Natural and ad hoc hierarchies in Hungarian morphophonology

Dániel Arató, Péter Rebrus
Eötvös Loránd University, HUN-REN Hungarian Research Centre for Linguistics
{arato.daniel, rebrus}@nytud.hun-ren.hu

Abstract

In Hungarian morphology, suffixed word forms display vowel/zero alternation. The distribution of the unstable vowels at the root–suffix morpheme boundary, commonly referred to as linking vowels, appears to be conditioned by a multitude of factors including several interacting aspects of the phonology of the word and its morphology (Siptár & Törkenczy 2000); in fact, the conditions of the occurrence of linking vowels are unpredictable, i.e. lexically determined. Beyond the basic dilemma of binary occurrence versus non-occurrence, linking vowels also exhibit occasional overabundance, i.e. cases where both possible forms are accepted realizations of the same paradigm cell (Kálmán et al., 2012). The loci of such overabundance are not trivial to predict either.

Rebrus (to appear) observes that mapping out word paradigms in a suitably ordered two-dimensional table of roots and suffixes reveals the distribution of linking vowels to be less haphazard than it might seem at first: two strikingly separate areas of certainty emerge in the table, one with linking vowels and the other without linking vowels, although with a not-so-neat staircase-like border between them. Along the border we find some overabundant cells. He suggests that such a pattern may be the result of a gradual erosion of an earlier cleaner, more distinct and unambiguous separation of the two areas over the history of the language – a process of erosion driven by analogical attraction between neighboring forms.

We propose a rough computational model to simulate what such a gradual transformation driven by analogy might unfold like over time, and use it to evaluate the hypotheses that local[1] analogical pressures will 1) almost always dismantle an initial rectangular division of the two rival areas; and 2) usually keep the convex, "monotonic" shape of the division intact. We find support for these assumptions under reasonable free parameter settings.

A tangentially related investigation into a subset of linking vowels' distribution from a purely phonological viewpoint found the same staircase shaped separation, this time arising from a natural ordering of phonological forms (Blaskovics 2024), hinting that these kinds of patterns may be more widespread than initially assumed.

Introduction

In morphology, speakers of a given language need to be able to reliably supply the correct word form of any concrete lexeme coupled with any set of morphosyntactic features, i.e. for any cell in its full inflectional paradigm. For example Finnish *talo* 'house' in a context requiring plural number, inessive case and first person singular possessor becomes *taloissani* 'in my houses'. Compared to determining full inflected forms, a strictly easier task is to determine whether a certain morphological trait or element will occur in each paradigm cell of a certain lexeme or not. E.g. the *-i-* element in the previous Finnish example is found in some but not all plural cells in the paradigm of *talo*.

A highly challenging and intricate question of this kind is presented by the inconsistent occurrence of so-called **linking vowels** in Hungarian: the singular accusative of *gáz* 'gas', for example, is *gázt*; but the accusative of *ház* 'house' is *ház**a**t* (not **házt*). Predicting the

[1] "Local" in the sense of short-distance interactions between cells of a paradigmatic system, not "local" in the sense of short-distance effects around a phoneme or morpheme in a linguistic form.

presence or absence of such vowels in the inflected forms of Hungarian nouns and adjectives is a notoriously complicated problem. The distribution of linking vowels is subject to a number of relative, rather than absolute phonological as well as morphological criteria: most of these generalizations only partially determine the outcome. Importantly, large parts of the distribution cannot be fully captured by looking only at the root (e.g. *házat* 'house.ACC' but *háztól* 'from a house', not **házatól*) or only at the suffix (see again *gázt* vs *házat*). The quality, specifically the backness and openness of linking vowels varies too, but we shall not concern ourselves with that additional problem in this paper.

Root	ACC.SG	NOM.PL	Gloss
hal	*hal**a**t*	*hal**a**k*	'fish'
lap	*lap**o**t*	*lap**o**k*	'sheet (of paper)'
dal	*dalt*	*dal**o**k*	'song'
kocsi	*kocsit*	*kocsik*	'car'

Table 1, based on Kálmán et al. (2012)

Even though a number of sweeping generalizations exist (for instance no root ending in a vowel will ever take a linking vowel in its singular accusative form) trying to formulate the totality of the distribution of linking vowels in terms of such clean general categorical conditions seems almost futile. Of course, it is even more hopeless and absurd to propose the opposite: that speakers independently memorize every single cell of every single lexeme. This approach would not help us explain two remarkable facts about inflection in general: firstly, how speakers are able to deal with paradigm cells they have never encountered or produced before (whether vanishingly rare cells of existing words such as *novembereinktől* 'from our Novembers', or completely unattested made-up words like *zupter* : *zuptereinktől* 'from our (nonsense root)s'); and secondly why they sometimes exhibit vacillation a.k.a. overabundance (i.e. use and acceptance of both forms in the same paradigm cell, such as *rozst* ~ *rozsot* 'rye.ACC' or *sanszt* ~ *sanszot* 'chance.ACC'). That leaves us with a slightly less elegant third option for analysis: that the morphology of linking vowels is lexicalized only to a certain degree, and processes of on-line generation are also at play, again to a certain degree. More commonly attested and consequently better memorized word forms will likely be produced by simple lexical retrieval, while forms there is not much call for will presumably rely more on dynamic on-demand production[2] (Bybee, 2006). It is also not unreasonable to expect high-frequency forms to act as beacons of certainty that project their influence about them in the system and potentially sway other forms in their relative vicinity (Kálmán et al., 2012).

Literature review

There are some well known abstract binary criteria on the occurrence of linking vowels making reference to the phonology of the root morpheme, the phonology of the suffix, or lexically assigned arbitrary features; criteria that cover small parts of the complete

[2] Note that this is different but not entirely orthogonal to what's usually meant by a theory of "dual-route morphology" (Pinker, 1991): different because we do not claim that it's the irregular lexemes that are more lexicalized; but in practice not orthogonal either because there tends to be a strong correlation between irregularity and frequency of use.

distribution. E.g. the ablative suffix *-tól* ‘from’ never adds a linking vowel; and as mentioned before roots that end in a vowel also form their accusative without a linking vowel, without exception. As far as explanatory accounts go, some treatments employ rule-based procedural vowel insertion or deletion, while others suggest intricate internal representations, yet others a combination of the two. For a detailed description of each flavour of theory and of the empirical data themselves see Siptár & Törkenczy (2000). Given the complex and unpredictable nature of the data involved, the presentation of which goes beyond the scope of this paper, it may often be useful to consult the Hungarian Gigaword Corpus (Oravecz et al., 2014) about specific word forms.

As noted by Kálmán et al. (2012), all theoretical models relying on categorical binary representations and clear-cut binary rules responsible for decision-making in the production and processing of such word forms face a fundamental question: that of overabundance. Where does overabundance come from? Note that overabundance (also referred to as variation) only happens in quite specific regions of the grammar, it does not occur uniformly, so it cannot be due to an intrinsic imperfection or randomness of the production process. Strangely enough, it appears as though not just the presence or absence of an extra vowel but even the fact of overabundance were lexically encoded in speakers’ linguistic knowledge. The authors address this matter through the alleged mechanism of analogy at a relatively high level, claiming that several different lexical subclasses interact with each other, and overabundance obtains where their spheres of influence happen to overlap.

Root	ACC.SG	Prevalence of LV forms	Gloss
/mɛlɒnʒ/	/mɛlɒnʒt/ ~ /mɛlɒnʒ**o**t/	69.5%	‘mélange’
/pɛrformɒns/	/pɛrformɒnst/ ~ /pɛrformɒns**o**t/	58.5%	‘performance’
/peːnz/	/peːnzt/, */peːnzɛt/	0.1%	‘money’

Table 2, based on Kálmán et al. (2012)

Rácz et al. (2018) take a fundamentally similar approach in terms of the core cause of overabundance, and identify several different types of “attractors” – i.e. potential prototypes that may exert an analogical pull on other forms – for Hungarian verbs, some referencing the phonology, morphology or even transitivity of the root, some based on paradigmatic similarities between different verbs, and some having to do with frequencies. This more refined and more explicit model is used to generate quantitative predictions about what specific verbs will show overabundance.

In agreement with the previous sources, Rebrus et al. (2024) argue that efforts to establish discrete linear segmentations for all suffixed surface forms (which would definitively assign linking vowels to one individual morpheme or the other) is futile to begin with, and instead recommend a holistic treatment of word forms in an analysis of the quality of linking vowels where once again paradigmatic relationships and analogical similarities run the show. They explicitly emphasize that traditional formalist theories based on discrete features seem fundamentally unfit for describing and capturing the kind of gradual, non-categorical data that we are dealing with here.

Methodology

As mentioned before, the existence of one or more absolute implicative generalizations does not imply that the whole of the distribution can be covered by and derived from such generalizations. Just because a certain root (e.g. *hajó* 'ship') or a certain suffix (e.g. *-tól* 'from') can be asserted never to take a linking vowel, this does not mean that the rest of the distribution can be captured in such terms. Indeed this hardly seems to be the case.

Instead, in a two-dimensional table of fittingly ordered lexemes and suffixes (what we may call a paradigmatic system), pairs of unidirectional implications appear. In fact, Table 1 already hints at such an implication between singular accusatives and plural nominatives. Both of the following generalizations hold: **a1)** if in the paradigm of a specific lexeme the ACC.SG form has a linking vowel, the NOM.PL form will also have one; and **a2)** if the NOM.PL form of a lexeme does not have a linking vowel, the ACC.SG form won't have one either. Similar implicative interrelations exist between individual lexemes, for example: **b1)** if a given suffix attached to the noun *bár* 'bar' triggers the inclusion of a linking vowel, the same suffix will also trigger a linking vowel on *vár* 'castle'; and **b2)** if a suffix attached to *vár* does not trigger a linking vowel, it will also not trigger one on *bár*. Although these kinds of double implications abound, it is still strikingly common for pairs of morphemes not to behave in the same way throughout their paradigms: consult Table 3 where the pattern of the superessive *-n* 'on' is misaligned with both the singular accusative and the plural nominative; and likewise the paradigm of *lassú* 'slow' matches neither that of *bár* or *vár*. These specific morphemes are by no means unique in their behavior, nor do they exhaustively represent all the different morphological subcategories found in the grammar with respect to the linking vowel dilemma.

This state of matters suggests a kind of hierarchy among roots as well as among suffixes as to how "pro-linking-vowel", or conversely how "anti-linking-vowel" they appear to be. Note that this hierarchy is not binary in either the root dimension or in the suffix dimension: if Hungarian nominal morphemes truly have such a lexical quality associated with them, it is not a discrete two-valued feature. This is plainly seen in Table 3 below, where the full interparadigmatic pattern is such that it cannot be captured by a logical function in two arguments (the linking vowel preference of the root and that of the suffix), not by their conjunction, disjunction, exclusive disjunction, or even an arbitrary tailor-made logical function. (Again, this is but a very small sample of all Hungarian nominal suffixes and especially of roots, presented as an illustrative example.)

	-k 'PL'	*-n* 'SUE'	*-t* 'ACC'	*-tól* 'ABL
háború 'war'	*háború-k*	*háború-n*	*háború-t*	*háború-tól*
lassú 'slow'	*lassú-**a**k*	*lassú-n*	*lassú-t*	*lassú-tól*
bár 'bar'	*bár-**o**k*	*bár-**o**n*	*bár-t*	*bár-tól*
vár 'castle'	*vár-**a**k*	*vár-**o**n*	*vár-**a**t*	*vár-tól*

Table 3, based on Rebrus (to appear)

If we disregard the changeable quality of the linking vowels and merely look at their sheer distributions in Table 3 for a moment, a characteristic division emerges between two large contiguous homogeneous blocks of cells that behave in a uniform way with respect to the linking vowel dilemma. It turns out that these two contiguous areas in the table, and the border resembling a pyramid or a staircase between them, that guarantee the kind of pairwise implications we have just seen, are hardly an accident. In fact similar patterns (allowing adjacent identical rows and adjacent identical columns) hold for the entirety of the lexicon, i.e. for all roots and all suffixes – what's more, the same *a posteriori* hierarchy of roots and suffixes seems to reveal this same pattern even for completely different morphophonological phenomena (Rebrus, to appear). It bears stressing that this closed or “monotonic” staircase pattern is not a trivial property even if we are allowed to choose our own hierarchies i.e. ordering of rows and columns. Consider a hypothetical situation where you have a linking vowel in the top left corner cell (Table 4). Now it is entirely impossible to rearrange the rows and columns to get back the same pattern, even for this tiny table of selected morphemes.

	-k ‘PL’	*-n* ‘SUE’	*-t* ‘ACC’	*-tól* ‘ABL
háború ‘war’	0	0	0	**1**
lassú ‘slow’	1	0	0	0
bár ‘bar’	1	1	0	0
vár ‘castle’	1	1	1	0

Table 4, a made-up scenario where the monotonic pattern cannot be restored by any ordering

While some parts of the root and suffix hierarchies that successfully produce this convenient pattern have phonological correlates, by and large they still represent artificial, post hoc orderings. However arbitrary and unnatural these after-the-fact hierarchies may be though, the stubborn repeated emergence of the monotonic staircase pattern might not be entirely devoid of systemic reason or function after all. In fact, this “geometric” congruence of cells in a paradigmatic system may be more than an accidental epiphenomenon: indeed the fact that stable linking vowel forms and stable no-linking-vowel forms appear in sweeping unbroken lumps in the table might help speakers keep the grammar straight in their minds, by virtue of keeping the entropy (or, roughly speaking, the arbitrariness) of the system in check. As long as the pattern is maintained, the number of possible states and thereby the entropy of the paradigmatic system is limited well below its maximum under no restrictions. Similar monotonic patterns found in other aspects of morphology may also allow the speaker to deploy the same implicational techniques and sometimes even reuse the very same maps in the acquisition of several unrelated morphophonological alternations.

One way such a monotonic staircase pattern may hypothetically come about is through accidental local changes in the exact shape of the border due to the analogical influence of nearby cells. E.g. the corner cell highlighted in Table 5 has two immediate neighboring cells (also highlighted) on the other side of the linking vowel border splitting the table in two. If we suppose that this exposure can potentially destabilize the affected cell and result in variation/overabundance (a linking vowel preference strictly between 0 and 1) in the cell, and with time perhaps even prompt it to desert its own camp and join the no-linking-vowel side, then we might have a viable explanation as to the origin of the actually attested, more complicated pattern.

0	0	0	0
0	0	0	0
1	1	1	0
1	1	1	0

Table 5, a hypothetical cleaner original state for the attested arrangement in Table 3

One piece of evidence in support of this assumption is the fact that most currently overabundant cells are found directly adjacent to the border. Moreover, we also have evidence from earlier historical stages of the Hungarian language that such re-categorizations have indeed happened to a number of different paradigmatic system cells (Rebrus, to appear).

Based on these observations, we define a very simple abstract algorithmic model of the linking vowel problem (or, by virtue of abstraction, any similar binary morphological phenomenon sensitive to paradigms). We will not attempt to characterize the actual distribution of linking vowels, but only to check whether the proposed internal mechanism of local analogical attraction results in the kind of monotonic distributions we observe in practice. Specifically we will try and see if two concrete expectations are born out by the model: 1) an initial rectangular pattern as in Table 5 is disrupted and broken down by the pressures of neighboring cells, and 2) the monotonic staircase pattern does not break down.

We assume word forms to be neither purely statically retrieved from passive memory, nor purely dynamically created on the fly every time the need arises, but rather something in between: both static retrieval and dynamic assembly affect the resulting surface form, but neither one must necessarily fully determine it. In accordance with prevailing theoretical opinion, the insertion of linking vowels is still assumed to be richly lexicalized, in fact we assume that speakers retain large amounts of redundant information about which forms prefer a linking vowel and which forms don't. Rather than trying to eliminate redundancies and minimize the amount of data stored in speakers' memory (which used to be a central guiding heuristic for theories of morphology), our model assumes that speakers employ their considerable capacity for learned information (Landauer 1986) to memorize and reproduce morphological properties.

The model will notably be 1) stochastic, i.e. deliberately involving a degree of chance and nondeterminism, 2) lexicalized, i.e. the internal feature meant to track how much a cell prefers inserting a linking vowel will be stored at the most granular level possible: the level of the single paradigm cell, 3) gradual, i.e. the internal feature will be a continuous scalar variable, and 4) local, i.e. interactions are only assumed between neighboring cells. The model is open source and implemented in Python.

The concrete algorithm is as follows. User defined variables are boldfaced.

0. Start with any 2D matrix of preassigned scalar values from [0, 1].

1. Repeat for **max_steps** times:

- **1.1.** Select a cell at random, uniformly.
- **1.2.** Let p be the average of this cell and all cells in the same row or column within a distance of **effect_radius**, weighting the selected cell itself by **cells_own_weight**.
- **1.3.** Select a random outcome of "yes" with the probability of p or "no" with a probability of $(1 - p)$.
- **1.4.** Is the outcome "yes"?
  - **1.4.a.** yes → Add **delta** to the selected cell's value.
  - **1.4.b.** no → Subtract **delta** from the selected cell's value.
- **1.5.** If the cell's value is now over 1 or under 0, set it back to 1 or 0, respectively.
- **1.6.** Label all cells strictly over **tripartite_cutoff** as "green", all cells strictly under (1 - **tripartite_cutoff**) as "red", and all others as "grey".
- **1.7.** Are all rows and all columns in the order "green"–"grey"–"red" (i.e. "monotonic") from left to right and bottom up, respectively?
  - **1.7.a.** yes → Continue as is.
  - **1.7.b.** no → Can rows and columns be reordered to restore monotonicity?
    - **1.7.b.a.** yes → Reorder rows and columns to restore monotonicity.
    - **1.7.b.b.** no → Continue as is.

2. Check if all rows and columns are monotonic at the end.

For simplicity we fix the **effect_radius** as 1, meaning cells can only be affected by immediately adjacent cells in the paradigmatic system. We have tested **cells_own_weight** values between 1 and 5, to very little difference in the outcomes. A **delta** of 0.1 is used because it has been found that much smaller values produce negligible perturbations in the system, while significantly larger values result in drastic changes that hardly match actual observations (variable **delta**s have also been implemented but it is not trivial to choose a suitable monotonically decaying function that accurately reflects speakers' tendencies to fossilize their grammar). The default **tripartite_cutoff** is set at 0.8 (corresponding to 80%) which is low enough to categorize all those word forms (in the actual observed data) that display marginal variation as pertaining to one or the other extreme of the scale.

The value in each cell represents that particular cell's preference or "willingness" to have a linking vowel in the word form that corresponds to it. Since we're trying to test what happens to an initial clean rectangular distribution, the initial state we use will always be set up like that. Step 1.2 is intended to model the dual dependence of a given cell's realization on both its own stored preference and on the analogical sway of its neighbors. Technically, cells at the edge of the table would be less likely to be swayed by neighboring cells on account of having one less adjacent cell. The actual implementation of the model includes a feature to counteract this: it pretends that the edge cell has another neighbor, the value of which is the average of the three actual neighbors. Step 1.3 allows for the state of the system to evolve in a nondeterministic manner. Step 1.4 emulates feedback: linguistic forms that are repeatedly used are known to entrench and fix themselves in the language, likely resulting in more frequent use etc. Note that in a situation where most cells not only have an extreme value equal to 1 or 0, but are also surrounded by similar cells, in most iterations of the cycle the outcome will nudge the cell towards the same extreme, and so step 1.4 and step 1.5 together will produce no change. This is expected: cells whose forms can be reliably produced are unlikely to shift one way or the other. Step 1.6 relaxes the strictest possible criterion of monotonicity (the one where all values between 1 and 0 in all rows and columns are expected to appear in monotonic order) to one allowing partial non-monotonicity within the "green", "grey" and "red" sections of a row or column (although the software implementation of the model also includes an option for strict evaluation of monotonicity). Thus the following row of values will still be considered monotonic under the **tripartite_cutoff** = 0.8 setting: [0.9, 1.0, 0.3, 0.5, 0.1].

It is of course true that in reality not all paradigm cells are created equal with respect to how they are implemented internally: in languages with case systems, for example, it is much more reasonable to presume that nominative word forms are mostly stored "verbatim" and oblique forms are more likely to be assembled on-line than the other way around. In the state of the model reported in the present paper we actually exclude this potentially drastic asymmetry; we simply assume that all cells use the exact same production mechanisms. This kind of differentiation may be included in later versions of the model if a sufficiently well-founded empirical metric for placing cells along the stored versus generated continuum can be established.

Regarding the empirical motivation for a stochastic model, a certain degree of nondeterminism in the on-line production of specific word forms is already evidenced both by the diachronic tendency of some lexemes to change their inflection class (consider Latin *fugere* 'flee' which shifted to *fugire*, or even English *dwarfs*, eventually replaced by *dwarves* (patterning after *wives* and *thieves*)) and by the occasional ad hoc lapses in the on-line production of native speakers (such as *runned* in place of *ran*, or *criterias* pro *criteria*).

A screenshot of the actual application before the simulation is started is shown in Figure 1. Note that the specific roots and suffixes assigned as labels to each row and column are only there for illustrative purposes and to enable the user to track any reorderings of the table.

They neither behave exactly like the initial state suggests, nor play a role in the operation of the model. The two signs at the top of the window, currently both green, show whether the current state of the model is "conjunctive"[3] (i.e. out of the pattern of 1's and the pattern of 0's at least one is rectangular) and whether it is "monotonic" (i.e. complying with the empirically observed staircase pattern, possibly with a layer of "grey" cells along the border).

Figure 1, the software implementation of the model in operation

Results and discussion

We report the predicted diachronic outcomes simulated with the custom software based on the abstract model above with a variety of different user settings. All simulations were run with a 5x5 starting paradigmatic system state with a contiguous 3x3 block of cells in a corner initially assigned a 1.0 current preference, and the rest of the cells assigned a 0.0 current preference (see Figure 1). All results presented in the following table are percentages of the number of conjunctive and the number of monotonic model states after **max_steps** = 1000 iterations, over several repetitions of the whole simulation with the same parameter settings.

settings			results: final state…	
cells_own_weight	**delta**	**tripartite_cutoff**	**…conjunctive?**	**…monotonic?**
1	0.1	0.8	5%	94%
2	0.1	0.8	11%	94%
3	0.1	0.8	17%	92%

[3] The name "conjunctive" reflects the fact that the logical conjunction of two underlying variables would result in such a rectangular pattern in the paradigmatic system. Note that all paradigmatic systems that are conjunctive are by definition also monotonic.

4	0.1	0.8	21%	93%
5	0.1	0.8	33%	95%
1	0.2	0.8	3%	76%
2	0.2	0.8	4%	65%
3	0.2	0.8	2%	63%
1	0.1	0.9	0%	74%
2	0.1	0.9	1%	60%
3	0.1	0.9	2%	63%

Table 6, simulation outcomes with some sensible model parameter settings

Turning to our two main hypotheses mentioned earlier, we expect to see marked tendencies in the results columns in Table 6. The second to last column shows the number of simulations where the initial rectangular or conjunctive pattern did not break down or if it did it was not reconstituted (possibly in a different height or width) by the end of the simulation. As seen in that column, the initial rectangular or conjunctive pattern tends not to survive, mirroring our preliminary expectations. This means that analogical attraction as formalized in the model does indeed result in the probable breakdown of an original conjunctive pattern and its degradation to a monotonic staircase pattern or something even more arbitrary.

As to the monotonic property, the last column tells us that it is much more resilient to the effects of analogical pressures than the conjunctive property, indeed 65% or more of all simulation runs ended up in a monotonic state in all model settings tried above. This finding seems to corroborate our second hypothesis that the monotonic property tends to defy the complicating effects of analogy. It should be noted though that strict monotonicity (where values are expected to be strictly in order, not just according to their classification as “green”, “grey” or “red”) typically holds less than 50% of the time at the end for most parameter settings, far less often than the relaxed monotonicity applied by default in our experiments.

Aside from the main hypotheses, we may draw some minor additional conclusions from the results. The model turns out to be quite robust with respect to the **cells_own_weight** option: other than making the conjunctive pattern slightly less resilient, lower **cells_own_weight** settings do not seem to destabilize the monotonic pattern significantly compared to higher settings; in fact for some unclear reason the opposite seems to be the case, perhaps because the occasional defecting cell is brought back into the fold easier by its neighbors when **cells_own_weight** is low. Conversely, a higher **delta** results in considerably increased instability in the model, which makes sense because a higher **delta** means it takes a cell fewer iterations to be tipped over into a different category. Tighter **tripartite_cutoff**s naturally result in an even lower amount of conjunctive end states than before, because they allow edge cells less room to change their values before they are reclassified as overabundant “grey” cells and the resulting compromised state is considered non-conjunctive. Monotonicity is also affected, apparently because it becomes easier for “grey” cells to end up inside “non-grey” territories.

In an unforeseen accidental addition to this research, the linking vowel distribution in the singular accusative forms of nominal lexemes ending in two consonants specifically, examined by our colleague Blaskovics (2024), has provided a valuable supporting argument to the claim that staircase patterns in Hungarian morphophonology are not a mere accident. Nouns and adjectives ending in consonant clusters are somewhat rare in Hungarian, but several different kinds of word-final consonant combinations exist. Observe the following table showing the presence or absence of linking vowels in accusative forms depending on the type of the last and the second to last phoneme of the root (single-consonant roots are also included in the bottom row for completeness):

C_1↓ C_2→	**stops**	**sibilants**	**nasals**	**[l]**	**[r, j]**
stops	*matt-**ot***	*mátri*[ks]-***ot***			
sibilants	*ko*[st]-***ot***	*pa*[sː]-(***o***)*t*			
nasals	*pánt-**ot***	*sa*[ns]-(***o***)*t*	*finn-t*		
[l]	*Olt-**ot***	*Wales-*(***e***)*t*	*Lincoln-t*	*modell-t*	
[r, j]	*rajt-**ot***	*Wars-*(***o***)*t*	*Western-t*	*fájl-t*	*Knorr-t**
vowels	*bot-**ot***	*oázi*[ʃ]-*t*	*kinin-t*	*textil-t*	*radar-t*

Table 7, based on Blaskovics (2024)

Even though this data comes from a quite special subset of the lexicon of Hungarian, and the two dimensions of the table map the phonology of the root as opposed to the morphology, and has nothing to do with paradigms in the normal sense, we still see a highly similar monotonic pattern emerge as in the tables of roots versus suffixes before. Moreover, this time the monotonic pattern arises naturally by itself, with the penultimate and final phonemes appearing in their *a priori* order according to their sonority, no *ad hoc* orderings needed. The empty cells are empty due to phonotactic constraints on word-final clusters.

Two complications should be mentioned though. First, Table 7 shows only the productive side of morphophonology, with special closed lexical classes sometimes exhibiting different behavior. Second, the cell marked with an asterisk at the right edge of Table 7 is problematic, not because it features overabundance, but because it contains some lexemes that definitely take a linking vowel in their accusative form and some lexemes that definitely don't, e.g. *Tarr-t* '(a proper name).ACC' but *férj-**et*** 'husband.ACC'. Apart from these complications the linking vowel and no-linking-vowel areas display remarkable homogeneity once again in spite of the lack of a clear conjunctive divide.

Conclusions

Like other grammatical phenomena that not only show graduality and intra-speaker variation, but seem to elude categorical generalizations altogether, the morphological distribution of linking vowels in non-monomorphemic Hungarian word forms not only represents a concrete hard problem in morphology but also a theoretical challenge to some mainstream theories of grammar that suppose that speakers rely on the mechanical application of predetermined, discrete formal rules to generate non-trivial word forms. A potentially promising attempt at modeling this phenomenon is to suppose liberal lexicalization of a gradient preference of each paradigmatic system cell for including a linking vowel. Under reasonable assumptions as to how such a system might operate and change over time, this model seems to reproduce the kind of convenient paradigmatic arrangements observed in the actual data. Such an abstract model based on assumptions of analogy and feedback might also find uses beyond the scope of morphophonology.

It may be worth reiterating that the analogical mechanism we propose does not directly subserve the production or processing of concrete legal linguistic forms, as normally expected from elements of grammar; if anything, it complicates them. Rather, it is a consequence of the nondeterministic nature of analogy in morphological production and the constant feedback or the forms produced or encountered, each one changing the speaker's private morphology ever so slightly. Put another way, ordering-dependent local analogical attraction is not a factor of "grammar" in the sense generally construed by mainstream modern formalist theories. Rather, the behavior of linking vowels in Hungarian nominals, or at least part of it, may be considered a side-effect of the cognitive implementation of grammar, and of speakers' natural expectations of uniformity and regularity, their concrete distribution emerging as a consequence of the way human memory and recall seem to operate. The unexpectedly similar contiguous two-dimensional shape we find in the distribution of those same linking vowels when looking only at the phonology of roots ending in consonant clusters seems to corroborate this view and suggest that this characteristic monotonic staircase pattern may turn up in many unrelated areas of grammar for similar reasons.

Acknowledgement

This research was funded by NKFIH grant #139271 titled "The role of paradigm structure in Hungarian morphology and phonology with typological comparisons".

References

Blaskovics, Á. (2024). *Variation in the accusative form of Hungarian loanwords ending in coronal consonant clusters.* (Conference presentation at the "Nonuniformity in Morphophonology across Frameworks" workshop, Trier.)

Blevins, J. P. (2016). *Word and paradigm morphology.* Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199593545.001.0001

Bybee, J. (2006). *Frequency of Use and the Organization of Language.* Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780195301571.001.0001

Bybee, J. (2006). *From usage to grammar: The mind's response to repetition.* Language, 82(4):711-733. Linguistic Society of America. http://dx.doi.org/10.1353/lan.2006.0186

Kálmán, L., Rebrus, P. & Törkenczy, M. (2012). *Possible and impossible variation in Hungarian.* Current Issues in Morphological Theory: 23-49. John Benjamins, Amsterdam. http://dx.doi.org/10.1075/cilt.322.02kal

Landauer, T. K. (1986). *How much do people remember? Some estimates of the quantity of learned information in long-term memory.* Cognitive Science 10 (4):477-493. Wiley. https://doi.org/10.1207/s15516709cog1004_4

Oravecz, Cs., Váradi T. & Sass, B. (2014). *The Hungarian Gigaword Corpus.* In: Proceedings of LREC 2014. https://clara.nytud.hu/mnsz2-dev/en/

Pinker, S. (1991). *Rules of language.* Science, 253 (5019):530–535. https://doi.org/10.1126/science.1857983

Rácz, P., Rebrus, P. & Törkenczy, M. (2018). *Attractors of variation in Hungarian inflectional morphology.* Corpus Linguistics and Linguistic Theory 17 (2):287-317. De Gruyter Mouton. https://doi.org/10.1515/cllt-2018-0014

Rebrus, P. (to appear). *Regularity and exceptions in paradigmatic systems.* In: The Wiley Blackwell Companion to Diachronic Linguistics. John Wiley & Sons Ltd.

Rebrus, P., Szigetvári, P. & Törkenczy, M. (2024). *No lowering, only paradigms: A paradigm-based account of linking vowels in Hungarian.* Acta Linguistica Academica 71(1-2):137-170. http://dx.doi.org/10.1556/2062.2023.00674
Siptár, P. & Törkenczy, M. (2000). The Phonology of Hungarian. Oxford University Press, Oxford. http://dx.doi.org/10.1093/oso/9780198238416.001.0001

The implementation of the model introduced in this paper can be found on GitHub: https://github.com/nilthehuman/pyramid