Attractors of variation in Hungarian inflectional morphology

Abstract

We use algorithmic learning and statistical methods over a form frequency list (compiled from the Hungarian web corpus) to investigate variation in Hungarian verbal inflection. Our aims are two-fold: (i) to give an adequate description of this variation, which has not been described in detail in the literature and (ii) to explore the range and depth of lexical attractors that potentially shape this variation. These attractors range from closely related ones, such as the shape of the word form or the behaviour of the verb's paradigm, to broad ones, such as the behaviour of similar verbs or the phonotactics of related verb forms. We find that verbal variation is predominantly determined by similarity to related verb forms rather than by word shape or by word frequency. What is more, the effect of similarity is better approximated using inflected forms as opposed to base forms as points of comparison. This, in turn, supports a rich memory model of morphology and the mental lexicon.

keywords: morphophonology, language variation and change, similarity, frequency, Hungarian

Morphophonological variation is generally driven by stylistic or social factors (see e.g. Tagliamonte & Baayen 2012). Beyond these factors, such variation also uncovers parts of linguistic structure where exponents are underdetermined – where, in a way, multiple solutions are available for the same problem. In these cases one may encounter otherwise unseen attractor biases including segmental co-occurrence preferences (see Frisch et al. 2004) or the pull of related forms showing similar behaviour (see Krott et al. 2001). (In a dynamical system, an *attractor* is a part of the state space towards which the system tends to evolve, see e.g. Milnor 1985.) Variation effectively reveals wider aspects of linguistic structure.

English stress placement offers a simple example of such attractor biases affecting variation. Primary stress on the noun stem is preserved as secondary stress when the stem is combined with a stressed suffix, such as '-ation' (*accrédit – accrèditátion*; *imágine – imàginátion*). If this secondary stress would surface adjacent to the suffixed form's primary stress, it typically moves (*consérve – cònservátion*; *confírm – cònfirmátion*).

For certain stems, where these patterns clash, secondary stress placement is variable. Secondary stress either shifts from its original position, or occurs adjacent to primary stress (*condénse – còndensátion / condènsátion*). (Examples are from Pater's (2000) analysis of English stress placement.) While the larger picture is more complex, here we can say that words like 'condensation' reveal competing biases in the language.

In this paper we explore variation in a class of Hungarian verbs. We identify an underdetermined pattern of suffixation, and look at the various attractors that come

into play in variant selection. The focus is on the lexical aspects of variation, using the underdetermined pattern to shed light on the structure of the mental lexicon. We look at (i) specific lexical attractors that follow from a given verb's paradigm, (ii) broader attractors that follow from generalisations across verbal paradigms, and (iii) even broader ones that hold over the entire lexicon.

Instances of variation in morphophonology provide useful insight into the organisation of word-formation patterns in language (Bybee, 1985) and the structure of the mental lexicon (Hay & Baayen, 2005; Pierrehumbert, 2012). Our aims are (i) to give an adequate description of this Hungarian verbal pattern, one that can be generalised to similar verbal inflection patterns, and (ii) to draw inferences on the overall structure of the mental lexicon.

We use a webcorpus to approximate the ambient language, and an instance-based learning algorithm and hierarchical generalised linear regression to model lexical attractors. The resulting model does not provide an exhaustive description of morphophonological variation in Hungarian. It does, however, offer a heuristic tool to study lexical structure: it points us to attractors that are relevant to variation and provides a measure of their relative importance.

1 Lexical representation and variation in the community

Variation in the speech community is the sum of variable individual behaviours. In turn, individual variation is both influenced by patterns in the ambient language and the structure of the individual's mental lexicon. That is, variable behaviour, community-level patterns, and lexical organisation are all connected to each other.

Psycholinguistic research shows that individual linguistic behaviour is shaped by lexical organisation. Word forms are primed by related forms (Grainger et al., 1991) and the strength of the prime reflects structural relatedness (Gonnerman et al. 2007; this effect is less clear for affixal primes, see Dominguez et al. 2010; Duñabeitia et al. 2011). Word length and lexical neighbours have also been shown to affect word activation (see Grainger 1990; Carreiras et al. 2006).

Individual behaviour is also influenced by patterns in the speech community. This can be seen in studies of sociolinguistic variation (Labov, 2011). For example, individual style shifting follows the established patterns of formal and informal language use in the community. Language use in the community determines the frequency and predictability of words in the ambient language (the linguistic patterns that the listener encounters day-to-day), which are reflected in the individual's linguistic behaviour (Rumelhart & McClelland, 1986; Bresnan et al., 2007).

Conversely, lexical organisation affects patterns in the speech community. Types of sound change are influenced by similarity between words, analogy shapes morphological change, and gangs of similar words are more prone to resist change in general (Paul, 1880/1995; Wang, 1969; Bybee, 1995; Cuskley et al., 2014).

What follows is that lexical organisation and community-level variation are interrelated. For a given linguistic pattern, there is a trade-off between individual biases and community-level conventions (see e.g. Christiansen & Kirby 2003; for a trade-off based treatment of word length and word use, see Kirby et al. 2015).

Models of language recognise this relationship by assuming that certain patterns of

community-level use (such as word predictability) are represented in the individual's mental lexicon. There is an extensive literature on this topic that we cannot cover in detail (but see e.g. Skousen 1989; Bybee 1995; Colé et al. 1997; Baayen 2007 for some discussion). In this paper, we focus on the depth and resolution of these representations.

Morphophonological processes are, to some degree, reflections of this lexical organisation (Hay & Baayen, 2005; Rácz et al., 2015). They are probably not best modelled using a purely lexical approach, but it is clear that they rely on lexical organisation to a large degree (Albright, 2009; Hayes & Wilson, 2008; Pierrehumbert, 2016). Observed morphological patterns cannot be based on exhaustive lists of form-meaning pairs, nor can they be described purely in terms of abstract rules.

This claim raises some interesting questions about the relationship between patterns in the language community and the individual's lexicon. Which aspects of community-level variation is represented in the speaker's lexicon? What degree of detail is available to the speaker as they select a variable morphophonological exponent? In short, what are the relevant lexical attractors of variation?

One way to approach these questions is to take an example of community-level morphophonological variation and test a range of lexical attractors as potential predictors of this variation. If a lexical attractor is useful in explaining how the pattern behaves at the community-level, then it is likely available for individuals who are ultimately responsible for community-level variation. In a way, individual behaviour mirrors community behaviour. It has been demonstrated that a wide array of lexical attractors are active in morphophonology (see e.g. Dąbrowska 2008; Myers & Li 2009, for an overview, see Rácz et al. 2016). As a consequence, morphophonology provides an ideal testing ground for the relationship of community-level and individual-lexical variation.

In what follows, we address this relationship in two steps. We use an operationalisation which relies on a corpus to model community-level use and lexical attractors, and we investigate relationships between them through a range of exploratory methods that include category learning and numeric prediction. This approach does not favour hypothesis testing, but works very well in generating hypotheses about the structure of the mental lexicon.

Our emphasis is on the relative strength of variants in a given instance of variation, rather than the ontologically much more diffuse question of why certain forms are stable while others show variation in the first place.

2 The problem space

2.1 The CVC/CC class of Hungarian verbs

In what follows, we use a concatenative terminology to describe the problem space: *suffixation in Hungarian verbal inflection*. This terminology is adequate to give a simple description of Hungarian verbal suffixation. (It makes no commitments over either the individual's model of Hungarian verbal inflection or the description of Hungarian inflection in its entirety.) We loosely follow the Leipzig Glossing Rules (Bickel et al., 2008) in our examples (verb indefiniteness is not marked).

We can describe Hungarian verbal inflection as the concatenation of suffixes to stems, where the mechanisms of concatenation depend on the shape of the stem and the category of the suffix. Suffixes can be *analytic*, *quasi-analytic*, or *synthetic*. *Analytic* suffixes (C-) are always consonant-initial, regardless of the final segment(s) of the stem. However, some

suffixes show C/V alternation: *synthetic* suffixes (V-) are vowel-initial after consonant-final stems, whereas *quasi-analytic* suffixes (C/V-) are consonant initial when the stem ends in a single consonant but vowel-initial after cluster-final stems (Rebrus, 2000). For the purposes of this description, it suffices to say that *quasi-analytic* suffixes are variably C- or V-initial (C/V-).

inflection then depends on (i) whether the stem ends in a consonant cluster (CVC/CC) and (ii) whether the specific suffix is analytic (C-), synthetic (V-), or quasi-analytic (C/V-).

On one hand, if a quasi-analytic suffix is attached to a CC-final stem, a *linking vowel* appears between the stem and the suffix (e.g. [hord] + [nɒ] = [hord-ɒnɒ], 'carry-3SG.COND'). A CVC-final stem and a quasi-analytic suffix can combine without a linking vowel (e.g. [aːpol] + [nɒ] = [aːpol-nɒ], 'nurse-3SG.COND').

On the other hand, for a *class of stems*, the vowel of the stem-final CVC sequence does not appear when the stem is followed by a V-initial suffix (e.g. [søpør] + [øk] = [søpr-øk], 'sweep-1SG.IND'). The stem vowel is, at least in some stems, lexically specified and the linking vowel is determined by vowel harmony. This means that the behaviour of concatenated forms is partly phonological and partly lexical. In sum, stem and suffix combinations vary according to the presence of a linking vowel and the variability of the final stem vowel. Our focus is on the variable stem vowel (see below).

Table 1 provides a schematic outline of these classes using a representative set of verbal suffixes (for an overview, see Lukács et al. 2010). The table shows a set of representative verbs (rows) across a set of representative suffixes (columns); suffix type (no suffix / V-initial / C-initial / C/V-initial) is indicated in the header.

From top to bottom: (1) The *stable* class always has a stem vowel irrespective of the suffix. (2) The *weak* class loses the stem vowel when followed by a V-initial suffix (as in the 1SG.IND) but retains it when followed by a C-initial suffix (as in the 3SG.IMP or the 3SG.COND). (3) The *variable* class behaves similarly to (2) in that the stem loses the vowel when followed by a V-initial suffix. However, for C/V- suffixes (where both variants of the suffix are available in this stem class), these stems allow for two repair strategies: the stem vowel is maintained *or* a linking vowel is added between the stem and the suffix. This variation can be seen in the 3SG.COND. (4) The *no vowel* class contains stems that end in a consonant cluster that is never broken up by a vowel, irrespective of the suffix. These stems either use no linking vowels (with analytic suffixes, as e.g. 3SG.IMP) or opt for a linking vowel between stem and suffix (as e.g. in 3SG.COND). Finally, (5) *defective* stems tend to end in consonant clusters of increasing sonority, which cannot be broken up by a vowel (unlike in classes 2 or 3). When they combine with analytic suffixes, which do not allow for a linking vowel between stem and suffix (such as the 3SG.IMP), the derived form cannot be repaired in any way, resulting in a paradigmatic gap.

Verbs in classes (3,5) all take the suffix '-ik' in 3SG.IND. This suffix is only present for a subclass of Hungarian verb forms. The '-ik' suffix favours a CC variant of the form.

This paper focusses on Class (3), the class of CVC/CC verbs. These verbs show variation between two repair strategies if they are combined with C/V-initial suffixes. These forms either follow classes (1-2) and behave like *CVC-final* verbs, or follow classes (4-5) and behave like *CC-final* verbs. As suggested by the example in Table 1 (fyrødnɛ / fyrdɛnɛ), a single stem can show variation, even with the same suffix. The aim of this paper is to account for the process that generates the observed CVC/CC variation in Class (3) in the speech community.

It is important to note that the 3SG.IND form of CVC/CC verbs shows very little

class	3SG.IND (no suf.)	1SG.IND (V-suf.)	3SG.IMP (C-suf.)	3SG.COND (C/V-suf.)	*gloss*
1 stable	aːpol	aːpolok	aːpoljon	aːpolnɒ	nurse
2 weak	ʃøpør	ʃøprøk	ʃøpørjøn	ʃøpørnɛ	sweep
3 variable	fyrdik	fyrdøk	fyrødjøn	fyrødnɛ / fyrdɛnɛ	bathe
4 no vowel	hord	hordok	hordjon	hordɒnɒ	carry
5 defective	ʃiklik	ʃiklok	*ʃikljon	ʃiklɒnɒ	slide

Table 1: Hungarian verb classes

variation and is almost always of the CC form (followed by the V-initial 3SG.IND suffix '-ik', obligatory for this class though not for all verbs). Only a few counterexamples exist in the corpus, such as [tɒjteːkzik] / [tɒjteːkozik] 'throw a tantrum-3SG.IND' or [dohaːɲzik]/ [dohaːɲozik] 'smoke-3SG.IND'.

2.2 Identifying attractors of the CVC/CC class

In Section 1, we set out to test the scope of active lexical attractors using variation in the speech community. Here, we focus on variation in a corpus of Hungarian, reflecting community-level use to see which patterns of variation in the corpus are likely to be represented by the individual. Our focus is restricted to forms that show CVC/CC variation; the question of why certain forms remain stable has been reserved for future work. The restricted approach does, to some extent, provide an answer to this broader question.

We propose four sets of lexical attractors of CVC/CC verb variation: *phonotactic*, *across-paradigm*, *within-paradigm*, and *form-specific*.

Phonotactic attractors The widest attractors are generalisations that hold across all words in the Hungarian lexicon – phonotactic patterns. CVC/CC variation interacts with *phonotactics*. It effectively avoids CCC clusters, which are typically (though not uniformly) broken up at morphological boundaries (as well as most monomorphemic forms) in the Hungarian lexicon. This avoidance of marked clusters suggests a link with the phonotactics; that is, variation might respond to stochastic restrictions on consonant clusters across the entire lexicon. We can test for this by seeing whether phonotactic aspects of these clusters affect variation in a robust way.

One phonotactic aspect of consonant clusters is their *sonority*. It is, in the most neutral terms, the set of cross-linguistic preferences of consonants when they occur adjacent to each other – in various constituent parts of the syllable, as well as heterosyllabically (Clements, 1990). Falling sonority is more typical of clusters in syllable-final positions, while rising sonority is more likely to indicate a syllable-initial or a heterosyllabic position.

We expect that the sonority slope of the consonant cluster will influence CVC/CC variation (Siptár & Törkenczy, 2000). That is, CC sequences with a more falling slope will be less likely to be broken up, because these sequences are more likely to occur stem-finally across the Hungarian lexicon.

Though the theoretical and empirical validity of sonority as a concept is often criticised (see Harris 2006), it still provides a useful approximation of broad patterns that hold across the lexicon.

A second aspect of consonant clusters is *homorganicity*. Most members of the stable CC class in Hungarian have homorganic clusters (e.g. [kɛzd] 'start-3SG.IND', [tɒrt] 'hold-3SG.IND', [aːld] 'bless-3SG.IND'). This trend can reflect a more general tendency that homorganic clusters avoid breaking up. (This pattern would be further complicated by the behaviour of geminates, but no geminates occur in the cluster in CVC/CC verbs.)

Across-paradigm attractors Generalisations across the general class of verbs constitute a narrower set of attractors. We can think of these as generalisations over *across-paradigm* similarity within this general class. The variable behaviour of CVC/CC verbs maps onto the stable behaviour of two sets of verbs in Hungarian. One set consistently shows CVC behaviour in the relevant exponents, and the other shows consistent CC behaviour. (These are classes 1-2 and 4-5 in Table 1.) A given CVC/CC verb will be similar to both CVC and CC verbs.

This *formal* (as opposed to semantic) similarity across verb classes constitutes an attractor of variation (see Bybee & Slobin 1982; Dawdy-Hesterberg & Pierrehumbert 2014), which leads us to expect specific CVC/CC verbs to move towards the verb classes to which they are more similar.

We expect that CVC/CC verbs that are overall more similar to CVC verbs will show more CVC behaviour, while those that are more similar to CC verbs will show more CC behaviour.

Assuming an effect of across-paradigm similarity, our account needs to further specify the basis of across-verb similarity in the respective verbal paradigms. Similarity between verb classes can be based on basic forms in the paradigm, as in Albright & Hayes (2003) or Hahn & Nakisa (2000). Alternatively, similarity can be also based on *inflected forms*. The use of inflected forms commits us to a rich memory model of the mental lexicon, in which inflected forms are also available as bases of similarity-driven processes (Johnson, 2006; Rácz et al., 2015).

Within-paradigm attractors We can consider this trade-off between model complexity and available information specifically *within the verb's paradigm*. On the one hand, the most frequent form in the verb's paradigm is usually the 3SG.IND. It can be seen as the prominent basic form of the paradigm (see Baayen et al. 1997; Blevins 2001; Booij 1999). Variable CVC/CC verbs form the 3SG.IND with a CC stem and the V-initial suffix '-ik'. CC variants of CVC/CC forms will be more similar to this basic form. On the other hand, most variable CVC/CC forms are, in fact, CVC forms, because of an overall skew of C/V-initial suffixes to be C-initial. As a consequence, CVC variants of CVC/CC forms will be more similar to these suffixed forms.

We expect that the token frequency of the 3SG.IND *and the range of variation across V-initial suffixes will affect CVC/CC variation* – a more frequent base form results in more CC forms, a wider range of suffix variation results in more CVC forms.

Word-specific attractors Finally, we define certain attractors as *specific to the verb form* itself. These should have an effect on the extent to which the CVC/CC form preserves the stem. *Compound* verbs are verbs that have a lexicalised preverbal particle (such as [fɛl-hɒngzik] 'become audible, lit. up-sound'). *We expect these forms to have more CVC versus CC forms.*

Verbs that have a *free stem* should also have stronger stem identity compared to verbs that have no recognisable free stems. For example, [aːrɒm-lik] 'flow-3SG.IND' has the

nominal stem [aːrɒm] 'flow'. The noun form can be seen in the verbal stem. In contrast, [boml-ik] 'decay-3SG.IND' has no similar pair: the stem (*[boml]) is not attested in itself. What follows is that *we expect that verbs with free stems will have more CVC versus CC forms.*

Most CVC/CC verbs are intransitive. Therefore, *we expect that transitive verbs will also have more CVC forms.* Since CC forms make the stem harder to identify if it is shorter (László Kálmán, p.c. see also Rácz & Rebrus 2012), we also expect a correlation between the length of the stem and CVC/CC variation. As a result, we expect a higher ratio of CC forms for these verbs.

Most of our attractors, such as token frequency, word length, or the structure of the lexical neighbourhood, are common in studies of lexical variation. Our aim is to offer a taxonomy of lexical attractors based on attractor specificity, as we believe that the extent to which these various factors contribute to CVC/CC variation in Hungarian morphology is indiciative of how inflected forms are processed, stored, and organised in Hungarian. The mechanics of Hungarian inflection are, in turn, relevant for broader theoretical approaches to inflectional morphology and its relationship to the mental lexicon.

3 Analysis

3.1 Data source

We approximate variation in the speech community using a frequency dictionary derived from the Hungarian Webcorpus (Trón et al., 2006). The dictionary is based on a version of the corpus that is morphologically analyzed (Trón et al., 2005) and morphologically disambiguated on the inflection level (Halácsy et al., 2007). Using an online corpus comes with limitations as the data do not closely reflect spoken language (Rácz et al., 2016). At the same time, the corpus is very large and covers a range of written registers, both formal and informal (1.48 billion words unfiltered / 589 million words fully filtered by a morphological parser).

Our query of varying CVC/CC verbs is restricted in two ways.

Suffixes We selected five C/V-initial suffixes that are frequent enough to carry substantial CVC/CC verb variation. This allows us to balance the resulting dataset across a limited number of suffixes that do not vary widely in frequency of use across verb stems, rendering multi-level modelling feasible. The five suffixes are listed in Table 2 below.

function	C-variant	V-variant	CVC example	CC example
3SG.COND	-nA	-AnA	aːrɒmolnɒ	aːrɒmlɒnɒ
3PL.IND	-nAk	-AnAk	aːrɒmolnɒk	aːrɒmlɒnɒk
INF	ni	-Ani	aːrɒmolni	aːrɒmlɒni
3PL.PAST.IND	-tAk	-OttAk	aːrɒmoltɒk	aːrɒmlottɒk
2SG.IND	-tOk	-OtOk	aːrɒmoltok	aːrɒmlotok

Table 2: The five C/V-initial suffixes with [aːrɒmlik] 'flow-3SG.IND'

In the annotations, 'A' represents front-harmony ([ɒ] ∼ [ɛ]), 'O' represents rounding harmony ([o] ∼ [ø] ∼ [ɛ]).

Verbs The CVC/CC class is a fuzzy set. We chose 111 that have a variable final vowel in the stem (according to our speaker intuitions[1]). The envelope of true variation is considerably narrower than these verbs, so that it is unlikely that we missed large parts of the varying vocabulary. We excluded verbs that are defective (e.g. [hɒɲɒtlik] 'decline-3SG.IND / *[hɒɲɒtoljon] 'decline-3SG.IMP'), stems that show a semantic split between the CVC and the CC form (e.g. [fɛnɛkɛlhɛt] 'ground/spank-3SG.MOD' – see [fɛnɛklik] 'ground' / [fɛnɛkɛl] 'spank'), and four stems that were identified as nouns by the parser (e.g. [ɛlhɒngzottɒk] 'what has been said-3PL.IND / 3PL.ADV').

As we note in Section 1, we aim to focus on forms that do show variation. Therefore, we introduced one numeric threshold: each verb had to have at least one CVC form and one CC form with any suffix in Table 2. Since missing forms are likely due to data scarcity (these are low token frequency forms), we opted for a lexical definition of defectiveness and a low frequency threshold for inclusion in the dataset.

Data size Our semantic and frequency restrictions reduce the sample size to 44164 instances of 163 CVC/CC variants of 39 verbs. The number of all possible CVC/CC pairs of the 39 verbs with five suffixes is $39 * 6 = 195$.

In all type counts below, we include the CVC and CC variants of all suffixed forms of all verbs – 195 variant pairs – and set the token frequency of unattested variants to 0. (See in Table 3.)

Calculating odds ratios for CVC/CC verbs Since the Webcorpus parser tends to regard CVC and CC forms as separate lemmata, we calculate lemma frequency for the verbs by hand. Lemma frequency is strongly correlated with the frequency of the 3SG.IND ($r = 0.54$).

Following Janda et al. (2010), we calculate the odds of the CVC variant over the CC variant for each suffixed form, by adding 1 to both frequencies and then calculating the odds in order to avoid dividing by zero. We use logged odds in the analysis. Table 3 shows the variants of [aːrɒmlik] 'flow-3SG.IND' in the sample.

stem	suffix	CVC form	CC form	CVC freq	CC freq	odds	log_odds
aːrɒmol	-nAk	aːrɒmolnɒk	aːrɒmlɒnɒk	69	593	0.12	-2.14
aːrɒmol	-tAk	aːrɒmoltɒk	aːrɒmlottɒk	19	172	0.12	-2.16
aːrɒmol	-nA	aːrɒmolnɒ	aːrɒmlɒnɒ	14	42	0.35	-1.05
aːrɒmol	-tOk	aːrɒmoltok	aːrɒmlotok	0	0	1.00	0.00
aːrɒmol	-ni	aːrɒmolni	aːrɒmlɒni	75	161	0.47	-0.76

Table 3: Attested variant pairs of [aːrɒmlik] in the sample

3.2 Operationalising attractors of the CVC/CC class

We operationalised the attractors discussed in Section 2.2 as the following predictors:

1. phonotactic: sonority, homorganicity

[1]All three authors are native speakers of Educated Colloquial Hungarian, a language variant named by the third author.

	target	source	
type	CVC/CC verbs	stable CVC	stable CC
CVC	aːrɒmolnɒk	aːpolnɒk	
CC	aːrɒmlɒnɒk		hordɒnɒk

Table 4: Sources and targets in across-paradigm analogy

2. across-paradigm: 3PL.IND form's similarity to stable CVC verbs and to stable CC verbs

3. within-paradigm: frequency of 3SG.IND, lemma frequency of verb, type frequency of attested C/V-initial suffixed forms

4. word-specific: is the form a compound?, is a free base attested?, is the verb intransitive?, length of base in syllables

Phonotactic predictors For each verb, we tag the two consonants in the final CVC/CC cluster as hetero- or homorganic and calculate a sonority slope based on their relative sonority. The sonority of a given segment ranged from 1 (high sonority: the glide [j]) to 8 (low sonority: voiceless stops). The sonority of the pair is the sonority of the first segment subtracted from the sonority of the second segment. This means that a high value describes a falling slope of sonority (such as [jk]), while a low value describes a rising slope of sonority (such as [tr]). Clusters with a high value tend to prefer syllable codas, clusters with a low value tend to prefer syllable onsets.

Across-paradigm predictors Our expectation is that variable verbs behave like stable verbs to which they are more similar in form . The relevant classes are covered in Section 2.1. One class of verbs is stable CVC and takes C-initial variants of C/V-initial suffixes. Another class is stable CC and takes V-initial variants of C/V-initial suffixes. If a CVC/CC verb is more similar to stable CVC verbs, it will behave more like a stable CVC verb, whereas, if it is more similar to stable CC verbs, it will behave more like a stable CC verb.

The similarity argument seems circular if we focus only on the morpheme boundary between stem and suffix. In this case, e.g. a CC verb will be a CC verb by virtue of not breaking up the cluster in suffixed forms.

However, similarity holds across the entire form – CVC/CC *stems* have quantifiable similarity to stable CVC versus CC stems. At the same time, we want a similarity-based account to be based on real forms and not posit abstract stems or base forms as points of comparison. Extant forms of CVC/CC verbs, on the other hand, are either CVC or CC.

One way around this is to pick a suffixed form that occurs both in the CVC and CC form and that is frequent enough so that a similarity-based model will cover a large part of the potential parameter space.

We select the 3PL.IND, a frequent suffixed form that shows CVC/CC variation, and compare the 3PL.IND of variable CVC/CC forms to stable CVC and CC forms. We do so with both the CVC and the CC variants, separately. This is schematised in Table 4.

We can fit a learning algorithm that is trained on two source sets: stable CVC and stable CC verbs. It takes CVC/CC verbs as test forms and assigns category membership

in either the stable CVC or the stable CC class for each CVC/CC verb. The algorithm will also assign a weight to its judgement. The result is a distance measure that quantifies where a given CVC/CC form is between stable CVC and stable CC verbs, in terms of similarity of form.

We use the Generalised Context Model (GCM) to determine similarity between the verb stems in the sample and stable CVC and CC verbs in the language. The GCM takes a target form and compares it to individual training forms in two categories (in our case). It calculates the similarity between the target and the individual training forms based on the edit distance:

$$\eta_{ij} = \exp\left(-d_{ij}/s\right)^p$$

In the equation above, η_{ij} represents the similarity between form i and form j, while d_{ij} is the edit distance between the two forms. s and p are free parameters, here set to $s = 0.3$ and $p = 1$. The parameter s determines how quickly the similarity decreases as the distance between the forms increases. When p is set to 1, as here, the similarity function is an exponential, rather than a Gaussian function of the edit distance.

The GCM calculates a summed distance of the target form to the category. The overall similarity S_{iC_J} of a test form i to a set C_J is calculated by summing the similarity η_{ij} of each member j of class C_J to the test form i, and dividing by the summed similarity η_{ij} of each member k of class C_K (the class of all stored forms) to the test form i. If all stored forms are grouped in two sets, similarity to one group is complementary with similarity to the other group. Therefore, if a verb's similarity to stable CVC verbs is 0.4 then its similarity to stable CC verbs is 0.6, as the target verbs are either stable CVC or stable CC. This calculation is summarized in the following equation.

$$S_{iC_J} = \frac{\sum_{j \in C_J} \eta_{ij}}{\sum_{k \in C_K} \eta_{ik}}$$

We adapted the GCM from the framework of Nosofsky (1990) to compare word forms in the R language (R Core Team, 2016). This algorithm has been widely and successfully used in linguistic categorisation tasks (Krott et al., 2001; Albright & Hayes, 2003; Dawdy-Hesterberg & Pierrehumbert, 2014).

For this specific categorisation problem, we create a training class of stable CVC and CC verbs using a list of 3PL.IND forms in the Hungarian Webcorpus with a token frequency of 10 or higher. We exclude the 111 potential CVC/CC verbs involved in creating our target sample. We exclude a small set of suppletive forms, as well as forms that can be seen both as complex forms or separate lexical entries (such as [bɛ-fyːt-ɛnɛk] 'heat-3PL.IND.PFV', cf. [fyːt/ɛnɛk]) 'heat-3PL.IND'), and forms that end in the irregular derivational suffix [iːt] (such as [lɒpiːt-ɒnɒk] 'squash-3PL.IND').

This approach results in a combined training set of 5916 3PL.IND forms, 555 CC, 5361 CVC. (The Webcorpus frequency dictionary has 6423 3PL.IND forms with a token frequency of 10 or higher. There are 207136 verb types in the frequency dictionary in total with a token frequency of 10 or higher.)

We create two test sets. Both are based on the 39 CVC/CC verbs in the sample. One

set consists of the CVC forms of the 3PL.IND. The other set consists of the CC forms.

Take the example of [hɒjlik] 'bend-3SG.IND'. We can take 3PL.IND forms and then calculate the overall similarity of the CVC or the CC form of the verb to stable CVC and CC verbs. The CVC form of 'bend-3PL.IND' is [hɒjolnɒk]. Its similarity is 0.593 to stable CVC 3PL.IND forms (e.g. [sɛrɛpɛlnɛk] 'act-3PL.IND'; [dolgoznɒk] 'work-3PL.IND') versus stable CC forms (e.g. [tɛrmɛstɛnɛk] 'grow-3PL.IND'; [hɒlːɒtsːɒnɒk] 'be audible-3PL.IND'). The CC form is [hɒjlɒnɒk]. Its similarity is to stable CVC versus CC is 0.589.

The two values tilt towards the more populous stable CVC class. Overall, the similarity weight based on CVC [hɒjolnɒk] is higher than the weight based on CC [hɒjlɒnɒk]. If we interpret these values across all 39 verbs and compare them to the odds of the suffixed forms, we can determine which measure of similarity is a better predictor of the ratio of their CVC/CC forms.

We fit the GCM using both simple edit distance and distance based on segmental similarity between the forms. We found that the former yields better predictions and report calculations based on edit distance in the paper.

Within-paradigm predictors We use the lemma frequency of the verb and the token frequency of the 3SG.IND in the Webcorpus to approximate the 'strength' of the stem / base form. In order to assess the strength of the inflected forms in the paradigm, we create another predictor of morphological behaviour: the *number of existing CVC or CC suffixed forms for each verb*. This ranges between 2 (our minimum threshold) and 12 (an attested CVC *and* CC form for each six suffixed form), as it provides an approximation of the diversity of the verb form's CVC/CC variability across suffixes.

Word-specific predictors For each verb, we calculate syllable count by counting the vowels in the stem, that is, the 3SG.IND minus the '-ik' suffix. We hand-annotate verbs to determine whether they are compounds and whether they are transitive or intransitive. We use the Webcorpus to determine whether the verb stem is attested in itself – noting whether or not the verb has a free stem. Again, we define the stem as either the 3SG.IND without the '-ik' suffix, or else, if there is a recognisable derivational suffix present (as in [dohaːŋ-z-ik] 'smoke-3SG.IND', from [dohaːŋ] 'tobacco'), the nominal stem preceding this suffix.

4 Modelling

We use the R statistical environment for our analysis (R Core Team, 2016) and create plots using the `effects` package (Fox et al., 2009) and `ggplot` (Wickham, 2009).

Twelve specific predictors are defined based on sections 2.2 and 3.2: *sonority, homorganicity, the* 3PL.IND *form's similarity to stable CVC verbs, its similarity to stable CC verbs, frequency of* 3SG.IND*, lemma frequency of verb, type frequency of attested C/V-initial suffixed forms, whether the form is a compound, whether a free base is attested, whether the verb is transitive, and length of base in syllables.*

Our aim was to compare these predictors in a model of CVC/CC variation, effectively predicting the log odds of CVC and CC forms of the thirty-nine verbs with the five suffixes.

The main problem with an exploratory study of this kind is that the factors which determine variation are correlated. For instance, the token frequency of the 3SG.IND and

the lemma frequency of the verb will be very similar across verbs (in our data, r = 0.66).

To assess the extent of multicollinearity across predictors, we fit a mixed-effects logistic regression model predicting the odds ratio of CVC/CC, using all the predictors defined in Section 3, centralised, with the addition of a stem and a suffix random intercept, using the LME4 package in R (Bates et al., 2015). The kappa coefficient of this model is 6.74, which indicates moderate collinearity between predictors (see Baayen 2008).

This means that we cannot immediately co-opt the method used by Janda et al. (2010) and test predictor strength in a generalised linear mixed-effects model. While mixed-effects models are robust at handling nested data with predictor collinearity (Gelman & Hill, 2007; Jaeger, 2008), correlations across predictors can undermine modelling assumptions when such a diverse set of predictors is used. While top-down stepwise model selection would allow us to remove collinear variables, this method is criticised for either amplifying variation or, worse, generating significant trends out of noise (see Flom & Cassell 2007).

In order to further justify the choice of predictor variables in our final model, we fit a regression tree with random effects on the data, with all our predictors, using the `REEMtree` package (Sela & Simonoff, 2011). `REEMtree` fits a large number of regression trees on the data using various combinations of the predictor variables. We can extract the importance of the predictor variables across these iterations to gain a sense of how robust they are in predicting the outcome. We fit the model with the twelve predictors, along with a verb form and a suffix random intercept, using one hundred thousand iterations.

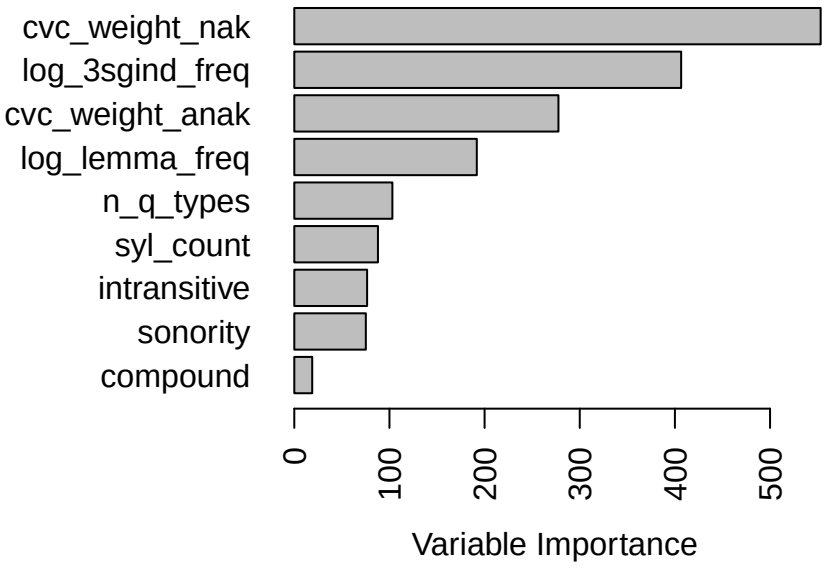

Figure 1: Variable importance in the regression trees

Figure 1 shows the most important predictor variables across all iterations. The most important predictor is the across-paradigm similarity of the verb's CVC 3PL.IND form to stable CVC or CC verbs (*cvc_weight_nak*). This is the [aːrɒmolnɒk] form in Table 4.

It is followed by the token frequency of the 3SG.IND (*log_3sgind_freq*), the similarity of the verb's CC 3PL.IND form to stable CVC or CC verbs (the [aːrɒmlɒnɒk] form in Table 4) (*cvc_weight_anak*), and the verb's lemma frequency (*log_lemma_freq*). The relevance of the remaining the predictors is diminished: the number of quasi-analytic suffix types attested (*n_q_types*), the verb's syllable count (*syl_count*), whether the verb is intransitive, the sonority of the consonant cluster, and whether the verb is a compound appear to not contribute much. Whether the verb has a free stem and whether the consonant cluster is homorganic have very little relevance.

We next fit a second mixed-effects logistic regression model on the dataset. The model predicts the log odds ratio of the CVC over the CC variant of the suffixed forms. Again, the counts of these variants are not independent – they are grouped under both the 39 verb stems and the 5 suffixes. The model accounts for the lack of independence by containing a random intercept for verb and for suffix. This approach largely follows Janda et al. (2010).

In selecting the predictors for this model, we relied on variable importance in the regression trees to build a model bottom-up, starting with the most important variable, and adding subsequent ones.

Model selection was informed by goodness of fit and the conceptual framework of the study. To select for goodness of fit we relied on the variance inflation factor, the Akaike Information Criterion, and model comparison using ANOVA in model selection. We included relevant random slopes in the final model. In terms of the conceptual framework our aim was to investigate *phonotactic*, *across-*, and *within-paradigm*, and *word-specific* attractors in determining CVC/CC variation. To consider these attractor types in a unified framework, we chose the most relevant predictor from each category to include in our model, irrespective of its stand-alone robustness.

This means that some of the predictors explain little variation in the data, but such a non-parsimonious model is not problematic given that the collinearity issues have been addressed.

5 Results

Our final model predicts CVC/CC variation based on four predictors. These are the across-paradigm similarity of the verb's CVC 3PL.IND form to stable CVC or CC verbs; the token frequency of the 3SG.IND; whether the verb stem is mono- or pollysyllabic; and the sonority slope of the consonant cluster. This model captures the relative importance of our attractor groups: phonotactic, across-paradigm, within-paradigm, and word-specific attractors. Not all of these are robust predictors of CVC/CC variation.

The marginal r^2 of the model is 0.28, the conditional r^2 is 0.74 (following Nakagawa & Schielzeth 2013). The kappa coefficient of this model with all predictors centered is 2.9, a considerable improvement on the model with all predictors (6.74).

The fixed effect estimates of the model can be seen in Table 5. Given the exploratory character of the analysis, we do not calculate p values for the estimates. It is, however, clear that the explanatory power of the predictors varies greatly.

The model intercept is not very meaningful in and of itself – it describes a word with a raw frequency of 1, for instance. Across-paradigm similarity (*cvc_weight_nak*) and word length (whether the word is mono- or *polysyllabic*) are robust predictors of CVC/CC variation, whereas phonotactics, specifically *sonority*, and the logged frequency of the

	Estimate	Std. Error	z value
(Intercept)	-5.43	1.66	-3.27
cvc_weight_nak	4.10	1.66	2.47
log_3sgind_freq	-0.20	0.21	-0.98
polysyllabicTRUE	2.74	0.88	3.11
sonority	-0.15	0.18	-0.87

Table 5: Estimated effects and standard errors, logistic model

3SG.IND (*log_3sgind_freq*) are not very relevant. A number of predictors are missing from the model. These either raise collinearity issues, have low explanatory value (as shown by the regression tree analysis), or are conceptually less interesting. For example, whether the word is a *compound* is a word-specific predictor of variable behaviour. As such, it is strongly correlated with and is far less relevant than word length.

5.1 Relevance of predictor groups

Using groupings outlined in sections 2.2 and 3.2, we now look at each predictor in the regression model.

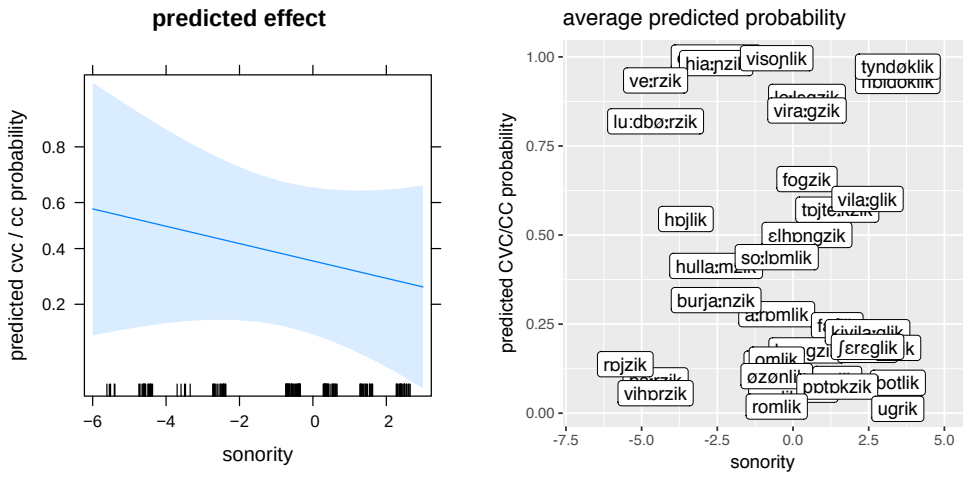

Figure 2: Sonority and CVC/CC probability across stems in the model

Phonotactic predictors Figure 2 shows the overall effect of sonority on the probability of a CVC form (to a CC form) in the model, with estimated standard error (left) and average predictions for verb forms (right).

Sonority (x axis) has a negative relationship with the ratio of CVC versus CC forms (y axis). Sonority slopes are never drastically rising in the clusters in the sample – sequences like [tr] or [kl] are not typical of the CVC/CC class. High values of sonority (that is, sonority difference) refer to clusters with a steeply raising sonority slope, like [tl] or [gr]. Low values of sonority refer to clusters with a less steep or a negative slope, like [ml] or [rz], respectively. These latter clusters are broken up more often.

The direction of the sonority effect is contrary to our expectations outlined in Section 2: clusters of rising sonority are *less* likely, rather than more likely, to be broken up.

At the same time, the estimate has a large amount of error, which means that in this model sonority has very little effect. We can add that the other phonotactic predictor, *homorganicity*, has not been pre-selected as a relevant predictor.

This indicates that the role of phonotactics in our model of CVC/CC variation is negligible. If lexical attractors operate on variable CVC/CC forms then their locus must lie elsewhere.

Across-paradigm predictors The predictor *cvc weight 'nak'* in Table 5 is a specific measure of similarity. It expresses the distance of the 3PL.IND form to stable CVC verbs, as compared to stable CC verbs. When it combines with CVC/CC verbs, the 3PL.IND suffix is [-nɒk] with CVC forms (e.g. [aːrɒmol-nɒk], 'flow-3PL.IND') and [-ɒnɒk] with CC forms (e.g. [aːrɒml-ɒnɒk], 'flow-3PL.IND'). We built two models of similarity, one based on the CVC form and one based on the CC form. These each provided a measure of similarity.

The measure that remains relevant in this model is the one based on the CVC rather than CC form of the CVC/CC verb: *cvc weight 'nak'* (as opposed to *cvc weight 'anak'*). The similarity effect is shown by Figure 3, which displays the overall effect of similarity to stable CVC verbs – as opposed to stable CC verbs – on the probability of a CVC form (versus a CC form) in the model, with estimated standard error (left) and average predictions for verb forms (right).

The similarity effect is in line with our expectations outlined in Section 2: a variable verb's similarity to stable verb classes affects its pattern of variation.

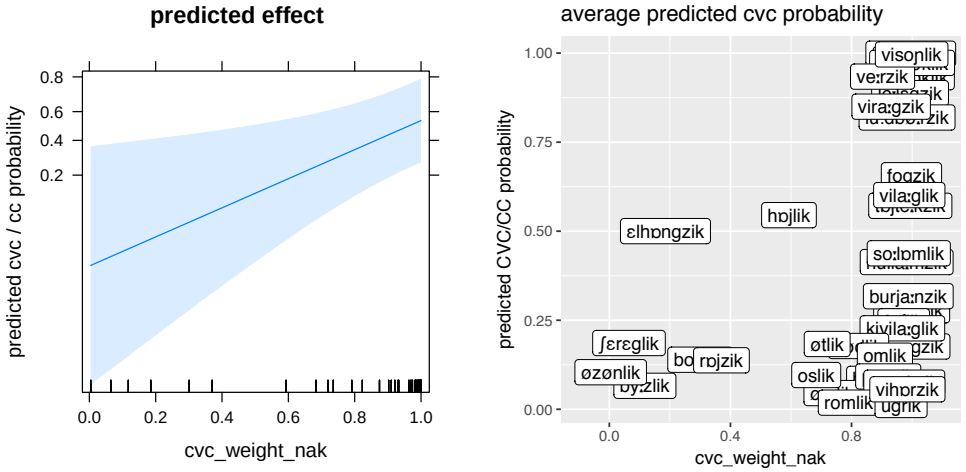

Figure 3: Similarity and CVC/CC probability across stems in the model

A variable verb's similarity to stable verb classes will affect variation. This is a robust finding of the analogical learning literature (Bybee & Slobin, 1982; Rumelhart & McClelland, 1986; Skousen, 1989). Our result corroborates this. It is noteworthy that the variable process in this case is essentially low-level allomorph selection resulting in a CVC or CC form. Similarity has been highlighted as a relevant factor in such low-level processes, as in Dutch linking vowels (Krott et al., 2001) or English hiatus resolution (Soskuthy, 2013).

At the same time, the analogical learning literature generally posits the base form as the target of analogical processes. The base form, the 3SG.IND, of CVC/CC verbs has a

V-initial suffix [-ik], and, as a consequence, is almost always CC. (Only six stems in our sample are attested with a CVC 3SG.IND, all with low token frequency.) It is the various other C-initial suffixes that result in a CVC form. In our model of similarity, the CVC form is a better target than the CC form – this strongly suggests that similarity operates across *a range of inflected forms*.

This, in turn, indicates that some inflected forms are available in the mental lexicon, and exert an influence on variable behaviour in verbal inflection (see Stemberger & MacWhinney 1986; Alegre & Gordon 1999; Lindsay et al. 2012).

In terms of attractor biases, the model suggests that patterns of similarity across verbs are important in predicting CVC/CC variation, whereas more general phonotactic patterns are not. While general phonotactics and word similarity are not statistically independent, the model supports the importance of the latter over the former.

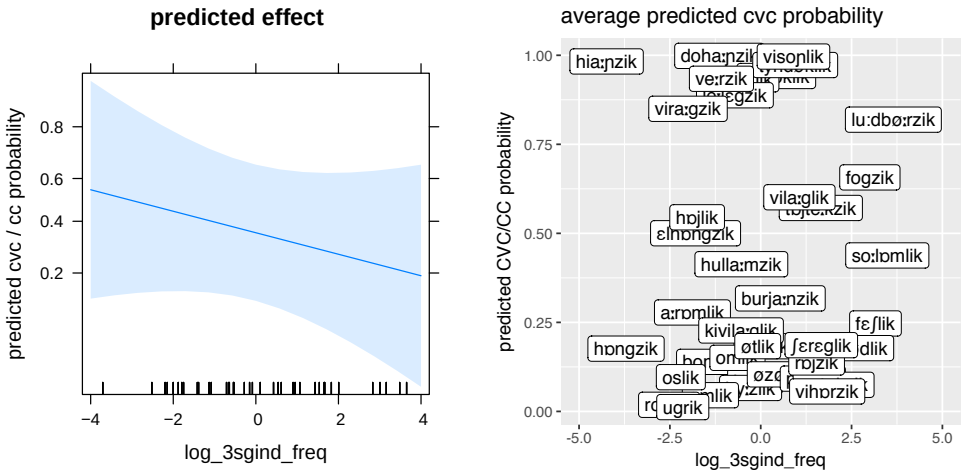

Figure 4: Token frequency of the 3SG.IND and CVC/CC probability across stems in the model

Within-paradigm predictors Figure 4 shows that there is a negative relationship between the frequency of the 3SG.IND form and the verb's CVC/CC ratio. This result is in line with the expectations in Section 2: more frequent verbs are more likely to be realised with CC variants. However, the error of the estimate is very large, which indicates that token frequency has negligible influence on CVC/CC variation.

The lemma frequency of the verb and the type frequency of suffixed variants are not pre-selected as relevant predictors in the model.

The lack of a type frequency effect indicates that the overall similarity of a form to stable CVC versus CC verbs in the lexical space is more relevant than type frequencies of its suffixed variants in determining variable behaviour.

The lack of a token frequency effect runs counter to our expectation that high token frequency forms to resist a shift to a majority pattern. This frequency effect is well documented for a wide range of types of morphological variation (see Bybee 1995). However, there is no obvious majority pattern within the set of CVC/CC verbs: across-paradigm similarity to stable verb classes is far more important than patterns of behaviour within the set. The lack of a clear majority pattern is an important aspect of this morphological variation, one that we return to in the discussion.

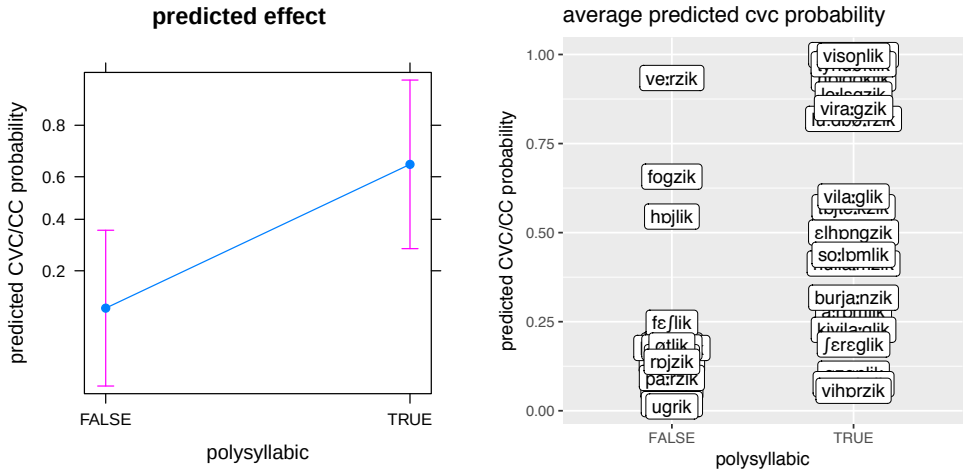

Figure 5: Stem length and CVC/CC probability across stems in the model

Word-specific predictors Verbs with longer stems prefer the CVC variant (Figure 5). (This result is the reverse of our expectation in Section 2.) This pattern indicates that longer forms tend to be even longer, contrary to broad correlations between length and frequency that also hold in Hungarian (Zipf, 1935; Németh & Zainkó, 2002).

A possible explanation of the positive effect of length is that the CVC variant preserves the lexically specified vowel in the cluster, and, as a consequence, the stem remains easier to identify. Identifiability is more important for verbs with a lower token frequency that are harder to activate in the first place. Alternatively, we see here a masked effect of morphology – longer verbs are more likely to have stems that occur across the paradigm and, as a result, are more robustly represented. At the same time, other word-specific predictors (e.g., whether or not the verb is a compound, or whether it has a free stem) do not contribute meaningfully to the model.

5.2 Implications for lexical attractors

Our aim was to describe a set of variable verbs in Hungarian and identify a very wide range of possible lexical attractors that can influence variation, ranging from very general ones – such as phonotactic patterns – to very narrow ones – such as the verb's paradigm or word-specific information. We used algorithmic learning, cluster analysis and regression modelling to identify strong trends in our sample within this range of complex and collinear attractors.

Results indicate that variation is governed by word-specific and across-verb attractors. Polysyllabic verb stems are more likely to have CVC variants over CC variants than monosyllabic ones in the sample. Variable verbs that are similar to stable CVC verbs are more likely to behave like stable CVC verbs. The basis for this similarity is, at least partially, the set of inflected forms.

We need to be mindful of existing caveats.

One issue is the model's operationalisation: Online language use, though less formal than writing and a rich data source, is not a stand-in for the ambient language. Our restrictions result in a principled dataset, but are the results of a series of analytical decisions. The way we operationalise lexical attractors follows the pre-existing psycholinguistics and corpus linguistics literature, but remains highly subjective. Our

similarity-based model, in particular, makes strong assumptions on lexical relations.

The other issue is the model's interpretation: the attractor types we discuss do not constitute disjunct sets. For example, it is trivially true that the *across-paradigm* similarity of forms will be affected by general phonotactics, since all forms conform to the phonotactics of the language. As a consequence, when we argue that the phonotactics have a negligible effect on variation, this should be interpreted in relation to the other factors we consider, such as similarity of form.

Still, our model presents a quantitative operationalisation of a complex problem in morphophonological variation, provides data from a non-Indo-European language, and offers insights into the mechanics of lexical variation. We have shown that, in our Hungarian data, similarity across verbs is more important in shaping CVC/CC variation than either broad phonotactic patterns or effects of the verb's frequency.

6 Discussion

We use concatenative terminology to describe CVC/CC variation in Hungarian inflectional morphology. A concatenative model has tremendous difficulties describing wider chunks of Hungarian morphology (Rebrus & Törkenczy, 2011; Rácz & Rebrus, 2012). However, it supplies a convenient metaphor to provide a simple description of CVC/CC variation. This metaphor has been well-used by linguists working with morphology since at least the time of Bloomfield (1926).

However, the corpus-based model proposed in this paper is not compatible with a genuine concatenation approach. The model indicates the presence of lexical attractors underlying variation. These attractors range from very broad ones, such as stochastic phonotactic patterns in the language, to more specific ones, such as the inflection behaviour of similar verbs.

The results have clear consequences for the description of the mental lexicon. The behaviour of variable CVC/CC verbs cannot be specified by rote, nor can it be accounted for using abstract generalisations. Instead, CVC/CC verbs point towards the relevance of the relationships between individual forms stored in the lexicon.

Hungarian inflectional variation, then, is compatible with a theory of morphology that posits a rich storage of forms, with connections between these forms (Bybee, 1985; Blevins, 2006). Such a theory must emphasise relationships between and within paradigms, driven by analogy and similarity (Milin et al., 2009; Dawdy-Hesterberg & Pierrehumbert, 2014). In this respect, our work shows close parallels with similar research on languages with rich inflectional paradigms, such as Finnish (Kidd & Kirjavainen, 2011), Serbian (Mirković et al., 2011), or Polish (Dąbrowska, 2008).

The results lend themselves to four primary conclusions. First, they give credit to a corpus-based heuristic which can be used to model community-level variation. Second, they provide additional support for a rich memory model of the mental lexicon. Third, they allow us to speculate on the amount of information available in the mental lexicon, a question subject to widespread debate (see e.g. Bresnan et al. 2007). Fourth, they inform our understanding of morphological variation in Hungarian.

In our model, the phonological shape of the word form, along with patterns of similarity across words in the lexicon, influences variation. These aspects appear to be stronger than simple measures of the word's frequency or its base form.

Our model set sheds light on specific aspects of CVC/CC variation. CVC/CC vari-

ation is unlike cases of morphological variation in which a majority productive pattern competes with multiple minority patterns. This variation is not productive, so it is predominantly affected by broader attractors (word shape and similarity to stable lexical classes) rather than competition across forms in its own class. This means that this exploratory analysis allows us to make predictions for types of variation in the future: while productive patterns will be affected by frequency, stable patterns will be shaped primarily by similarity.

These results can only be generalised to a limited extent. For instance, an entirely tree-based clustering analysis would have attributed more importance to form frequency. What they allow us to do, however, is generate hypotheses on the structure and influence of lexical structure on variation. These hypotheses can, in turn, lead to stronger assumptions in the design and analysis of subsequent corpus studies and psycholinguistics experiments.

References

Albright, Adam. 2009. Modeling analogy as probabilistic grammar. *Analogy in grammar* 185–213.

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

Alegre, Maria & Peter Gordon. 1999. Rule-based versus associative processes in derivational morphology. *Brain and Language* 68(1). 347–354.

Baayen, R Harald. 2007. Storage and computation in the mental lexicon. *The mental lexicon: Core perspectives* 81–104.

Baayen, R Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using r*. Cambridge University Press.

Baayen, R Harald, Ton Dijkstra & Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37(1). 94–117.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. doi: 10.18637/jss.v067.i01.

Bickel, Balthasar, Bernard Comrie & Martin Haspelmath. 2008. The Leipzig Glossing Rules. Conventions for interlinear morpheme by morpheme glosses. *Revised version of February* .

Blevins, James P. 2001. Paradigmatic derivation. *Transactions of the Philological Society* 99(2). 211–222.

Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42(3). 531–573.

Bloomfield, Leonard. 1926. A set of postulates for the science of language. *Language* 2(3). 153–164.

Booij, Geert. 1999. Lexical storage and regular processes. *Behavioral and Brain Sciences* 22(6). 1016–1016.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, R. Harald Baayen et al. 2007. Predicting the dative alternation. *Cognitive foundations of interpretation* 69–94.

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and cognitive processes* 10(5). 425–455.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*, vol. 9. Amsterdam: John Benjamins Publishing.

Bybee, Joan L & Dan I Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language* 265–289.

Carreiras, Manuel, Andrea Mechelli & Cathy J Price. 2006. Effect of word and syllable frequency on activation during lexical decision and reading aloud. *Human brain mapping* 27(12). 963–972.

Christiansen, Morten H & Simon Kirby. 2003. Language evolution: Consensus and controversies. *Trends in cognitive sciences* 7(7). 300–307.

Clements, George N. 1990. The role of the sonority cycle in core syllabification. *Papers in laboratory phonology* 1. 283–333.

Colé, Pascale, Juan Segui & Marcus Taft. 1997. Words and morphemes as units for lexical access. *Journal of Memory and Language* 37(3). 312–330.

Cuskley, Christine F, Martina Pugliese, Claudio Castellano, Francesca Colaiori, Vittorio Loreto & Francesca Tria. 2014. Internal and external dynamics in language: evidence from verb regularity in a historical corpus of English. *PloS one* 9(8). e102882.

Dąbrowska, Ewa. 2008. The effects of frequency and neighbourhood density on adult speakers' productivity with polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language* 58(4). 931–951.

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

Dominguez, Alberto, Maira Alija, Javier Rodriguez-Ferreiro & Fernando Cuetos. 2010. The contribution of prefixes to morphological processing of Spanish words. *European Journal of Cognitive Psychology* 22(4). 569–595.

Duñabeitia, Jon Andoni, Sachiko Kinoshita, Manuel Carreiras & Dennis Norris. 2011. Is morpho-orthographic decomposition purely orthographic? Evidence from masked priming in the same–different task. *Language and Cognitive Processes* 26(4-6). 509–529.

Flom, Peter L & David L Cassell. 2007. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In *NorthEast SAS Users Group Inc 20th Annual Conference: 11-14th November 2007; Baltimore, Maryland*, NorthEast SAS Users Group.

Fox, John, Jangman Hong et al. 2009. Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software* 32(1). 1–24.

Frisch, Stefan A, Janet B Pierrehumbert & Michael B Broe. 2004. Similarity avoidance and the OCP. *Natural Language & Linguistic Theory* 22(1). 179–228.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel-hierarchical models*, vol. 1. New York: Cambridge University Press New York.

Gonnerman, Laura M., Mark S. Seidenberg & Elaine S. Andersen. 2007. Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology: General* 136(2). 323–345.

Grainger, Jonathan. 1990. Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of memory and language* 29(2). 228–244.

Grainger, Jonathan, Pascale Colé & Juan Segui. 1991. Masked morphological priming in visual word recognition. *Journal of Memory and Language* 30(3). 370–384.

Hahn, Ulrike & Ramin Charles Nakisa. 2000. German inflection: single route or dual route? *Cognitive Psychology* 41(4). 313–360.

Halácsy, Péter, András Kornai & Csaba Oravecz. 2007. HunPos: an open source trigram tagger. In *Proceedings of the 45th annual meeting of the acl on interactive poster and demonstration sessions*, 209–212. Association for Computational Linguistics.

Harris, John. 2006. The phonology of being understood: Further arguments against sonority. *Lingua* 116(10). 1483–1494.

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

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry* 39(3). 379–440.

Jaeger, T Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language* 59(4). 434–446.

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(1). 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.

Kidd, Evan & Minna Kirjavainen. 2011. Investigating the contribution of procedural and declarative memory to the acquisition of past tense morphology: Evidence from finnish. *Language and Cognitive Processes* 26(4-6). 794–829.

Kirby, Simon, Monica Tamariz, Hannah Cornish & Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition* 141. 87–102.

Krott, Andrea, R Harald Baayen & Robert Schreuder. 2001. Analogy in morphology: modeling the choice of linking morphemes in dutch. *Linguistics* 39(1; ISSU 371). 51–94.

Labov, William. 2011. *Principles of linguistic change, cognitive and cultural factors*, vol. 3. Oxford: John Wiley & Sons.

Lindsay, Shane, Leanne M. Sedin & M. Gareth Gaskell. 2012. Acquiring novel words and their past tenses: Evidence from lexical effects on phonetic categorisation. *Journal of Memory and Language* 66(1). 210–225.

Lukács, Ágnes, Péter Rebrus & Miklós Törkenczy. 2010. Defective verbal paradigms in Hungarian—description and experimental study. In *Proceedings of the British Academy*, vol. 163, 85.

Milin, Petar, Dušica Filipović Djurđević & Fermín Moscoso del Prado Martín. 2009. The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Journal of Memory and Language* 60(1). 50–64.

Milnor, John. 1985. On the concept of attractor. In *The theory of chaotic attractors*, 243–264. Springer.

Mirković, Jelena, Mark S Seidenberg & Marc F Joanisse. 2011. Rules versus statistics: Insights from a highly inflected language. *Cognitive science* 35(4). 638–681.

Myers, James & Yingshing Li. 2009. Lexical frequency effects in Taiwan Southern Min syllable contraction. *Journal of Phonetics* 37(2). 212–230.

Nakagawa, Shinichi & Holger Schielzeth. 2013. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2). 133–142.

Németh, Géza & Csaba Zainkó. 2002. Multilingual statistical text analysis, Zipf's law and Hungarian speech generation. *Acta Linguistica Hungarica* 49(3-4). 385–405.

Nosofsky, Robert M. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology* 34(4). 393–418.

Pater, Joe. 2000. Non-uniformity in english secondary stress: the role of ranked and lexically specific constraints. *Phonology* 17(2). 237–274.

Paul, Hermann. 1880/1995. *Prinzipien der sprachgeschichte [principles of the history of language]*, vol. 6. Berlin: Walter de Gruyter.

Pierrehumbert, Janet B. 2012. The dynamic lexicon. In Abigail Cohn, Cécile Fougeron & Marie Huffman (eds.), *The Oxford handbook of laboratory phonology*, 173–83. Oxford: Oxford University Press.

Pierrehumbert, Janet B. 2016. Phonological representation: beyond abstract versus episodic. *Annual Reviews* .

R Core Team. 2016. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Rácz, Péter, Viktória Papp & Jennifer B. Hay. 2016. Chapter 25. Frequency and corpora. In Andrew Hippisley & Gregory Stump (eds.), *The Cambridge Handbook of Morphology*, Cambridge University Press.

Rácz, Péter, Janet B Pierrehumbert, Jennifer B Hay & Viktória Papp. 2015. Morphological emergence. In *The handbook of language emergence*, 123–146. Wiley Online Library.

Rácz, Péter & Péter Rebrus. 2012. Variation in the possessive allomorphy of Hungarian. In *Current issues in morphological theory:(ir) regularity, analogy and frequency. selected papers from the 14th international morphology meeting, budapest, 13 16 may 2010*, vol. 322, 51. John Benjamins Publishing, Amsterdam, NL.

Rebrus, Péter. 2000. Morfofonológiai jelenségek [morphophonological phenomena]. In Ferenc Kiefer (ed.), *Strukturális magyar nyelvtan: Morfológia [Structural Hungarian Grammar: Morphology]*, vol. 3, Budapest: Akadémiai Kiadó.

Rebrus, Péter & Miklós Törkenczy. 2011. Paradigmatic variation in Hungarian. In *Approaches to hungarian. papers from the 2009 debrecen conference*, vol. 12, 135–161.

Rumelhart, David E. & James L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel distributed processing (volume 2): Psychological and biological models*, 216–271. Cambridge, MA: MIT Press.

Sela, Rebecca J & JS Simonoff. 2011. Reemtree: Regression trees with random effects. *R package version 0.90* 3. 741–749.

Siptár, Péter & Miklós Törkenczy. 2000. *The phonology of Hungarian*. Oxford University Press.

Skousen, Royal. 1989. *Analogical modeling of language*. Springer Science & Business Media.

Soskuthy, Marton. 2013. Analogy in the emergence of intrusive-r in english. *English Language & Linguistics* 17(1). 55–84.

Stemberger, Joseph Paul & Brian MacWhinney. 1986. Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition* 14(1). 17–26.

Tagliamonte, Sali A & R Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language variation and change* 24(2). 135–178.

Trón, Viktor, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda & Eszter Simon. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of 5th international conference on language resources and evaluation (lrec'06)*, Genova, Italy.

Trón, Viktor, András Kornai, György Gyepesi, László Németh, Péter Halácsy & Dániel Varga. 2005. Hunmorph: open source word analysis. In *Proceedings of the workshop on software*, 77–85. Association for Computational Linguistics.

Wang, William S Y. 1969. Competing changes as a cause of residue. *Language* 9–25.

Wickham, Hadley. 2009. *ggplot2: elegant graphics for data analysis*. Springer-Verlag New York. http://ggplot2.org.

Zipf, George Kingsley. 1935. *The psycho-biology of language.* Houghton, Mifflin.

Appendix

The list of stems in the data in 3SG.INDEF is, in orthographic form, contained in Table 6. We consider both the CVC and the CC forms of the 3SG.INDEF – here, only the CC form is printed. Table 7 lists, for each form, the most frequent CVC form in the corpus. Table 8 lists, for each form, the most frequent CC form in the corpus.

A note on the orthography: á:[aː], é:[eː], a:[ɒ], e:[ɛ], ö:[ø], ü:[y], sz:[s], s:[ʃ], zs:[ʒ], ny:[ɲ]

form_3sg	ikfreq	cvc_freq_total	cc_freq_total	cvc_cc_odds	gloss_stem
áramlik	2773	177	968	1.18	flow
bomlik	1746	16	960	1.02	decompose
botlik	493	9	343	1.03	toddle
burjánzik	247	32	117	1.28	burgeon
bűzlik	560	1	107	1.02	stink
döglik	381	29	302	1.10	perish
dohányzik	1314	2253	3	752.33	smoke
elhangzik	2545	192	1061	1.18	be voiced
feslik	18	3	19	1.21	peel
fogzik	21	6	3	3.33	tooth
fuldoklik	276	220	21	11.52	choke
hajlik	2439	919	1363	1.67	bend
haldoklik	821	161	18	10.00	die
hangzik	17253	474	5550	1.09	sound
hiányzik	29456	11538	346	34.35	be missing
hullámzik	722	126	218	1.58	wave
kivilâglik	728	2	20	1.15	light up
ködlik	25	1	16	1.12	fog
lélegzik	862	1084	176	7.16	breathe
lúdbőrzik	11	7	2	5.00	get goosebumps
omlik	803	59	654	1.09	collapse
ömlik	1693	11	610	1.02	pour
oszlik	3818	27	1600	1.02	decompose
ötlik	460	5	118	1.05	occur
özönlik	229	7	716	1.01	surge
párzik	101	15	132	1.12	mate
patakzik	57	1	62	1.03	efflux
rajzik	91	6	183	1.04	swarm
romlik	5257	19	2480	1.01	degrade
sereglik	80	12	294	1.04	rally
szólamlik	13	1	3	1.67	voice
tajtékzik	81	12	15	1.87	tantrum
tündöklik	164	482	21	24.00	shine
ugrik	3620	12	3769	1.00	jump
vérzik	1248	479	24	21.00	bleed
viharzik	68	1	34	1.06	storm
világlik	146	20	21	2.00	lighten
virágzik	3133	814	250	4.26	bloom
viszonylik	172	2319	13	179.46	relate

Table 6: List of verbs. 'Form 3sg' is the 3SG.IND form, 'ikfreq' is the frequency of the 3SG.IND, 'cvc freq total' is the summed frequency of CVC forms, 'cc freq total' is the summed frequency of CC forms, 'cvc cc odds' is the odds ratio, 'gloss stem' is the gloss of the stem.

form_3sg	best_cvc_form	freq_cvc_form	gloss_suffix	gloss_stem
áramlik	áramolni	75	inf	flow
bomlik	bomolni	6	inf	decompose
botlik	botolnak	3	3pl.ind	toddle
burjánzik	burjánoztak	16	3pl.past	burgeon
bűzlik	bűzölni	1	inf	stink
döglik	dögölni	18	inf	perish
dohányzik	dohányozni	1493	inf	smoke
elhangzik	elhangoztak	89	3pl.past	be voiced
feslik	feseltek	1	3pl.past	peel
fogzik	fogaznak	4	3pl.ind	tooth
fuldoklik	fuldokolni	130	inf	choke
hajlik	hajolni	400	inf	bend
haldoklik	haldokolnak	72	3pl.ind	die
hangzik	hangoztak	201	3pl.past	sound
hiányzik	hiányoznak	7012	3pl.ind	be missing
hullámzik	hullámoztak	59	3pl.past	wave
kivilághk	kivilágolnak	1	3pl.ind	light up
ködlik	ködölni	1	inf	fog
lélegzik	lélegezni	677	inf	breathe
lúdbőrzik	lúdbőrözni	4	inf	get goosebumps
omlik	omolnak	23	3pl.ind	collapse
ömlik	ömölni	4	inf	pour
oszlik	oszolni	11	inf	decompose
ötlik	ötölni	2	inf	occur
özönlik	özönölnek	4	3pl.ind	surge
párzik	pározni	14	inf	mate
patakzik	patakoznak	1	3pl.ind	efflux
rajzik	rajoznak	5	3pl.ind	swarm
romlik	romolni	9	inf	degrade
sereglik	seregeltek	4	3pl.past	rally
szólamlik	szólamolnak	1	3pl.ind	voice
tajtékzik	tajtékoznak	9	3pl.ind	tantrum
tündöklik	tündökölni	208	inf	shine
ugrik	ugornak	12	3pl.ind	jump
vérzik	vérezni	232	inf	bleed
viharzik	viharoznak	1	3pl.ind	storm
világlik	világolnak	8	3pl.ind	lighten
virágzik	virágoznak	378	3pl.ind	bloom
viszonylik	viszonyulnak	1339	3pl.ind	relate

Table 7: List of verbs with most frequent CVC form. 'Form 3sg' is the 3SG.IND form, 'best cvc form' is the most frequent CVC form of the verb, 'freq cvc form' is the frequency of this specific form, 'gloss suffix' is the gloss of the suffix, 'gloss stem' is the gloss of the stem.

form_3sg	best_cc_form	freq_cc_form	gloss_suffix	gloss_stem
áramlik	áramlanak	593	3pl.ind	flow
bomlik	bomlanak	710	3pl.ind	decompose
botlik	botlanak	168	3pl.ind	toddle
burjánzik	burjánzanak	76	3pl.ind	burgeon
bűzlik	bűzlenek	42	3pl.ind	stink
döglik	dögleni	136	inf	perish
dohányzik	dohányzani	2	inf	smoke
elhangzik	elhangzanak	937	3pl.ind	be voiced
feslik	feslenek	9	3pl.ind	peel
fogzik	fogzani	2	inf	tooth
fuldoklik	fuldoklani	11	inf	choke
hajlik	hajlanak	916	3pl.ind	bend
haldoklik	haldoklanak	11	3pl.ind	die
hangzik	hangzottak	3019	3pl.past	sound
hiányzik	hiányzanak	326	3pl.ind	be missing
hullámzik	hullámzanak	87	3pl.ind	wave
kiviláglik	kiviláglanak	12	3pl.ind	light up
ködlik	ködlenek	10	3pl.ind	fog
lélegzik	lélegzeni	144	inf	breathe
lúdbőrzik	lúdbőrzenek	2	3pl.ind	get goosebumps
omlik	omlanak	256	3pl.ind	collapse
ömlik	ömlenek	253	3pl.ind	pour
oszlik	oszlanak	1132	3pl.ind	decompose
ötlik	ötlenek	55	3pl.ind	occur
özönlik	özönlöttek	320	3pl.past	surge
párzik	párzanak	88	3pl.ind	mate
patakzik	patakzottak	35	3pl.past	efflux
rajzik	rajzanak	160	3pl.ind	swarm
romlik	romlottak	979	3pl.past	degrade
sereglik	sereglettek	174	3pl.past	rally
szólamlik	szólamlani	2	inf	voice
tajtékzik	tajtékzottak	8	3pl.past	tantrum
tündöklik	tündöklenek	14	3pl.ind	shine
ugrik	ugrani	2151	inf	jump
vérzik	vérzenek	9	3pl.ind	bleed
viharzik	viharzottak	18	3pl.past	storm
világlik	világlottak	11	3pl.past	lighten
virágzik	virágzanak	187	3pl.ind	bloom
viszonylik	viszonylanak	13	3pl.ind	relate

Table 8: List of verbs with most frequent CC form. 'Form 3sg' is the 3SG.IND form, 'best cc form' is the most frequent CC form of the verb, 'freq cc form' is the frequency of this specific form, 'gloss suffix' is the gloss of the suffix, 'gloss stem' is the gloss of the stem.