Acta Linguistica Academica 70 (2023) 3, 397–409
DOI: 10.1556/2062.2023.00620

AKADÉMIAI KIADÓ

Sound patterns, frequency and predictability in inflection

ANDRÁS CSER[1,2]*

[1] Hungarian Research Centre for Linguistics, Hungary

[2] Pázmány Péter Catholic University, Hungary

Received: October 15, 2022 • Revised manuscript received: April 26, 2023 • Accepted: April 26, 2023
Published online: June 13, 2023

ABSTRACT

The paper investigates the relations between phonological form and information content within Latin verbal inflection from two interrelated points of view. It looks at conditional entropy relations within the present paradigm to see how these relate to the textual frequency of the individual forms; and it seeks to answer the question to what extent the phonological form of stems and endings has the potential to lead to ambiguity in morphological marking. The latter issue is approached from the angle of the information content that word forms taken in themselves have about their morphological status. The broader question of potential ambiguity is broken down into two separate questions: one concerns stems where intra-paradigmatic ambiguity would be possible; the other concerns stems that include phonological material that could itself be interpreted as a morphological marker. The absence of potential ambiguity in the first sense, and its severe restriction in the second sense is interpreted here as an emergent mechanism to enhance the information content of verb forms.

KEYWORDS

morphology, phonology, Latin, conditional entropy, inflection

1. INTRODUCTION

The relation between linguistic form and information content has, in a very general sense, always been central to the study of language. However, morphology is a field in which this

* Corresponding author. E-mail: cser.andras@btk.ppke.hu

issue has come to prominence over the past two decades to a degree not seen elsewhere. One of the most spectacular manifestations of this shift is the increased interest in the investigation of entropy in morphological systems, and more specifically of conditional entropy, i.e. the capacity to predict a certain form when another form is known (e.g. Ackermann, Blevins & Malouf 2009; Ackermann & Malouf 2013; Stump & Finkel 2013; Sims & Parker 2016 on inflectional morphology, Bonami & Strnadová 2019 extending the approach to derivation).

The present paper investigates the relations between phonological form and information content that obtain within Latin verbal inflection from two partly interrelated points of view. In 2, we look at conditional entropy relations that hold within the present paradigm (this being the most varied) and see how these relate to the textual frequency of the individual forms. While conditional entropy within Latin verbal inflection has been studied previously (most notably in Pellegrini 2020), its relation to frequency has not. We then turn to the question to what extent the phonological form of stems and endings has the potential to lead to ambiguity in morphological marking. We approach this issue from the angle of the information content that word forms taken in themselves have about their morphological status. The broader question of potential ambiguity is broken down into two separate questions: the one discussed in 3 concerns stems where intra-paradigmatic ambiguity would be possible; the other, discussed in 4, concerns stems that include phonological material that could itself be interpreted as a morphological marker (ending). The absence of potential ambiguity in the first sense, and its severe restriction in the second sense is interpreted here as an emergent mechanism to enhance the information content of verb forms.

Throughout the paper we use the terms stem and ending, but these are only convenient labels to denote phonological material that either stays consistently present throughout a morphologically well-defined set of forms or is changed only by productive phonological rules. Analyses of morphological exponence and of allomorphy within the Latin inflectional system have been presented by numerous authors (Matthews 1974; Lieber 1980; Aronoff 1994; Cser 2015, 2020, 124–152), partly in constituent-based frameworks. The present paper crucially assumes no strong constituency, or morphemic structure, and anything that is said using such terminology for convenience can easily be translated into a word-based framework.

In terms of time periods, the discussion is deliberately broad in that it is not restricted to Classical Latin in the narrow sense, but it is restricted to the native Latin period lasting until later Antiquity. Data coming from different periods will be duly pointed out.[1]

2. FORM FREQUENCY AND CONDITIONAL ENTROPY IN THE PRESENT PARADIGM

2.1. The frequency of verb forms according to person and number

We have counted the occurrences of the six present forms of 31 verbs in the Packard Humanities Institute database (latin.packhum.org). There were two criteria for the selection of the verbs: (i) they should be frequent enough to contribute meaningfully to any quantitative analysis; in practice this meant they should each have at least one form in the paradigm with a textual

[1]On questions of the periods in the history of Latin see Adamik (2015).

Table 1. The frequency of verb forms in the present paradigm (based on the forms of 31 verbs in the Packard Humanities Institute database)

1sg	6549
2sg	4326
3sg	20,581
1pl	2781
2pl	942
3pl	8079

frequency of at least 200 (in fact the smallest number was 232); (ii) they should present no more than a manually manageable amount of homography (hence the omission of some very frequent verbs such as *esse* 'be', *ire* 'go' or *velle* 'want').[2] The resulting numbers are given in Table 1.

Two generalisations are immediately apparent: (i) in any person there are more singular than plural forms, and (ii) in either number the amount of person forms decreases in the order 3>1>2. Obviously we are dealing here with a written corpus, albeit a large and varied one, where register and genre distort these proportions somewhat. Note, however, that the same ratios emerge e.g. from the Hungarian National Corpus (http://clara.nytud.hu/mnsz2-dev/, cf. Oravecz, Váradi & Sass 2014). Since we do not at present see a way to get closer to the realities of spoken Latin in these specific terms,[3] we shall be referring to these numbers when correlating frequency and conditional entropy at the end of the next section.

2.2. Conditional entropy

We calculated the unweighted conditional entropy of the forms in the active present paradigm of regular verbs. All the three restrictions are important and require at least a brief explanation – not least because they represent a major difference between the present paper and the most important one on the topic to date, Pellegrini (2020). We restrict ourselves to the present paradigm because it is the most varied morphophonologically and the least predictable on average. All other paradigms in the Latin verbal inflection are much more self-contained. In fact, with the partial exception of the perfect paradigm, they consist of forms that are fully predictable from any other form within the paradigm, thus conditional entropy is zero.

[2]The list of verbs included is the following: *accipere* 'receive', *adferre* 'take', *agere* 'drive', *amare* 'love', *cogere* 'force', *continere* 'contain', *credere* 'believe', *dare* 'give', *debere* 'must', *docere* 'teach', *efficere* 'effect', *existimare* 'existimate', *gaudere* 'rejoice', *gerere* 'carry', *habere* 'have', *jubere* 'order', *mittere* 'send', *movere* 'move', *negare* 'deny', *ponere* 'put', *premere* 'press', *quaerere* 'ask', *recipere* 'receive', *reddere* 'give back', *redire* 'go back', *stare* 'stand', *tenere* 'hold', *timere* 'fear', *tradere* 'hand over', *vetare* 'forbid', *videre* 'see'.

[3]We did a count of the forms of the verb *facere* 'do' in the comedies of Plautus, which consist largely of dialogues; the numbers found are 1sg: 68, 2sg: 82, 3sg: 79, 1pl: 6, 2pl: 16, 3pl: 35. Note that while generalisation (i) above still holds, second person forms outnumber first person forms, in the singular even third person forms; but 1 and 2pl forms are still the least frequent. While *facere* is a verb of high frequency, it is only one verb, and these numbers are still too small to be representative in any realistic sense.

The structure of the perfect paradigm is interesting and non-trivial (pace Pellegrini), but we will not be concerned with it in this paper.

Secondly, we restrict ourselves to regular verbs. This is because we are primariliy interested in the general structural aspects of entropy relations, but at a next stage in the research project irregular verbs should definitely be included, as indeed they are in Pellegrini (2020), where the resulting entropy relations are more fine-grained than ours.

Thirdly, we do no weighting in our calculations. While Pellegrini (2020) weights with the relative lexical frequency of the verb classes (conjugations), and this is obviously important to arrive at a realistic picture, it is not at all clear at this stage what exactly one should weight with. To simply take verb classes is certainly not enough, as the well-known English example shows: whereas regular verbs far outnumber irregular ones, if a verb in the present has the form *C(C)ing*, it has a far greater likelihood of being irregular than regular (*sing, ring, cling, sting* etc.). We take a Latin example too to illustrate this point. The two largest verb classes are *a*-stems[4] (first conjugation) and C-stems (the majority of the third conjugation); they account for roughly 44% and 34% of the number of verbs, respectively.[5] However, if the verb stem includes a consonant cluster consisting of a nasal and a stop in this order (the most frequent type of cluster in the language), an interesting asymmetry emerges. Most of those verbs in which the stop is voiceless belong to the set of *a*-stems (*cantare* 'sing', *truncare* 'truncate', *runcare* 'weed out' etc.), whereas those in which the stop is voiced belong to the set of C-stems (*jungere* 'join', *cumbere* 'lie down', *pandere* 'open', *lambere* 'lick' etc.). While there are exceptions (*rumpere* 'break', *vincere* 'conquer'), this appears to be a strong generalisation. What this means is that putative 1sg forms ***pambo* and ***panco* would not be associated with 2sg forms with a probability that follows from weighting with the lexical frequency of verb classes (which would be ***pambas,* ***pancas* 56.4% vs. ***pambis,* ***pancis* 43.6%);[6] ***pambo* is much more likely to be associated with ***pambis*, while ***panco* is much more likely to be associated with ***pancas.*[7] Since in this paper we do not attempt to answer the question what factors exactly one should take into consideration in weighting, while not debating its importance in theory, for the time being we will simply leave this issue aside.

Given these preliminaries, we used the standard formula in (1) to calculate the unweighted conditional entropies that obtain within the present paradigm of regular verbs. This formula

[4]Throughout the paper the term stem refers to what is called the present stem (or infectum stem) of verbs. The other stem on which finite forms are based is the perfectum stem; in order to keep the discussion streamlined we do not include such forms at all, but they neither contradict nor dilute any of the claims to be made.

[5]We calculated these percentages from Lewis & Short (1879), a non-digitised dictionary. Note that in the LiLa Knowledge Base (lila-erc.eu), an online database that contains a wealth of lexical and morphological information (partly based on Lewis & Short 1879), the percentage values given are somewhat different. This is because LiLa includes lexical items that are postclassical (medieval or even modern), and verbal neologisms were usually created in the *a*-stem class (e.g. *acquietare* 'to acquiesce').

[6]These percentage numbers correspond to the 44% and 34% above, respectively; 1sg forms ending in -C*o* are incompatible with the other verb classes.

[7]One more reason we did not use the numbers given in Pellegrini (2020) is that the implicational relations as presented in that paper are incomplete; to wit, the existence of verbs such as *hiare* 'gape', *pipiare* 'peep', *creare* 'create', *meare* 'wander' indicate that 1sg *-io, -eo* endings are compatible with *a*-stems, i.e. 2sg *-ias, -eas,* 3sg *-iat, -eat* etc., a fact overlooked in that paper.

Table 2. Unweighted conditional entropies in the present paradigm of regular verbs

	1sg	2sg	3sg	1pl	2pl	3pl	Avg
1sg		0.4	0.6	0.4	0.4	0	**0.36**
2sg	1.2		0.6	0	0	0.4	**0.44**
3sg	1	0		0	0	0	**0.2**
1pl	1.2	0	0.6		0	0.4	**0.44**
2pl	1.2	0	0.6	0		0.4	**0.44**
3pl	1	0.4	0.6	0.4	0.4		**0.56**
Avg	**1.12**	**0.16**	**0.6**	**0.16**	**0.16**	**0.24**	

expresses the uncertainty regarding the value of Y given the value of X. The results are as given in Table 2.

(1) $$H(Y|X) = -\sum_{x\in X}P(x)\sum_{y\in Y}P(y|x)\log_2 P(y|x)$$

The table gives the entropy (i.e. uncertainty values) associated with the contents of the paradigm cells on the left when the contents of the paradigms cells at the top are known. For instance, there is no uncertainty whatever about the 3sg form if the 3pl form is known (entropy = 0); the reverse is not true because there is some uncertainty (entropy = 0.6) about the 3pl form even if the 3sg form is known, and even greater uncertainty about the same form if only 1sg is known (entropy = 1).

In terms of the structural sources of uncertainty, the morphophonology of Latin verbal inflection is very simple. Almost all uncertainty, i.e. conditional entropy higher than zero, results from either of two factors. One is a phonological rule that deletes a back nonhigh vowel before other vowels, an exceptionless process in derived environments (Cser 2020, 112–114): this deletes stem-final [a] in the 1sg (suffixed with *-o*), thus rendering such forms highly uninformative with regard to other forms. The other is the heteroclisy of two stem types, C-stems and *i*-stems. There is a sizable class of verbs whose forms coincide partly with the corresponding C-stem forms, partly with the corresponding *i*-stem forms in a systematic fashion (Kaye 2015; Cser 2020, 128).[8] A third, minor source of uncertainty is the shortening of all vowels before final [t], which renders some 3sg forms (suffixed with *-t*) less informative with respect to other forms.

In the rightmost column we see the average predictability of the forms on the left; a higher number means less predictability (i.e. higher entropy). In the bottom row we see the average predictive power of the forms at the top; a higher number means less predictive power. A number of interesting observations can be made about the relation between form frequency (see Table 1) and conditional entropy relations (Table 2); at this point we draw attention to three of these.

[8]In the traditional terminology of Latin grammar this class is called *ī*-stems, as opposed to *ī̆*-stems; in the present paper we refer to the former as heteroclitic, to the latter as *i*-stems.

The most predictable paradigm cell is the most frequent one, 3sg (Avg 0.2). We are not yet quite certain what the significance of this is, but it very clearly stems from the fact that, apart from the 1sg, all the other forms unambiguously predict this form, i.e. the conditional entropy is 0 in the four remaining cells of the 3sg row.[9]

The paradigm cells of the three least frequent forms, 2sg, 1 and 2pl are fully predictive of each other, that is, the conditional entropy in their relations is zero. Furthermore, it is generally true that these three cells have the highest predictive power (Avg 0.16).[10] Again we are not certain how to interpret this fact yet, but we shall see below that these three cells are connected by other features pertaining to their exponence too.

3. PHONOLOGICAL FORM AND INFORMATION CONTENT: PERSON MARKERS

The question we turn to in this section is how much information the sound shape of verb forms conveys in itself. To take a simple example, if any Latin polysyllabic word form ends in an unrounded vowel + [t], it is certain to be an active verb in 3sg, but no further information is conveyed apart from these features. Such a form can in fact represent any of the tense/aspect/mood combinations found among the verbal categories (apart from the imperative):

(2) present: *amat* 'love', *videt* 'see', *agit* 'drive'
subjunctive: *amet* 'love', *agat* 'drive'
future: *aget* 'drive', *veniet* 'come'
perfect: *egit* 'drive', *vidit* 'see'
future perfect: *venerit* 'come', *egerit* 'drive'
subjunctive perfect: *venerit* 'come', *egerit* 'drive'
imperfect: *amabat* 'love', *videbat* 'see'
pluperfect: *viderat* 'see', *egerat* 'drive'
subjunctive imperfect: *videret* 'see', *ageret* 'drive'
subjunctive pluperfect: *vidisset* 'see', *egisset* 'drive'

It is interesting to look at the active person markers from this perspective; here we shall do this by and large in order of decreasing information content. The 3pl ending is *-nt*, which has very high information content: all active 3pl verb forms without exception end in this sequence, and no form of any other Latin word ends in the same. This ending thus has a very strong capacity to convey information: any form ending in this sequence can be immediately identified (i) as a verb and (ii) as a 3pl form. The 3sg ending is *-t*, which in regular inflection is always preceded by a vowel. There are a handful of function words that also end in [t] (e.g. *et* 'and', *ut* '(so) that', *sicut* 'as'), and there are two related nouns (*caput* 'head', *sinciput* 'half a head'), but apart from these

[9]Note that the second most predictable cell is 1sg (Avg 0.36), but this is only because 1sg, in fact, greatly increases the conditional entropy of all the other cells, and it clearly cannot increase its own conditional entropy.

[10]The first and the third observations can be made from the data in Pellegrini (2020, 209), but the numbers are slightly different because of weighting and the inclusion of irregular verbs. The second observation cannot be made from Pellegrini's data for the same reasons.

only 3sg verb forms end in [t].[11] The ending *-t* thus indicates the category of verbs strongly (though not quite as strongly as *-nt*), and within the category of verbs it unambiguously indicates 3sg. Note that even in those irregular verbs in whose 3sg form the *-t* is not preceded by a vowel (*vult* 'want', *est* 'be', *ēst* 'eat', *fert* 'carry' and their prefixed forms) it is never preceded by [n] so there is no room whatsoever for confusion with 3pl. It is to be further noted that the forms that have the highest information content, *viz.* 3sg and 3pl, are also the two most frequent forms.[12]

The 1sg forms end in *-o*, which, in contrast to the third person endings, does not signal the category of verb very well; many nouns and adjectives (in various forms) as well as adverbs end in the same vowel (*homo* 'man', *domino* 'lord' DatSing, *magno* 'big' DatSingMasc, *paulo* 'a little' etc.).[13] Even within the category of verbs, 1sg is not marked by *-o* unambiguously: imperative forms also end in *-to* (*facito* 'do', *ito* 'go', *videto* 'see').[14]

The three remaining forms, 2sg, 1 and 2pl have the endings *-s, -mus* and *-tis*, respectively. Similarly to the *-o* ending, these also do not indicate the category of verb; nouns, adjectives and partly even adverbs can show the same word-final sequences (*anas* 'duck', *infimus* 'lowest', *mitis* 'mild', *penitus* 'thoroughly' etc.). Within the category of verbs, in theory, these forms have the capacity to give rise to ambiguity, since the two plural endings fully include a phonological form identical to the 2sg ending, viz. [s]. Such potential ambiguity could be illustrated e.g. by an English sequence [tæks], which can either be *tax* or *tacks*. But a better example would come from a morphologically richer language, e.g. Hungarian, where [ʃiːnɛk] can either be the dative of *sí* 'ski' or the plural (nominative) of *sín* 'rail', since the dative ending *-nek* fully includes a phonological sequence identical to the plural ending *-ek*. The question now is whether the same relation holds between the 1 and 2pl endings on the one hand, and the 2sg ending on the other.

The question of the 1pl ending *-mus* can be put to rest very easily, since the 2sg ending *-s* is never preceded by a round vowel in Classical Latin, and thus no verb form ending in *-mus* could ever be potentially interpreted as a 2sg form whose stem happens to end in *-mu-*. The situation with respect to 2pl is not as straightforward because there are 2sg verb forms ending in the phonological sequence *-tis* (*petis* 'strive'). However, for morphological as well as lexical reasons, the relevant personal endings can only be preceded by the vowels [aː eː iː i]; which means that ambiguity between 2sg and 2pl could only arise with verb stems ending in the sequence [aːt], [eːt], [iːt] or [it]. For instance, a putative form ***upitis* could be either the 2sg of a verb ***upitere* or the 2pl of a verb ***upere*. But the fact is that such verbs do not exist in Latin at all.

[11] I disregard interjections (e.g. *attat*) throughout the discussion.

[12] It goes without saying that this explanation simplifies the issue of the signifying potential of word forms to some extent. It is quite certain, for instance, that word boundaries were not phonologically indicated in Latin and the resyllabification of word-final consonants was regular; thus e.g. the portions [sinetawroː] of the following two phrases would be homophonous: *sinet auro emere* 'he will allow to buy for gold' and *sine tauro arare* 'to plough without a bull'. We note, however, that since word stress was regularly counted from the end of the word, speakers did have an important prosodic cue regarding boundaries.

[13] Depending on tense and mood, active 1sg forms may also end in a nasal vowel, denoted by an etymologically motivated orthographic *-m*. In the present paradigm, however, this suffix variant is only found in the highly irregular *sum* 'be' and its prefixed variants. Similarly to *-o*, the nasal vowel ending is also found in nouns, adjectives and adverbs.

[14] This is the so called second imperative; it is rarer than the first imperative (*fac, i, vide* for the above verbs, respectively), but is nevertheless amply attested for numerous verbs.

Why is it an interesting statement about the language that no verb stems (not even heteroclitic stems) end in the phonological sequence [a:t], [e:t], [i:t] or [it]? Why is this not simply seen as a lexical accident? We believe there are good reasons to think that the absence of such stems is noteworthy.

Nearly 40% of all verbs are consonant stems, including heteroclitic stems.[15] The stop [t] is the second most frequent consonant in Latin (for exact numbers and how they were arrived at see Cser 2020, 186–188). The vowel [i] is also very frequent (though we cannot give precise numbers for it), and its frequency in word-internal syllables was greatly increased by a sound change in early Latin, whereby short vowels in open syllables were weakened to [i] (Leumann 1977, 79–91, Meiser 1998, 67–73, Weiss 2020, 116–121, Sen 2015, 80–88). This change is well illustrated, among others, by unprefixed vs. prefixed variants of etymologically identical stems:

(3) *facere* 'do' ~ *conficere* 'accomplish'
legere 'gather' ~ *eligere* 'choose'
agere 'drive' ~ *subigere* 'subjugate'
sedere 'sit' ~ *assidere* 'sit by'

In some cases the weakening led to [i:]:

(4) *caedere* (-[aj]-) 'cut' ~ *occīdere* 'cut down'
quaerere 'ask' ~ *conquīrere* 'seek for'

It is interesting, however, that the weakening never produced stem-final [it] or [i:t] sequences; it either failed to apply (or perhaps was reverted without trace), or produced a different vowel in irregular fashion:

(5) *petere* 'strive' ~ *appetere* 'reach after', *repetere* 'strike again'
quatere 'shake' ~ *concutere* 'strike together', *percutere* 'strike through'

It is also interesting that in the preclassical era there was one single verb stem, used with a variety of prefixes, which ended in [i:t]. This was *bitere* (perhaps earlier *baetere*, but there is no accepted etymology, see de Vaan 2008 s.v. *baeto*), and it did indeed produce near-homophonous forms with other verbs in precisely the same way as explained above (see 6). While this verb was often used by Plautus and his contemporaries (3rd–2nd c. BC) in unprefixed form as well as in various prefixed forms, it completely disappeared by the end of the 2nd c. BC, and is no longer part of the lexicon in the Classical Latin period.

(6) *abitis* [a:bi:tis] 'leave' 2sg from *a*+*bitere* ≠ *abitis* [abi:tis] 'leave' 2pl from *ab*+*ire*

What we thus have is a system of verbal inflection in which the category of person/number is not simply indicated by adding the appropriate endings to the stem. The full verb forms, i.e. the stem+ending sequences are of such a phonological shape that even the possibility of ambiguity

[15] 34% are pure consonant stems (see above), and about 5% are heteroclitic stems.

is apparently avoided, and thus their information content is enhanced. Whether this is some form of lexical optimisation is hard to say with certainty; it is certain, however, that the net result is a system in which the distribution of information is bound to exponents in a non-trivial way.[16]

4. STEM-FINAL CONSONANT CLUSTERS

A related phenomenon can be observed in the distribution of nasal+stop clusters in verb stems. Such clusters are the most frequent kind of clusters in Classical Latin (as indeed they are cross-linguistically; for typological background see Greenberg 1965; Côté 2000; Vallé et al. 2009; Gordon 2016, 97ff.). Their textual and lexical frequency shows a parallel distribution according to place of articulation as well as voicing (see Table 3).[17] Nasal + voiceless stops clusters are more frequent than their nasal + voiced stop counterparts; coronal clusters are the most frequent and labial clusters are the least frequent. Both observations are in line with cross-linguistic generalisations. The one relating to voicing is not true of languages that have an active or a historical post-nasal voicing process (e.g. Modern Greek, see Kümmel, 2007, 53ff.); Latin had neither.

What is particularly noteworthy in Latin is that the distribution of nasal + stop clusters is entirely different in the final position of verb stems. There is an apparent preference for the voiced clusters, which is marked for the velar clusters, less marked but visible for the labials (of which there are very few at any rate); but the most striking feature is the complete absence of the single most frequent cluster [nt] from this position.

The occurrence of consonant clusters in such a specific position (end of verb stem) is certainly not expected to reflect the general patterns obtaining in the language; and the numbers are so small as to preclude a proper quantitative analysis.[18] Nevertheless we believe there is good

Table 3. Frequency of nasal + stop clusters in Latin

	Lexical	Textual	Verb stem-final
nt	3,623	503,317	0
nd	1,679	181,214	119
mp	1,219	97,902	13
mb	412	13,527	18
ŋk	1,524	105,648	8
ŋg	737	45,000	127

[16]We do note, however, that irregular verbs add a handful of marginal counterexamples: *vertis* 'turn' 2sg vs. *fertis* 'carry' 2pl, *sistis* 'stop' 2sg vs. *estis* 'be' 2pl.

[17]Textual frequency was calculated from the Packard Humanities Institute database (latin.packhum.org), lexical frequency was calculated from the online dictionary of the Perseus Digital Library (perseus.tufts.edu).

[18]On an etymological basis the numbers are even smaller, since prefixed forms are counted as separate lexemes.

reason to assume that the absence of [nt] from stem-final position is a fact worth reflecting upon. Our arguments are the following.

(i) As was said above, [nt] is the most frequent consonant cluster in the language.
(ii) Consonant-stem verbs are the second largest class of verbs; without heteroclitics, they include about 34% of all verbs; with heteroclitics this number goes up to about 39% (see 3 above).
(iii) An infix/suffix *-n-* was used to form present stems from roots in Proto-Indo-European (Clackson 2007, 153–154). This affix survived in many verbs in the Old Indo-European languages, among them in Latin, where it proved particularly stable before stops (*jungere* 'join', *cumbere* 'lie', *vincere* 'win' etc., cf. Leumann 1977, 533–535 and Weiss 2020, 431).
(iv) The number of *t*-final roots in Proto-Indo-European, from which *-nt*-final present stems could potentially be formed, was considerable. In particular, in Rix et al. (2001) there are 43 *t*-final and 16 *tH*-final roots, to which the 4 *-nt*-final roots can be added. These together make up 6% of the roots reconstructed for Proto-Indo-European and listed in Rix et al. (2001).
(v) Two Latin verbs are of particular interest in this context. One is *pandere* 'open', which is clearly related etymologically to *patere* 'be open', and is reconstructed as deriving from PIE $^{*}peth_2$-with nasal affixation; the other is *mandere* 'eat' from PIE $^{*}menth_2$- (see de Vaan 2008 s.vv. *pando, mando;* the latter alternatively from PIE $^{*}meth_2$-with nasal infixation according to Rix et al. 2001 s.v.). It is not easy to explain why their stems both end in [nd] rather than [nt]. It is quite certain that there was no *[nt] > [nd] change in the early history of Latin. A complex derivation is presented in Schrijver (1991, 222, 498–504), which involves an interplay of several sound changes as well as morphological levelling. Another derivation involving a sound change *[tn] > [nd] is assumed by Weiss (2020, 183) (see also Schrijver 1991, 500 for a discussion and the history of this idea); the problem is that there are virtually no examples for such a change apart from these two verbs. The upshot of this is that apparently there could have been at least two verbs with [nt]-final stems, but for some reason not securely identified they developed in a phonologically different direction.

These arguments lead us to assume that the absence of [nt]-final stems is a fact to be accounted for. The explanation that most readily suggests itself is that the sequence [nt] is itself a phonological form with a morphological function, viz. it functions as the exponent of 3pl; moreover, it is the only consonant cluster in the entire inflectional morphology of Latin that does so for any category. It attaches to stems ending in a nonhigh vowel as [nt]; to stems ending in a consonant or [i] it attaches in the form [unt] (*ama-nt* 'love', *vide-nt* 'see', *ag-unt* 'drive', *veni-unt* 'come'). Note that the present participles are also formed with a *-nt-* suffix (*ama-nt-* 'loving', *vide-nt-* 'seeing', *ag-ent-* 'driving'). Participles are not part of the inflectional system in the same sense in which person-marked forms are, and their endings' morphotactic position is also different in that they are always followed by a case ending, but they attach to stems similarly. The absence of this particular consonant cluster from stem-final position – the most frequent one in a language where stem-final clusters are not rare at all – suggests that configurations in which [nt] would be repeated at a short interval were avoided.[19]

[19]On repetition avoidance in general see Walter (2007).

Within stems, however – though not in final position –, [nt] is found in quite a number of verbs. In several, the cluster emerges via concatenation with the productive prefixes *in-* and *con-* (e.g. *integere* 'cover', *conterere* 'grind'). In some verbs it is found inside unprefixed stems; the stems of these verbs all end either in [i] or in [a] (*sentire* 'feel', *cantare* 'sing', the former a rare type). The 3pl forms of these verbs thus include two instances of [nt], but these are separated by a considerable amount of phonological material: at least two syllable nuclei in the first and second types (*integunt, conterunt*, *sentiunt*), and by the most sonorous – and historically long – vowel in the third (*cantant*), which gives the hearer enough time to process the sequence and disentangle the two occurrences of the potentially morphemic [nt]. With the high vowels [i] [u] this would probably be much harder, and this is indeed borne out by the data: the Packard Humanities Institute database does not include a single instance of **ntunt* or **ntint* sequences.[20]

If the absence of stem-final [nt] was an isolated phenomenon it could perhaps still be seen as fortuitous. But when we look at the other two person markers that are always polysegmental, 1pl *-mus* and 2pl *-tis*, we note that not only do they not occur in stem-final position, in fact they do not occur within verb stems at all (except for the clearly onomatopoeic *muss(it)are* 'murmur, mutter'). Their historically earlier forms, **-mos* and **-tes* are also unattested within verb stems (except for the denominal *testari* 'testify' and its prefixed variants).[21] Thus there seems to be a certain consistency in the avoidance of those sequences in verb stems that are identical to person markers: for the 1 and 2pl endings this seems to be general, whereas for the 3pl ending it appears confined to stem-final position. And while it is unrealistic to assume that similar tendencies would prevail regarding the monosegmental endings (1sg *-o*, 2sg *-s*, 3sg *-t*), we do note the curious fact that no verb stem ends in a round vowel in Latin,[22] and that stems ending in [s] – the most frequent consonant in Latin – are exceedingly rare.[23]

5. CONCLUSION

We have looked at two types of phenomena that in a broad sense belong to the domain of the relation between phonological form and information content within Latin verbal inflection. We have shown that there is a relation between conditional entropy and the textual frequency of the individual forms; and while we have looked at regular paradigms only, with a focus on structural aspects, we are quite confident that our findings can be generalised to the whole of verbal inflection. The other phenomena we looked at concerned sequences within verb stems that had the potential to lead to ambiguity in person marking. We looked at phonological

[20]The latter, **ntint* is actually etymologically impossible, i.e. it could not historically emerge via affixation.

[21]The verb *petessere* 'strive after' is a variant of *petere;* it is only found five times in the classical era, four out of the five in hexametre-final position, where the normal forms of *petere* would be metrically impossible, it can thus be safely dismissed. The numerous verbs ending in orthographic *-tescere* (e.g. *putescere* 'rot') include [teːs], not [tes].

[22]Even if one analyzes the type traditionally called *u*-stems as stems indeed ending in [u] instead of [uw], it is still true that [o(ː)] is the only vowel quality never found in verb stem-final position (for discussion and references see Cser 2020, 130).

[23]The only verb stems ending in [Vs] are *visere* 'view' and its prefixed variants; in those few verb stems that end in [Cs] (*texere* 'weave', *arcessere* 'summon') this sequence is never identical to any 2sg ending, not even irregular ones (e.g. *fers* 'carry').

patterns within stems that could potentially lead to intra-paradigmatic ambiguity; we also looked at whether stems included phonological material that could itself be interpreted as a morphological marker. The absence of potential ambiguity in the first sense, and its severe restriction in the second sense is interpreted here as an emergent mechanism to enhance the information content of verb forms.

Conflict of interest: András Cser is the Editor-in-Chief of Acta Linguistica Academica and has not been part of the study review process.

ACKNOWLEDGEMENT

This research was supported by the ELKH KSZF-14/2023 grant.

REFERENCES

Ackerman, Farrell, James P. Blevins and Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In J. P. Blevins and J. Blevins (eds.) Analogy in grammar: Form and acquisition. Oxford: Oxford University Press. 54–82.

Ackerman, Farrell and Robert Malouf. 2013. Morphological organization: The low entropy conjecture. Language 89. 429–464.

Adamik, Béla. 2015. The periodization of Latin. In G. V. M. Haverling (ed.) Latin linguistics in the early 21st century. Uppsala: Uppsala Universitet. 640–652.

Aronoff, Mark. 1994. Morphology by itself. Cambridge, MA: MIT Press.

Bonami, Olivier and Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. Morphology 28. 167–197.

Clackson, James. 2007. Indo-European linguistics: An introduction. Cambridge: Cambridge University Press.

Côté, Marie-Hélène. 2000. Consonant cluster phonotactics: A perceptual approach. Doctoral dissertation, MIT, Cambridge, MA.

Cser, András. 2015. The nature of phonological conditioning in Latin inflectional allomorphy. Acta Linguistica Hungarica 62. 1–35.

Cser, András. 2020. The phonology of Classical Latin. Oxford: Wiley–Blackwell.

Gordon, Matthew K. 2016. Phonological typology. Oxford: OUP.

Greenberg, Joseph. 1965. Some generalizations concerning initial and final consonant sequences. Linguistics 18. 5–34.

Kaye, Steven. 2015. Conjugation class from Latin to Romance: Heteroclisis in diachrony and synchrony. Doctoral dissertation. University of Oxford, Oxford.

Kümmel, M.J. 2007. Konsonantenwandel: Bausteine zu einer Typologie des Lautwandels und ihre Konsequenzen für die vergleichende Rekonstruktion. Reichert, Wiesbaden.

Leumann, Manu. 1977. Lateinische Laut- und Formenlehre. Munich: Beck.

Lewis, Charlton T. and Charles Short. 1879. Latin dictionary. Oxford: Clarendon Press.

Lieber, Rochelle. 1980. On the organization of the Lexicon. Doctoral dissertation. MIT, Cambridge, MA.

Matthews, Peter H. 1974. Morphology: An introduction to the theory of word structure. Cambridge: Cambridge University Press.

Meiser, Gerhard. 1998. Historische Laut- und Formenlehre der lateinischen Sprache. Darmstadt: Wissenschaftliche Buchgesellschaft.

Oravecz, Csaba, Tamás Váradi and Bálint Sass. 2014. The Hungarian gigaword corpus. In C. Nicoletta, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (eds.) LREC 2014 – Ninth International Conference on Language Resources and Evaluation. Lisbon: European Language Resources Association. 1719–1723.

Pellegrini, Matteo. 2020. Patterns of interpredictability and principal parts in Latin verb paradigms: An entropy-based approach. Journal of Latin Linguistics 19. 195–229.

Rix, Helmut, Martin Kümmel, Thomas Zehnder, Reiner Lipp and Brigitte Schirmer. 2001. Lexikon der indogermanischenVerben: Die Wurzeln und ihre Primärstammbildungen. Wiesbaden: Reichert.

Schrijver, Peter. 1991. The reflexes of the Proto-Indo-European laryngeals in Latin. Amsterdam-Atalanta: Rodopi.

Sen, Ranjan. 2015. Syllable and segment in Latin. Oxford: Oxford University Press.

Sims, Andrea D. and Jeff Parker. 2016. How inflection class systems work: On the informativity of implicative structure. Word Structure 9. 215–239.

Stump, Gregory and Raphael A. Finkel. 2013. Morphological typology. Cambidge: Cambridge University Press.

Vaan, Michiel de. 2008. Etymological dictionary of Latin and the other Italic languages. Leiden: Brill.

Vallée, Nathalie, Solange Rossato and Isabelle Rousset. 2009. Favoured syllabic patterns in the world's languages and sensorimotor constraints. In F. Pellegrino, E. Marsico, I. Chitoran and C. Coupé (eds.) Approaches to phonological complexity. Berlin: Walter de Gruyter. 111–140.

Walter, Mary Ann. 2007. Repetition avoidance in human language. Doctoral dissertation. MIT, Cambridge, MA.

Weiss, Michael. 2020. Outline of the historical and comparative grammar of Latin. Ann Arbor, MI: Beech Stave Press.