https://doi.org/10.7592/FEJF2023.90.kaalep

BASE FORM AND OTHER FORMS OF THE ESTONIAN VERB

Heiki-Jaan Kaalep
Associate Professor
Institute of Computer Science
University of Tartu, Estonia
heiki-jaan.kaalep@ut.ee

Abstract: The article discusses the implicational patterns present in the Estonian verb paradigm: which paradigm slot acts as the base form, which slots act as other principal parts, and what does their dependency hierarchy look like. The argument relies on data from three different types of source: acquisition of Estonian as the first language by children; verbal inflectional classes that are reconstructed from the 17th century Tallinn variety of Estonian; and statistics from different contemporary corpora. The article arrives at a different implicational schema than that which is generally accepted in the Estonian grammar tradition. The article suggests that the base form is the bare stem (which is used as the 2nd person imperative and prohibitive, as well as the negation of present indicative), and that the other three principal parts are: the infinitive; the wordform representing simultaneously the past participle impersonal and past indicative negative impersonal; and third-person singular past indicative. The supine, which is traditionally regarded as the base form, is relegated to being dependent on the third-person singular past indicative. The article acknowledges that the proposed schema causes difficulties with the algorithm of generating paradigm slots for words that now exhibit strengthening gradation pattern, traditionally considered to be unproductive for Estonian words, and even completely missing for verbs.

Keywords: child language, corpus linguistics, historical linguistics, linguistic variation, morphology, old Estonian, paradigm structure, text statistics, verbal inflection

INTRODUCTION

A description of a language has to indicate how inflectional forms of its words are formed. In addition to explicating rules, linguists have traditionally presented exemplary paradigm tables in grammar books and dictionary guides. Exemplary paradigms provide an intuitive insight into similarities and differences both between inflectional forms and also between inflectional classes – declinations and conjugations.

Over time (often over centuries), starting from the very first description of a language, linguists have continued to search and argue for rule sets and exemplary paradigms that would match the language better. A specific set of questions has concerned intra-paradigm implicational relations between inflectional forms: which ones are more basic and which are their dependents, how is a form inferred from a more basic one, and which forms have to be memorised. This is also the focus of the present paper, looking at Estonian verbal paradigm.

It is customary to use supine (*ma*-infinitive) as the base (and index form in dictionaries) for Estonian verbs. This article, however, agrees with those authors who are dissatisfied with this tradition (Ehala 1997; Help 2004).

One may propose different algorithms for generating the complete set of inflectional forms of Estonian verbs. The algorithms may differ in what is chosen as the starting point for the generation (base form or forms), what the applicable rules are, and how the vocabulary should be divided into inflection classes sharing identical (sub)sets of rules. To check that an algorithm is not merely a speculative proposal by a linguist, but is actually used by people, it is common to look at evidence from children who are acquiring their first language, the history of the language, and also inflectional form usage statistics. These three perspectives are used in this article.

It is reasonable to expect that the base form of a paradigm should be the form that changes the least over time, the one that children acquire first when learning their first language, and the one that is more frequent than other forms of the same word in texts. The first form acquired by Estonian children is a bare stem that expresses command, prohibition, and negation (Salasoo 1995; Kohler 2003: 52–68; Argus 2008: 14; Argus & Bauer 2020). When researching morphology, it is customary for children's language researchers to consider this as the base form. Therefore, the task of this article is to make its argument through language history and usage frequency.

The plan of the article is the following. Section 1 gives an overview of sound gradation as a mechanism of Estonian word inflection and its role in the decisions of earlier linguists as to what might be the base form of a verb. Section 2 describes the Estonian verbal paradigm as a set that consists of a small number of subsets of uniformly inflecting paradigm cells, i.e. conjugational series each of which is modelled after a principal part. This view of a paradigm as an arrangement of different conjugational series will be the one used subsequently in the article. Sections 3 and 4 present empirical evidence relevant to determining the base form of the verbal paradigm. Specifically, Section 3 describes conjugations from the 17th century, juxtaposing them with contemporary conjugations, and Section 4 looks for the paradigm slot with the highest prevalence in the vocabulary of different corpora. Section 5 arrives at

a suggestion of paradigm slots for base form and other principal parts, and shows how their intra-paradigm hierarchy plays out with different conjugations, with Section 5.4 paying special attention to deciding on the dependency relationship between the supine and past tense in the indicative mood. Section 6 in turn presents algorithmic and descriptive issues that arise from the proposed base form selection. Section 7 gives conclusions and hints at new perspectives that subsequently arise for historical, lexicographic and morphology theory-related issues relating to Estonian verbs.

1. BACKGROUND AND TRADITION

Inflectional forms of a word are not used with uniform frequency; certain forms are encountered and produced less likely than others. (We use **word** to denote an item of vocabulary; when used, a word has to be inflected, i.e. appear as an inflectional word form.) A speaker needing to use an unseen form has to create it in a transparent manner on the basis of a known form. Intra-paradigm inheritance structure is thus an inevitable result of language being used by people (Bybee 1995: 237).

Choice of the base form and correspondingly the index form for representing a verb in a dictionary is closely tied to the manner in which words are (non) transparently inflected, and particularly to grade changes in word stems.

In Estonian, in a stress-initial disyllabic sequence, i.e. foot, the first syllable may be characterised as being in quantity degree 1, 2 or 3; the number is related to the syllable length and weight. If this foot happens to be a disyllabic stem (or a word form), then this quantity number is also used to characterise the **quantity grade** (**first**, **second** or **third**) of the word form. Inflecting a word may result in sound changes in its stem. If it is only the length of some phones that alternates, causing the stem to be in either the second (i.e. **weak**) or third (i.e. **strong**) **grade**, then the pattern is called **quantitative gradation**. If the change in stem involves (dis)appearance of a consonant (which is possibly accompanied by phonologically conditioned changes in neighbouring phones), then it is called **qualitative gradation**.

For both the quantitative and qualitative gradation, **weakening gradation** means that the base form is in the strong grade, and **strengthening gradation** means that the base form is in the weak grade.

A disyllabic word may exhibit strengthening, weakening or no gradation, and although native speakers have no difficulties in choosing the right pattern, linguists have found it to be challenging to explicate the rules that the speakers appear to be following.

Anton Thor Helle, the author of the first Estonian–German dictionary, published in 1732, chose supine as the index form and suggested rules for how other verb forms should be inferred from it (Helle 1732: 29–31). In the middle of the 19th century, Eduard Ahrens said that "the verb stem is an imperative from which all other forms are made" (Ahrens 1853: 85). However, despite this, he still justified the choice of the supine as the base form by stating that "…it is much easier to form a weak form from a strong form than a strong form from a weak one, and as the illative (*ma*-infinitive) is always a strong form, it seems expedient to treat the illative (without the *ma* suffix) as the verb stem…. But it should not be forgotten that this way is followed only for practical and not theoretical reasons" (Ahrens 1853: 88).

F. Wiedemann also opined that it is not good for a verb's index form to be sometimes in a weak and sometimes in a strong grade (although this is the case for declinable words). He therefore wrote: "in the interests of consistency, I have decided to use the verbal noun ending in *m* (whose usage partially corresponds to the German infinitive) as the base form, and the verbs are accordingly arranged in an alphabetical order. After all, such a verbal noun formed from any verb is always in the strong grade, and it is easier to form a weak grade from a strong grade than the other way around" (Wiedemann 1875: 438).

The same tradition is continued by several contemporary approaches that either merely state that the choice of the base form is traditional (EKK: 204) or add their own semantic reasoning to the traditional view (EKG: 121).

2. PARADIGM SLOTS, WORD FORMS AND ANALOGY GROUPS

This section introduces the reader to the view of the Estonian verb paradigm as an arrangement of different conjugational series, a view that will used subsequently in the article.

A paradigm is an orderly set of grammatical category value bundles, expressible as word forms (EKK: 289). These bundles are commonly referred to as **paradigm members** or **slots**, reflecting the intuition that paradigm has fixed structure, and every word should have inflectional forms that correspond to category value bundles.

It is known that certain word forms of a verb can be used as models for generating other inflectional forms of that verb: using simple analogy rules, one may infer a set of forms from one basic form, a **principal part** of the paradigm. All the forms connected by these analogy rules form an **analogy group**, i.e. a conjugational series. However, it is not immediately obvious which member of an analogy group should be considered its basic form, and how many groups

are there overall. Specifically, languages have words the morphology of which is more or less irregular, the forms of which are distributed between analogy groups differently, and/or the full paradigms of which require more principal parts than others. Excluding these irregular words in various ways makes simpler language descriptions possible, although this comes with the cost of decreased vocabulary coverage. For example, in his grammar, Elmar Muuk (1927) considers that seven principal parts are necessary, Ülle Viks (1992: 47) thinks that four primary, plus eight secondary, principal parts are needed, and the EKG and (Viht & Habicht 2019: 108–109) regard four principal parts as essential. This raises the question of whether the number of principal parts (and thus analogy groups) is something that is decided by the describer of the system of analogy rules, or whether there are other arguments aside from the elegance of the description of the system itself.

One possibility is to assume that there is a natural connection between the expression of grammatical categories and the grouping of forms into analogy groups, and that with a good description the groups emerge on their own. Therefore, the question is whether inferring forms via principal parts is related to the system of grammatical categories or whether it is independent of it. In other words, are the forms that share a certain category value also similar? For example, all forms that express past simple could be similar, regardless of voice or mood (though not in Estonian).

A way to explicate what categories and values are bundled, and how they are expressed in word forms, is by drawing a table where every grammatical category forms a separate column and the rightmost one contains the word form that expresses the set of values for these categories, one per row. If the ordering of categories in this table is based on the principle that the order of the category columns should reflect the order of the morphs representing those same categories, i.e. starting from the stem and proceeding from left to right, then in the case of a perfectly agglutinative inflection system (one with no allomorphs, i.e. one grammatical category value is expressed by exactly one morph, and a bundle of values is expressed by the sequence of such morphs), the word forms that share common left morphs are grouped in aligned rows in the table. In other words, in such a table, similar word forms are grouped according to their grammatical meaning simply because each morph expresses a single grammatical meaning.

However, if the inflectional system is fusional, i.e., one morpheme can express values for a number of grammatical categories (for example, number along with person, as in Estonian) and one such bundle can be expressed by different allomorphs, then it is possible that two sequences of allomorphs (or formatives) encode completely different grammatical meanings, while the forms of these

sequences themselves coincide. In this case, it may happen that no matter how the category columns are arranged, similar word forms are still not grouped. Or, if we approach it differently and place similar word forms in rows close to each other, then the part of the table expressing categories and their values becomes chaotic.

Although the morphology of the Estonian verb is not perfectly agglutinative, after positioning category columns according to the order of morphemes, and arranging formatives that express category value bundles suitably into rows, we arrive at Table 1. Table 1 helps to see how the paradigm can be grouped by word form formation patterns. In terms of categories and the columns that reflect them, it follows Kaalep's (2015) approach to verb morphology. In this table, the word forms that are formed from one principal part, i.e., belong to one analogy group, are arranged into aligned rows and express similar grammatical meanings.

In Table 1, each such group is surrounded by a rectangle with rounded corners and is denoted by the formative of its principal part (Ø, *GE*, *MA*, *S*, *NUD*, *DA*, *TUD*), the highest frequency paradigm member among those belonging to the same analogy group. All forms of one group are formed from the same stem, and there is no allomorphic alternation of affixes for this stem inside that group. Rows with infinitive forms are situated between rows of finite forms; concord of grammatical meanings with analogy groups is thus achieved, albeit at the expense of systematicity in presentation of categories.

The attempt to present analogy groups by aligning rows, i.e. by similar grammatical meanings, leads to separation of some word forms that could be part of one analogy group. Only 5 to 22 irregular words such as *julgeda – julgenud / julenud – julgege* (dare); *näha – näinud – nähke* (see) have slots separated into groups *DA*, *GE* and *NUD*. The exact number of such words depends on whether some phonologically conditioned stem alternatives are counted as different stems. For example principal parts *tuua – toonud – tooge* (bring) could be regarded as having the same stem, just with long vowel heightening in front of *a* and with alternation of the quantitative grade.

Ovals denote two analogy groups that occur in a small number of paradigms and are clearly irregular in terms of the system. Firstly, only one word – *minema* (go) – has the 2nd person singular present imperative (*mine*) based on the stem that is not the same as the stem for the present indicative (*lähe*). Secondly, 16 words use allomorph *a* for the infinitive (e.g. *müüa* (sell)), and according to the written language norm these words also have the affirmative form of the impersonal indicative (*müüakse*) in that same *DA* analogy group, in contrast to the rest of verbal vocabulary where that paradigm slot (row with example *ela-TAKSE* in table 1) belongs to *TUD* analogy group.

Table 1. *Verb paradigm (both finite and infinite forms) with analogy groups and principal parts' formatives as their symbols.*

voice	tense	mood	number and person	aspect	example	symbol
personal	present	indicative	sg1, sg2, sg3, pl1, pl2, pl3	affirmative	*ela-N, -D, -B, -ME, -TE, -VAD*	
			unspecified	negative	(*ei*) *ela*	
		conditional	sg1, sg2, pl1, pl2, pl3	affirmative	ela-KSIN, -KSID, -KSIME, -KSITE	**Ø**
			unspecified	unspecified	*ela-KS*	
		imperative	sg2	unspecified	*ela*	
			unspecified	unspecified	*ela-GU*	**GE**
			pl1, pl2	unspecified	*ela-GEM, -GE*	
		quotative	unspecified	unspecified	*ela-VAT*	
		participle			*ela-V*	**MA**
	supine and its forms				ela-MA, -MAS, -MAST, -MAKS, -MATA	
	past	indicative	sg1, sg2, sg3, pl1, pl2, pl3	affirmative	*ela-SIN, -SID, -S, -SIME, -SITE*	**S**
			unspecified	negative	(*ei*) *ela-NUD*	
		conditional	sg1, sg2, pl1, pl2, pl3	affirmative	*ela-NUKSIN, -NUKSID, -NUKSIME, -NUKSITE*	
			unspecified	unspecified	*ela-NUKS*	
		imperative	unspecified	unspecified	*ela-NUD*	**NUD**
		quotative	unspecified	unspecified	*ela--NUVAT*	
		participle			*ela-NUD*	
infinitive					ela-DA	**DA**
gerund					ela-DES	
impersonal	present	indicative		affirmative	*ela-TAKSE*	
				negative	*ela-TA*	
		conditional		unspecified	*ela-TAKS*	
		imperative		unspecified	*ela-TAGU*	
		quotative		unspecified	*ela-TAVAT*	
		participle			*ela-TAV*	**TUD**
	supine and its forms				*ela-TAMA*	
	past	indicative		affirmative	*ela-TI*	
				negative	(*ei*) *ela-TUD*	
		conditional		unspecified	*ela-TUKS*	
~~-------~~	~~--------~~	~~imperative~~		~~---------~~	~~---------~~	
		quotative		unspecified	*tarvita-TANUVAT*	
		participle			*ela-TUD*	

3. VERB CONJUGATION CLASSES AT THE BEGINNING OF THE 17TH CENTURY

This section juxtaposes contemporary conjugation patterns with those from 400 years ago highlighting what has changed and what has stayed the same.

Change in the manner a word is inflected should mean that the base form remains the same while the formation of other forms will be different. If the change included the base form, it would be changing the word itself. Changes in inflectional system do not take place evenly throughout the lexicon, but via changes in conjugation classes as well as by words moving from one conjugation class to another. There are words the paradigm formation of which does not change over time, and analogy groups that persist over time within paradigms. Thus, one could see here a process of analogical change that has certain regularities.

This section, however, does not attempt to provide an exhaustive overview of the verb morphology of the early seventeenth century. Rather, it attempts to capture a moment in the development of the paradigms of individual verbs. This discussion is based on Georg Müller's sermons (Müller 2007).

3.1. Müller's verb conjugation classes

Georg Müller (c. 1570–1608) was an assistant pastor at the Church of the Holy Spirit in Tallinn. An electronic text corpus (VAKK) and an author's dictionary have been created based on 39 manuscripts of his Estonian sermons from 1600 to 1606 (Habicht et al. 2000). All the tokens in the corpus are morphologically tagged and provided with a modern Estonian dictionary keyword: for example, the modern equivalent of *neütis* is 3rd person past indicative of *näitama* (show). There are 99,000 text tokens in the corpus (whereas the elements of compounds are counted as different tokens), of which 18,000 are verbs (including compound words with the verb form at the end, for example, *vlles+toußnut* (up+risen, i.e. ascended)). The vocabulary size of the corpus is 1,800, including 370 verbs of which 30 are unknown today, for example, *günnima, ihastama, luulma*.

The orthography is variable according to the custom at the time, for example *ieema/iema/iæhma/iæma/iehma* (stay); *iooxma/ioxma/iohxma/ioxsma* (run); *leututh / leutut / leudtuth / leuduth* (found). In the following, however, contemporary orthography will be used to present Müller's verb forms. Problems in interpreting and translating historical orthography into modern forms have been extensively discussed elsewhere (Habicht et al. 2000; Prillop 2003, 2004) and will thus not be further elaborated on here. However, it must be acknowl-

edged that sometimes it is impossible to unambiguously decide how the word form was pronounced at the time.

From the point of view of verb morphology, it is important today whether the letter *d* or *t* is used – for example, *astuda* (step; infinitive) vs. *astuta* (step; impersonal present indicative negative). (The distinction between *t* and *d* marks a length contrast in Estonian, not voicing. Orthographic *t* represents a long voiceless stop /t:/, and *d* a short counterpart /t/.) However, according to Prillop (2020: 168), "the length of a consonant does not play a differentiating role in word meaning in German and is thus not given much attention." It is therefore natural that the Low German orthography followed by Müller does not require consistency in denoting the length of the consonant. Thus, it is impossible to tell the quantitative grade of disyllabic word stems with a long first syllable: for example, whether the forms *hackame*, *hackada* and *hackadta* of modern *hakka* (begin) were in the second or the third quantitative grade. It should also be noted that the *d* / *dt* alternation in Müller is also not reliable for differentiation. Therefore, it is impossible to determine whether or not in Müller's discourse the quantitative grade changed in *hakka* (and if so, was it with a weakening or strengthening pattern). In other words, did it belong to the contemporary *kasva* (grow) or *hüppa* (jump) conjugation class.

The procedure for grouping Müller's verbs was the following. As every token in the Müller corpus is accompanied by a corresponding modern Estonian dictionary headword and by the set of grammatical categories the word form represents, it was possible to gather automatically all the instances of one word, count the frequency of each of its forms and allocate the forms into analogy groups. The (dis)similarities of the analogy groups of different words became clear, and it was possible to determine whether a word inflects similarly to another one from the same corpus, i.e. belongs to the same conjugation. The result is available in https://www.cl.ut.ee/ressursid/mylleri_verbid/.

Defining conjugation classes based on the Müller corpus was similar to describing the grammar of a previously undescribed language or describing for the first time the grammar of the verb of a dialect. The problem of correct typology is indeed complex, as is well known from the historical attempts to describe conjugation class systems of Estonian, see (Viht & Habicht 2019: 365–373) for an overview of these historical attempts.

Considering a comparison with contemporary conjugation particularly important, the classification resulted in an impressionistic system of conjugation classes. In the case of several classes, the option of defining parallel forms was opted for, because due to the lack of corpus data, it is not possible to be certain that all actual forms of the word used at that time are represented in the corpus. In short, we cannot rule out that many words had parallel forms.

Following the example of analogy groups known today, Müller's verb forms can be divided into seven analogy groups, except for one difference: the third-person present indicative was sometimes formed with a strong grade stem, for example *istvad* (today *istuvad* (they sit)). This was a common practice in written Estonian until 1872 when the Society of Estonian Literati (Eesti Kirjameeste Selts) decided that in standard Estonian, this form should be based on the same stem as the rest of the indicative present tense forms. (Kask 1984: 139) However, as the difference in the formation of this form compared to the present day does not change the conjugation class of a single word, it is ignored in the present discussion.

The *GE* analogy group is relevant only for 59 of Müller's verbs and was disregarded because it does not affect inflection class affiliations – there are no such classes where only difference is the way the forms of *GE* analogy group are created. The remaining analogy groups appear for the following number of verbs: *Ø* – 258, *DA* – 198, *MA* – 195, *S* – 106, *TUD* – 152, *NUD* – 215, but only 43 verbs have them all. Therefore, some verbs are assigned to a conjugation class even if some of its analogy groups are empty, but other forms and the phonological form of the base form do not contradict the classification. As a final result, almost 200 of the 370 verbs used by Müller were classified. Too few word forms are available to reliably classify the rest.

Table 2 shows conjugation classes via principal parts of the paradigm. Forms that differ from contemporary ones are presented in bold. The forms given as principal parts are the word forms Müller used, so it may be that the principal parts of one conjugation class are represented by different word forms – the corpus simply did not have all the forms of the sample word. The size of a given conjugation class is also indicated for each example paradigm.

Table 2. *Conjugation classes identified in Müller's sermons.*

Ø	*DA*	*MA*	*S*	*TUD*	*NUD*	No. of words
ela	*elada*	*elama*	*elas*	*elatud*	*elanud*	19
kirjuta	***vihasta/*** *vihastada*	*kirjutama*	*kirjutas* / ***kirjutis***	***kirjutud***	*kirjutanud*	63
valitse	*valitseda*	*valitsema*	***valitsis***	*valitsetud* / ***valitsud***	*valitsenud*	10
ütle	*ütelda*	*ütlema*	***ütlis***	*üteldud*	*ütelnud*	5
otsi	*otsida*	*otsima* / ***otsma***	*uppus* / ***istis***	*otsitud*	*sattunud* / ***satnud***	30
hinga	*hinga[**d/t**]a*	*hingama*	*hingas*	*hingatud*	*hinganud*	9
pööra	*pöörda*	*pöördma*	*pöördis*	*pöördud*	*pöördnud*	14
kanna	*kanda* / ***kandada***	*kandma*	*kandis*	*kannetud*	*kandnud*	27

Many words that have irregular inflection rules today had either completely or almost the same inflection rules back then. For example, monosyllabic stem verbs (*joo* (drink), *jää* (stay), *käi* (walk), etc.) were conjugated as they are today, and were thus also irregular in terms of the conjugation system that existed 400 years ago. Irregular verbs are not reflected in Table 2.

Comparing the conjugation classes then and now, it is evident that the analogy group of a bare stem is the only one that remains unchanged in all classes for 400 years, and that when the word changes its conjugation class, the forms of this bare stem analogy group remain the same while others change. It can thus be concluded that the base form of the word is some form belonging to the bare stem analogy group.

4. PARADIGM SLOT FREQUENCY PROFILES IN TEXT CORPORA

Paradigm slot statistics based on a text corpus makes it possible to narrow the set of hypotheses concerning intra-paradigm dependencies.

Not all inflected forms are kept entirely in human memory; some are formed on the basis of other intra-paradigmatic forms (usually only one) of the same word. This principal part can only be a slot that is already known, i.e. it has been previously encountered. This means that we should not encounter a word's inflectional form representing a dependent paradigm slot without also encountering its principal part in that same synchronic corpus, and this should be true for every analogy group, as well as for the whole vocabulary.

To put it differently, the **type frequency** (i.e. the number of unique word forms) per any paradigm slot should not exceed the type frequency per its governing principal part, and the type frequency per the paradigm slot that acts as the base should not be smaller than type frequency per any other slot.

Counting and comparing type frequencies per paradigm slots reveals that different corpora are very similar in noun paradigm slot frequency profiles, and very dissimilar in verb paradigm slot frequency profiles. It can be said in advance that both the differences in verb paradigm slot frequency orders and the difficulties in associating them with other regularities (the history of language changes, the correlation between the length and frequency of forms, and the order in which children acquire the forms) force us to look at a number of different corpora.

4.1. Declinable words

In the case of declinable words, the corpus data are straightforward. On the basis of one from the family of UT corpora, the 0.5-million-token morphologically tagged corpus of written language, Kaalep (2018) has found that type frequency per singular nominative exceeds the type frequency per any other paradigm slot. This is followed (in decreasing order) by singular genitive, singular partitive, plural nominative, etc. Type frequency per singular case exceeds that of the same case in plural.

In four of the five 0.1-million-token sub-corpora of this corpus (journalism, Estonian literature, the popular science magazine *Horisont*, George Orwell's *1984*), the type frequencies per different cases coincide with the pattern found in the whole corpus. Only the 0.1-million-token corpus of legal texts is different, as its type frequencies per singular and plural genitive are unusually large. The order is: genitive, nominative, and partitive singular followed by genitive and nominative plural. Additionally, a 0.1-million-token conversation corpus representing spoken language has a nominal paradigm slot type frequency profile similar to that of standardised written language. The morphologically tagged 0.1-million-token chatroom corpus word usage differs from standardised written language only in that the type frequency per singular partitive is a little higher than per singular genitive.

In addition to the UT corpora, one might consider the 0.4-million-token caregiver language portion of CHILDES Estonian. It represents the language that children hear at the language learning age. This means that the frequency characteristics of this corpus are the ones which children base their language knowledge on. After automatic morphological analysis and disambiguation (for a description of the tools, see Kaalep & Vaino 2000), it turned out that the order of the topmost cases by their type frequencies was the same as in the corpus of written language described above.

The intra-paradigm implicational hierarchy is in accordance with the order of the type frequencies of its slots: nominative singular at the top, followed by genitive and partitive, and then the other cases (Kaalep 2018).

The order of the type frequencies is also in line with the rule that more frequent items are shorter. The singular nominative case has no ending, a theme vowel is added to singular genitive stem, the partitive must have a theme vowel and/or case ending *-d* / *-t,* etc.

4.2. Verbs

In stark contrast to the almost non-existing inter-corpus variation in type frequency order of the nominal paradigm slots, the type frequency order of verb paradigm slots shows great inter-corpus variation between the same corpora.

Figure 1 shows for five corpora (a subset of those eight described in section 4.1) the proportion of the verbal vocabulary (in percentages) that turns up to realise this paradigm slot. (The percentages add up to more than 100% because one verb can turn up in several different inflectional forms.) The slots are denoted by their affix formatives (as in Table 1). Importantly, they represent all seven analogy groups, with the top slots of every group in terms of their type frequencies.

Figure 1 shows that the frequency profile of even the most prevalent paradigm slots is quite different in different corpora. In the literature corpus, 55% of the verbs turn up as the third-person past indicative mood (*S*), 35% as *nud*-participles or negations of past tense indicative mood (*NUD)*, 35% as third-person singular indicative mood (*B*). In contrast, in the CHILDES corpus, more than 45% of the verbs turn up as the second-person present imperative or negations of the indicative mood (bare stem with no ending, marked as Ø), 45% as (*B*), 35% as the second-person singular indicative mood (*D*), and 30% as infinitive (*DA*). In the chat room corpus, too, almost 40% of the verbs turn up as bare stems (Ø), more than 30% as (*B*) and almost 30% as (*DA*). In the conversations corpus, 35% of the verbs turn up as (*B*), 35% as (*DA*), and almost 30% as supine (*MA*). In the journalism corpus, about 50% of the verbs turn up as (*B*), about 45% as (*DA*), and almost 40% as (*S*).

There are slots the lines of which do not really stand out in the figure (for example, *MA*) or do not stand out well. This means that the proportion of the vocabulary realising these slots is similar in different corpora.

Paradigm slot statistics for the different corpora provides conflicting evidence as to what could be the base form – *B*, Ø, or *S*. The recognition that the genre characteristics of the text corpus influences the choice of verb forms more and differently than that of noun forms means that the nature of the corpus must be carefully considered when examining the system of verb morphology. Only in the CHILDES and chat room corpus is the shortest verb form also the most common. One argument supporting the idea that literature and journalism corpora are not a suitable basis for studying language as a learnable system is that they include genre-specific texts the creation of which needs to be specially studied for. On the other hand, chat rooms, although having written communication instead of oral, seem to represent natural usage of language, i.e., language that does not require one to learn genre-specific features.

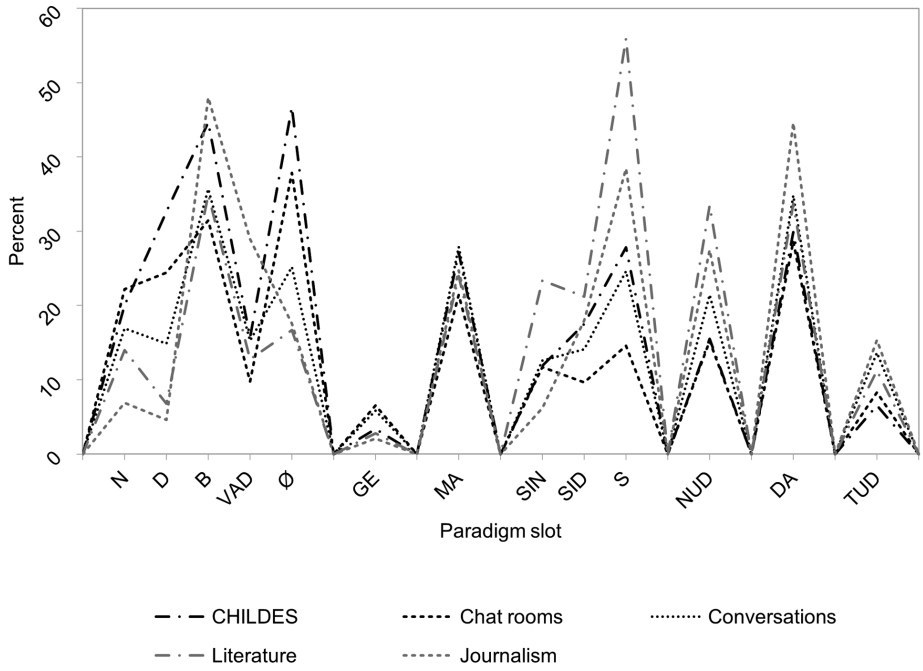

Figure 1. *Proportion of vocabulary realising a paradigm slot in different corpora.*

CHILDES should reflect the natural use of language on which language proficiency is based. On looking at the usage of verb forms, it could be said that there is a lot of talk about wishes, commands, and refusals in this corpus. This is typical of situations where people do something together, whether they are a mother and a child or builders building a house together (*anna haamer*, *võta ise*, *nii ei saa*, *ära siia astu* (give me a hammer, take it yourself, cannot do that, don't step here)). Interestingly, the usage statistics for the verb paradigm slots of the chat room corpus are very similar to CHILDES, indicative of a functional style that is used for similar (albeit mental, not physical) interactions. The frequency profile of verb paradigm slots in the literature corpus reflects the fact that it contains many narratives (descriptions of how someone once did something). The journalism and conversation corpora are both heterogeneous in terms of genre and thus their frequency profiles are not easy to interpret. For better interpretation, the corpora might need to be divided into even smaller sub-sections (for example, opinion articles, news; storytelling, talk during problem-solving).

Unsurprisingly, the caregivers' language in the CHILDES corpus is in harmony with the order in which children acquire inflectional forms. The corpus

of chat rooms, in turn, shows that the functional style represented by the caregivers' language is not different from the communication style of adults, insofar as it is manifested by the frequencies of verb paradigm slots.

The takeaway message from this section is that not every corpus is in accordance with evidence about language learning stages by children, although relevant corpora are.

5. IMPLICATIONAL HIERARCHY OF PARADIGM SLOTS

Having established (in sections 3 and 4) that the best candidate for the base form of the paradigm is the bare stem (Ø), we can start to sketch the rest of the paradigm hierarchy: how are the principal parts related to each other, and how do these relationships reveal themselves in conjugation classes?

Figure 2 shows a hypothetical implicational hierarchy of principal parts. The direction of the arrow is from the basis of inference towards the inferred. However, this does not mean that every form of every word should be inferred from its immediate head. The form may simply be memorised, or it may be based on another form higher up the same line of hierarchy. For example, in the case of *lepi* (agree), *S* is based on *DA* (*leppida*> *leppis*), and *TUD* is based on Ø (*lepi*> *lepitud*), but in the case of *hakka* (start), *S* is based on Ø (*hakka*> *hakkas*), and *TUD* is based on *DA* (*hakata*> *hakatud*). This means that Figure 2 is suitable for narrowing down the choice of possible principal parts, but for making the final choice, something specific to conjugation class or to the word still needs to be known. What it is, and how would possible additional rules and restrictions allow a more precise and stricter hierarchy to be presented will not be described here.

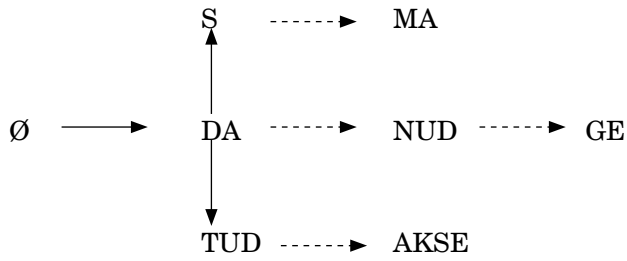

Figure 2. *Proposition for implicational hierarchy of principal parts.*

The principal parts linked with bold arrows represent the four traditionally distinguished analogy groups (for example, Viks 1992; EKG; Viht & Habicht

2019) with the exception that instead of the third-person singular (*B*) and supine (*MA*), the bare stem (*Ø*) and past tense (*S*) are used to represent them here.

The endpoints of dashed arrows represent secondary principal parts. They stand for analogy groups the very existence of which becomes apparent only in paradigms of a few words and in view of a few irregular forms. For the vast majority of words, the word forms of these groups may even be included in the analogy group higher in hierarchy. For example, a need to infer some word forms from *S* and some from *MA* is necessary for only 15 words the past indicative forms of which are formed by the (allo)morph *i* (*sõi – sööma* (eat), *pesi – pesema* (wash)). Using different rule sets and bases for *TUD* and *AKSE* analogy groups is necessary for only the standardised forms of 16 words (*viiakse,* not **viidakse* (take)). Separating *DA*, *NUD* and *GE* is necessary for 5–22 words (*süüa – söönud* (eat), *joosta – jooksnud – jooske / jookske* (run)), the exact number being dependent on what counts as differences in word forms, as explained in section 2.

The present approach differs from the traditional one in that it proposes that first, principal parts form a hierarchy, and second, that the bare stem is at the top and the supine is relegated to a secondary principal part.

This hierarchy becomes apparent only when we look at inflectional classes. Agglutinative morphological systems (such as Turkish noun declination) are not allomorphic or only have allomorphs that are phonologically motivated, i.e., each paradigm slot can only be realised in one way. All slots are easily predictable and there is only one conjugation class. However, if it is possible to express some grammatical category value by means of different allomorphs, i.e., for each word there is a need to choose between them, the question of basis of this choice immediately arises.

In the simplest case, the choice depends only on the phonological form of the base form. For example, all words with *a*-terminal base forms inflect in one way and all other words in another way (this one condition may affect the choice of allomorphs of several different morphemes, i.e., two conjugation classes differ in many forms). However, it is possible that some slots of the paradigm are not predictable from the base form: over time, the phonological structure of individual forms has changed and/or the rules that previously allowed one word form to be inferred from another are no longer applicable. In this case, language users simply have to memorise these word forms.

In fusional languages, words are divided into conjugation classes. Some conjugation classes are productive, i.e., new words can be added to them, and all forms can be constructed from the base form by rules. Typically, they already have a large number of words. Productive classes differ from each other by the set of rules that are applicable for inferring all the word forms, and also by the phonological-derivational structure of the base forms of their member words;

it is this difference in base form structures that allows a speaker to pigeonhole words into the correct classes.

In addition to productive classes, there are non-productive ones, i.e., no new words are added to them, and not all forms can be inferred from the base form. Non-productive classes vary in size and regularity. For some, it is enough to know two principal parts; in the case of the most irregular, even seven may not be enough (for example, *söö* (eat), *näe* (see), *mine* (go), *ole* (be)).

Before showing how the proposed hierarchy 'works', a few words on the phonological shape of words belonging to all five productive Estonian conjugational classes are necessary.

The word form by which to determine class membership is the bare stem. One must consider whether the word is a derivation, the number of syllables of the stem, the final vowel, and in the case of disyllabic stems, its quantity grade.

1) If the stem is non-derived, disyllabic, ends with *a*, and has the third quantity grade, it changes like *hüppa* (jump). One must also take into account whether the word can be changed into having the second quantity grade at all: for example, *koonda* (aggregate) is in the third quantity grade, but by the rules of the quantitative grade alteration, it cannot be shortened into the second quantity grade, and thus it cannot belong to this class.

2) If the stem is non-derived, disyllabic, ends with *i*- or *u*, and is in the second quantity grade, it inflects as *õpi* (learn).

3) If the stem is disyllabic, ends with *le*, and is in the third quantity grade, it inflects like *hüple* (jiggle).

4) If the stem is trisyllabic and ends with *ele*, it inflects like *rabele* (struggle) (i.e., like both *hüple* and *ela*).

5) In all other cases the word inflects like *ela* (live).

As the listed classes are productive, no dictionary can contain all the words belonging to them. However, based on the *Filosoft* speller dictionary, *ela*-class contains 5,000, *lepi*-class 1,800, *hakka*-class 400, *hüppa*-class 150, and *rabele*-class 40 words (of the words with grade alteration only those with quantitative grade alternation are considered here because qualitative grade alternation is a non-productive phenomenon, i.e., both the strong and the weak grade form must be memorised). There are about 350 words in the speller dictionary that do not belong to the five classes listed above, i.e., they belong to non-productive conjugation classes. 250 of these words exhibit qualitative grade alternation.

Below, the ways of inferring word forms from a slot higher in the hierarchy are outlined.

5.1. Bare stem (Ø)

The top of the hierarchy, or the base form, is a bare stem. In the class without grade alternation, i.e. *ela*-class, all other principal parts can be easily derived from this. In fact, formally one could choose any slot of the paradigm to be the base form for this class because producing any form is trivial: just append an affix to the bare stem. The building of *ela*-class forms is not discussed below.

5.2. Infinitive (*DA*)

In productive conjugation classes, the infinitive can be obtained from the base form in the following way.

The default affix allomorph is *-da*; in the *hüppa-class*, *-ta* (*hüpata*).

In the *hüppa*-class, a weak grade must be formed from the stem of the base form. In the *lepi*-class, a strong grade must be formed. This is further discussed in section 6 Problems.

Words belonging to the *hüple*-class are disyllabic words with a stem in the strong grade, i.e. in the third quantitative grade. When forming the infinitive, the stem is converted into the weak grade and the final *e* deleted. To eliminate pronunciation difficulties, an *e* is inserted in front of *l* (*hüpelda*). *Rabele* can be interpreted in two ways in terms of its phonological form. It is trisyllabic, so it has no grade alteration and belongs to the *ela*-class (*rabeleda*). However, considering it ends in *le*, it may belong to the same conjugation class as *hüple*, which means that when making the infinitive, the final *e* of the stem is deleted (*rabelda*).

If a word belongs to a conjugation class with grade alteration, meaning in *lepi-, hüppa-,* or *hüple-class*, and its grade-altering is qualitative, then speakers simply have to remember the form of the opposite-grade infinitive. Otherwise, they form the infinitive with quantitative grade alteration.

There are about a hundred words belonging to other non-productive conjugation classes (for example, *naera* (laugh), *seisa* (stand), *sööda* (feed), *nuta* (weep), *tule* (come), *too* (bring), *vii* (take)) and their infinitive is not predictable from the bare stem. It must simply be memorised.

5.3. The third-person singular of the past indicative mood (*S*)

By default, the past tense marker *-s* is added to the bare stem (Ø), although in the *lepi-class* it is added to the infinitive stem (*DA*) instead. The same rule applies to qualitative grade-changing words of the *lepi-, hüppa-*, or *hüple*-classes.

In any case, the final vowel of the stem remains the same as at the end of the bare stem.

If the stem of the infinitive (*DA*) ends with a consonant, i.e. the word belongs to an unproductive class *seisa* (stand), *naera* (laugh), *saada* (send), *leia* (find), *peta* (deceive), or *jäta* (leave), then the stem of (*S*) must be in strong grade and its creation requires knowledge of the bare stem as well as the infinitive form. The theme vowel of the stem will be *i*.

There are about twenty words in other non-productive conjugation classes and their forms simply need to be memorised.

5.4. The position of supine (*MA*) and past indicative (*S*) in the hierarchy

The morphology of the supine and *s*-allomorphic past tense forms is fairly similar across conjugation classes. Traditional approaches deeming *MA* to be the base form of the word think the forms of the past tense are derived from it. The rule would be as follows:

The forms of *MA* and *S* are based on the same stem. If *MA* has a consonant-final stem (*naerma* (laugh)), then *S* is formed by adding the vowel *i* (*naeris)*. If *MA* has a vowel-final stem (*uskuma* (believe)), then *S* is formed simply by adding the tense marker to the same stem (*uskus*). These regularities are valid no matter the conjugation class of the word. They do not apply only to 17 irregular monosyllabic words (*saama* (get), *tooma* (bring), etc.) and to 11 disyllabic words whose past tense marker is *i* (*pesi* (washed), *tegi* (did), *o*li (was), *lasi* (let), etc.). *Kaitsema / kaitsma – kaitses* (defend) and *maitsema / maitsma – maitses* (taste) are also exceptions to that rule.

However, if we assume that *MA* is not the word's base form (because Ø is), then the question arises whether *S* could be the base of *MA*. The rule would be symmetrical to the above rule based on *MA*, for only the base and the derived would have exchanged positions. If the stem vowel of *S* is the same as the vowel of Ø, then *MA* has the same stem. If the vowel of *S* is different (in this case it is the vowel *i*), then *MA* is based on the *S* stem minus the vowel *i*, i.e., on the consonant-final stem. In parallel to the case when inferring was assumed to be based on *MA*, this rule does not apply to irregular monosyllabic stems and to words with the *i*-marked past tense.

The exception and problem for both bases – *MA* and *S* – is the existence of parallel forms for *MA – kaitsema / kaitsma* (defend), *maitsema / maitsma* (taste), and singular for *S – kaitses, maitses*. The only theoretically plausible pairs of *MA* and *S* would be *(m | k)aitsema – (m | k)aitses* and *(m | k)aitsma – (m | k)aitsis.*

In earlier times, *(m|k)aitsis* was used indeed, according to VAKK, for example, A. Thor Helle 1739, Fr. R. Kreutzwald 1840). Now the question is: what was the chain of events that led to the contemporary forms? Did *(m|k)aitsis* change into *(m|k)aitses*, and this induced *(m|k)aitsema*, with *(m|k)aitsma* remaining as a remnant from the past? Or, alternatively, did *(m|k)aitsma* develop an alternative form *(m|k)aitsema* which in turn induced *(m|k)aitses*? This alternative seems unlikely, because how come the ability of *(m|k)aitsma* to induce *S* disappeared completely, as evidenced by the lack of *(m|k)aitsis* in contemporary Estonian?

To clarify the issue, we can turn to words that are currently leaving their conjugation class. In this case, alternative word forms for the same paradigm slot, with different frequencies, are used. Table 3 shows the frequencies in the etTenTen13 corpus of the principal parts of the three words *naase* (return), *veena* (convince) and *mööna* (concede) that historically belong to the *naera*-class. To form all the principal parts of the *naera*-class (except for the bare stem), it is necessary to remember the infinitive (*DA*). The top row of a cell contains the historical (which is also the contemporary normative) form, and the bottom row the form according to its new conjugation class. The numbers reflect the token frequency of the old/new form, respectively.

Table 3. *Frequencies of principal parts of words in the process of changing their conjugation class (etTenTen13).*

Ø	*DA*	*S*	*MA*	*NUD*	*TUD*	*GE*
naase 215	*naasta* *naaseda* 2002/95	*naasis* *naases* 1154/519	*naasma* *naasema* 415/44	*naasnud* *naasenud* 1462/18	*naastud* *naasetud* 217/0	*naaske* *naasege* 9/0
veena 546	*veenda* *veenata* 3523/3	*veenis* *veenas* 549/19	*veenma* *veenama* 502/2	*veennud* *veenanud* 274/2	*veendud* *veenatud* 102/0	*veenge* *veenake* 18/0
mööna 16	*möönda* *möönata* 346/0	*möönis* *möönas* 1785/43	*möönma* *möönama* 366/1	*möönnud* *möönanud* 121/2	*mööndud* *möönatud* 10/0	*möönge* *möönake* 1/0

These words have started to move from their historical conjugation class to one corresponding to the phonological structure of their base forms (bare stems), in which it is not necessary to memorise anything other than the base form for deriving any forms: the disyllabic *naase* (return), which ends with *e* and has a long first syllable, is on the road to the non-gradational *ela*-class, while similar *a*-final words *veena* (convince) and *mööna* (concede) are on the road to becoming gradational *hakka*-class words. For all three words, only the analogy group with the bare stem remains unchanged.

The frequency differences between the words themselves and between the principal parts (Ø, *DA*, *S*, etc.) are irrelevant, but the frequencies of possible alternative realisations are significant. Specifically, the numerical relations between the old and the new forms show that the analogy groups are moving to a new conjugation class at a different pace: some have barely started, i.e., the forms are still old-fashioned, while other groups use many new forms instead of old ones. New forms are most widely used in the past indicative (*S*), i.e., this group has reached the farthest point in its transition. In the remaining groups, new forms are far less likely to replace the old ones, i.e., they are much more conservative. For example, when comparing the ratios of old/new forms of *S* and *MA*, the likelihood of meeting an innovative *S* form is greater than the likelihood of meeting an innovative *MA* form: for *naase,* five times, for *veena* and *mööna*, ten times. (*Naase* seems to have gone further with conjugation class change than *veena* and *mööna*: the new forms of *naase* are now more likely to replace old ones than those of *veena* and *mööna*.)

From the point of view of the intra-paradigm hierarchy of paradigm slots, the logic is that the way the principal part inflects must change before the way the inferred form inflects: language users must see the new principal part before they can form anything on its basis. Thus, since *S* moves toward the new way of inflecting faster than *MA*, the latter cannot be the basis for the former. On the contrary, *S* must be the basis of *MA*.

In addition, during the period of learning to speak, children start using the indicative past tense (*S*) forms earlier than the supine (*MA*). (Vihman & Vija 2008; Argus & Bauer 2020)

It turns out that the supine is only a secondary principal part, i.e. occupies a far less prominent position in the paradigm hierarchy than the linguistic tradition assumes.

6. PROBLEMS

If the base form of the verb is always assumed to be a strong-grade stem, then the only possible grade alteration pattern would be towards the weak grade. Indeed, this has been the approach this far. However, if the data on language acquisition and change over time show that the base form may be in the weak grade, as in the *lepi*-class, then there is a need to describe how a productive grade alteration pattern towards the stronger grade can take place at all. This is a serious problem for the morphologist (which has been successfully ignored for 150 years through the choice of the base form), with no solution currently visible.

The most serious objection to the productive grade alteration pattern towards the stronger grade is that today a strong grade word form can be inferred from a weak grade one in several ways. For example, the strong grade infinitive (*DA*) of a word ending with …*angu* could be either …*anguda* (like *manguda* (cadge, scrounge)) or …*ankuda* (like *vankuda* (falter, teeter)). However, language usage data shows that possibly faulty alternatives are not created, i.e., language users unanimously choose the same correct way. This could mean that they do not create this strong grade form of this type of words according to a rule, but already have it in memory. This, in turn, would mean that *DA* still cannot be formed from Ø. All in all, this is a logical contradiction.

This logical contradiction is of the same type as seen when taking a closer look at nominative singular and genitive singular forms of declinable words. Consonant-ending nominative is moulded into genitive by appending a theme vowel, such as …*eit* > …*eide* (such as *eide* (hag)) or > …*eidi* (such as *kleidi* (dress)). The choice of vowel seems to be unpredictable. However, language usage data show that despite this seeming impossibility of predicting, users very rarely make mistakes, as if they had the right version in their memory. In this case, the solution comes from an observation that (depending on the phonological structure of the word) some vowels are appended to very few words. This means that the users really do have the right genitive versions stored in memory, although not necessarily for all the words, but rather just for a few of them. The users have to remember the short list of irregular words that inflect with exceptional vowels. For other words, the rule is that the suitable vowel is the one usually used for similar words (Kaalep 2012).

Would a similar solution be possible when choosing the right infinitive form (*DA*) for *lepi*-class words? There is some hope in the fact that for many words, the strengthening grade alteration can still be applied according to some rules and in that even if it cannot, the alternatives have unequal probabilities. In the *lepi*-class, there are 1,800 verbs with altering quantitative grade. The only group among them for which the formation of a strong grade is not unambiguously determined are words with the structure C* V [V | L] G [*u* | *i*], i.e. disyllabic *u*- or *i*-ending words in the second quantity grade, and their internal phones are either two vowels (a long vowel or a diphthong; V) or a vowel and a sonorant (L = l, m, n, r) followed by a short stop (G = g, b, d), for example, *vangu*, (falter, teeter) *mangu*, (cadge, scrounge) *räägi*, (talk) *määgi* (bleat). Their strong grade can be formed either by lengthening the short stop – *vankuda, rääkida*, which is done in 150 words, or by lengthening the sonorant (not shown in orthography) – *manguda*, *määgida*, which is done in 25 words. This means that, in principle, it would be easier to memorise the rarer pattern word by word.

If these 25 words were common and old, it would be very plausible that their irregular morphology was memorised. However, many of the 25 words have recently entered the language, for example, *svingi*, (swing) *hängi* (hang) and thus the existence and mechanism of a specific grade alteration pattern is not clear yet.

Another problem arises with qualitative grade altering words, such as *pöa – pügada* (shear, prune) the bare stem of which was originally in weak grade and is being replaced by a strong grade stem, i.e., *pöa* > *püga*. Such a development is incompatible with the claim that the bare stem is the base form: how can it be that the base form changes but other forms remain the same? As a solution, it could be suggested that for some words, the analogy group forms of a bare stem may be so rare that the bare stem really cannot function as the basis for other forms, and becomes similar to one of the more common ones (especially the infinitive). The word *pöa*/*püga* may serve as an example: in etTenTen13 *pügada* occurs 500 times, *pöetud* 130 times, and *pöa*/*püga* only 20 times.

7. SUMMARY AND FURTHER THEORETICAL PROBLEMS

This article offers explanations for some phenomena in the Estonian verb morphology, although in turn some things that were not unexplainable according to previous theory, have become so. Thus, the article highlights the need for clarification and rather asks questions than gives exhaustive answers.

The article presents structure of the verb paradigm by grouping paradigm slots according to grammatical categories, as well as by analogy groups related to the principal parts. The analogy groups might be called building blocks of the paradigm: to describe how to infer the word forms of a paradigm, it is sufficient to describe how to infer one word form of every analogy group. In this article, analogy groups make up the prism, or the method of grouping usage-based data, through which to look at language change over time, acquisition of the first language by children, and word form usage statistics. The choice of the paradigm slot that represents an analogy group, in turn, is arbitrary from the point of view of paradigm structure description. However, in this article, the slot with the highest type frequency has been chosen as this representative.

The article suggests that the bare stem should be considered as the base form of the verb. The other principal parts are the infinitive, the third-person singular past indicative, and the past participle of the impersonal voice/negation of the indicative mood. These suggestions are based on the phenomena that become visible in language usage: morphology changes since the 17th century, the order of the verb forms children acquire when learning their first language, and form usage frequencies in text corpora. In addition, the algorithmic possibility of

inferring one word form from another is also taken into account. If such an algorithm could not be proposed (for example, to infer every strong grade stem for *lepi*-class words), then failure is not considered a sufficient argument to immediately rule out the existence of an implicational relationship between word forms.

If the hierarchy of verb paradigm slots is as proposed in this article, several traditional positions need to be reconsidered. Further research should show whether the new perspective is better than the traditional one.

1) According to tradition (EKK), there are no strengthening grade alteration conjugation classes among verbs (which is surprising, as they exist among declinable forms). It now turns out that the productive *lepi*-class has strengthening grade alteration.

2) The claim that there has been an internal loss of phones corresponding to universal sound change in verbs (for example, *laulamaan* > *laulma* (sing)) (Kettunen 1962: 162) needs to be reconsidered, as it concerns one of the principal parts of secondary importance. Perhaps these are simply analogy shifts within the paradigm, which in turn are caused by the fact that allomorphs used to express the infinitive and the past tense have been replaced by others (*laulaa* > *laulada* > *laulda*; *lauloi* > *laulis*).

3) Some of the keywords in dictionaries that have very similar meanings to other words and are tagged as archaic are actually not different words but exhibit different conjugation patterns of the same word. For example, keywords *koolma* and *koolema* are supine forms of the same word *koole* (die); *ulguma* and *uluma* are supine forms of *ulu* (howl). (It is apparent in both cases that the word is moving into the *ela*-class.) Perhaps the tradition that supine is the index form of a verb should be changed?

ACKNOWLEDGEMENTS

The work was partly supported by the Centre of Excellence in Estonian Studies (CEES, TK 145).

WEB RESOURCES

CHILDES = Children's and custodians' language corpus https://childes.talkbank.org/access/Other.

etTenTen13 = A corpus of Estonian language webpages downloaded from the Internet on Keeleveeb www.keeleveeb.ee/dict/corpus/ettenten/about.html.

UT corpora = Corpora. The University of Tartu computer linguistics research group https://www.cl.ut.ee/korpused.

VAKK = The corpus of old written language https://doi.org/10.15155/TY.0005.

REFERENCES

Ahrens, Eduard 1853. *Grammatik der Ehstnischen Sprache Revalschen Dialektes*. Reval: Kluge und Ström.

Argus, Reili 2008. *Eesti keele muutemorfoloogia omandamine*. [Acquisition of Morphology in Estonian.] (Humanitaarteaduste dissertatsioonid 19.) Tallinn: Tallinna Ülikooli Kirjastus.

Argus, Reili & Bauer, Annika 2020. Muutevormide ilmumine eesti keelt esimese keelena omandavate laste kõnesse. [Emergence and Productive Use of Inflectional Forms in Early Estonian.] *Philologia Estonica Tallinnensis*, No. 5, pp. 17–57.

Bybee, Joan L. 1995. Diachronic and Typological Properties of Morphology and Their Implications for Representation. In: Laurie Beth Feldman (ed.). *Morphological Aspects of Language Processing*. Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 225–246.

Ehala, Martin 1997. Eesti morfoloogia olemus. [The Spirit of the Estonian Morphology.] *Keel ja Kirjandus*, No. 6, pp. 370–383.

EKG = Erelt, Mati & Kasik, Reet & Metslang, Helle & Rajandi, Henno & Ross, Kristiina & Saari, Henn & Tael, Kaja & Vare, Silvi 1995. *Eesti keele grammatika I. Morfoloogia. Sõnamoodustus*. [Estonian Grammar I. Morphology. Word Formation.] Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.

EKK = Erelt, Mati & Erelt, Tiiu & Ross, Kristiina 2007. *Eesti keele käsiraamat. Kolmas, täiendatud trükk*. [Handbook of Estonian. Third Edition.] Tallinn: Eesti Keele Sihtasutus.

Habicht, Külli & Kingisepp, Valve-Liivi & Pirso, Urve & Prillop, Külli 2000. *Georg Mülleri jutluste sõnastik*. [Dictionary of Georg Müller's Sermons.] (Tartu Ülikooli eesti keele õppetooli toimetised 12.) Tartu: Tartu Ülikool.

Helle, Anton Thor 1732. *Kurzgefasste Anweisung Zur Ehstnischen Sprache*. Halle: Stephan Orban.

Help, Toomas 2004. *Sõnakeskne keelemudel. Eesti regulaarne ja irregulaarne verb*. [Word-Centered Language Model. The Regular and Irregular Verb in Estonian.] (Dissertationes philologiae estonicae Universitatis Tartuensis 13.) Tartu: Tartu Ülikooli Kirjastus.

Kaalep, Heiki-Jaan 2012. Eesti käänamissüsteemi seaduspärasused. [Patterns in the Declination System of the Estonian Language.] *Keel ja Kirjan*dus, No. 6, pp. 418–449.

Kaalep, Heiki-Jaan 2015. Eesti verbi vormistik. [Estonian Verb Paradigm.] *Keel ja Kirjandus*, No. 1, pp. 1–15.

Kaalep, Heiki-Jaan 2018. Statistika koht keelemudelis. [Place of Statistics in a Language Model.] *Keel ja Kirjandus*, No. 8–9, pp. 713–727.

Kaalep, Heiki-Jaan & Vaino, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. [The Complete Morphological Analysis of a Text in the Toolbox of a Linguist.] In: Tiit Hennoste (ed.). *Arvutuslingvistikalt inimesele.* [From Computer Linguistics to People.] (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1.) Tartu: Tartu Ülikool, pp. 87–99.

Kask, Arnold 1984. *Eesti murded ja kirjakeel*. [Estonian Dialects and the Standard Language.] (Eesti NSV Teaduste Akadeemia Emakeele Seltsi toimetised 16.) Tallinn: Valgus.

Kettunen, Lauri 1962. *Eestin kielen äännehistoria. 3. Trükk*. [Phonetic History of the Estonian Language. Third edition.] (Suomalaisen Kirjallisuuden Seuran toimituksia 156.) Helsinki: Suomalaisen Kirjallisuuden Seura.

Kohler, Kaja 2003. Erwerb der frühen Verbmorphologie im Estnischen. Dissertation. Potsdam: Universität Potsdam.

Muuk, Elmar 1927. *Eesti keeleõpetus. Kd I. Hääliku- ja vormiõpetus.* [Teaching the Estonian Language. Vol 1. Vowels and Forms.] (Akadeemilise Emakeele Seltsi toimetised XII.) Tartu: Eesti Kirjanduse Seltsi Kirjastus.

Müller, Georg 2007. Jutluseraamat. [Sermon Book.] (Eesti mõttelugu 78.) Tartu: Ilmamaa.

Prillop, Külli 2003. Georg Mülleri teisenev keel. [The Changing Language of Georg Müller.] In: Valve-Liivi Kingisepp (ed.). *Vana kirjakeel ühendab.* [Old Written Language Connects.] (Tartu Ülikooli eesti keele õppetooli toimetised 24.) Tartu, pp. 242–260.

Prillop, Külli 2004. Kuidas märksõnastada vanu eestikeelseid tekste? [How to Tag Old Estonian Texts?] *Keel ja Kirjandus*, No. 2, pp. 90–99.

Prillop, Külli 2020. Lühike, pikk ja ülipikk häälik eesti kirjakeele ajaloos. [Short, Long and Overlong Vowel in the History of the Estonian Language.] In: Mati Erelt (ed.). *Emakeele Seltsi aastaraamat* 65. Tallinn: Teaduste Akadeemia Kirjastus, pp. 164–191.

Salasoo, Tiiu 1995. Morfoloogiliste tunnuste esmakasutus ühe lapse arenevas keeles. [The First Use of Morphological Characteristics in the Developing Language of a Child.] *Keel ja Kirjandus*, No. 4, pp. 239–252.

Vihman, Marilyn May & Vija, Maigi 2008. The Acquisition of Verbal Inflection in Estonian: Two Case Studies. In: Natalia Gagarina, Insa Gülzow (eds.). *The Acquisition of Verbs and their Grammar: The Effect of Particular Languages*. (Studies in Theoretical Psycholinguistics 33). Dordrecht: Springer, pp. 263–295.

Viht, Annika & Habicht, Külli 2019. *Eesti keele sõnamuutmine.* [Estonian Inflection.] (Eesti keele varamu IV.) Tartu: Tartu Ülikooli Kirjastus.

Viks, Ülle 1992. *Väike vormisõnastik. Kd 1. Sissejuhatus & grammatika*. [A Concise Morphological Dictionary of Estonian. First Volume. Introduction and Grammar.] Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn.

Wiedemann, Ferdinand Johann 1875. *Grammatik der Estnischen Sprache*. L'Académie Impériale des sciences. St. Petersbourg.

Heiki-Jaan Kaalep is an associate professor of language technology in the Institute of Computer Science at the University of Tartu. His main areas of research are morphology of Estonian (both theoretical and computational) and corpus linguistics (creating and tagging corpora, as well as using them as a basis for research). He has participated in creating software which is used in commercial spell checkers, as well as in procesing texts for linguistic research.

heiki-jaan.kaalep@ut.ee