Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 348–359
https://doi.org/10.22364/bjmc.2022.10.3.08

Looking into Estonian Syllabification

Heiki-Jaan KAALEP

University of Tartu, Ülikooli 18, 50090 Tartu, Estonia

heiki-jaan.kaalep@ut.ee

Abstract. The paper presents a detailed description of an algorithm of Estonian syllabification. The paper has a dual goal: justify the algorithm with references to phonology, and make it robust enough for using on real-life texts. The algorithm is presented as a commented set of finite state transducer expressions.

Keywords: Estonian, syllable, finite state transducers

1 Introduction

The current paper attempts to shed some light into Estonian syllabification and provide some speculative answers to syllable-related questions, while describing an approach to mark syllables in real texts. The approach makes explicit some assumptions, follows them, and provides the research community a test-bed for evaluating the borderline cases and debatable proposals. The goal of the paper is to make the problematic cases explicit and hope that others will challenge the paper's claims.

Special attention is devoted to determining the position of a syllable border within a vowel sequence and a consonant cluster. The proposed solution deviates from standard treatment of Estonian syllabification, and thus also from previous computational implementations of it.

The paper presents a rule-based syllabifier for Estonian that does not fail miserably on real texts, but when it does, its behaviour is at least explainable.

Both the software and corpus tagged with it are freely available[1].

2 Syllabification - simple, but vague

Human speech is a sequence of sounds, produced by a series of articulatory gestures by speech organs making the vocal tract smoothly alternate between more open and more

[1] https://cl.ut.ee/korpused/silbikorpus/

closed positions, resulting correspondingly in more and less sonorous sounds. This alternation is called the sequence of syllables: the syllable nuclei are where the sonority is at its maximum, and the syllable boundaries occur where the sonority is at its minimum. The division of speech into syllables seems perceptually simple for both speakers and listeners, and intuitively easy to capture. Notably, the very first sound-based writing systems were syllabic, e.g. Sumerian cuneiform that developed from Sumerian pictographs, and Brahmi, the basis for many Indian scripts.

According to (Räsänen et al., 2018), "Syllables are often considered to be central to infant and adult speech perception. Many theories and behavioral studies on early language acquisition are also based on syllable-level representations of spoken language."

Important aspects of Estonian - gradation and inflection - are traditionally described, using syllable as an explanatory notion.

Attempts to get hold of the syllables, however, turn out to be nontrivial. Räsänen et al. (2018) state that "despite this belief in the importance of syllables in language acquisition, adult-like syllabification depends on knowledge of phonological structure and of the specific language being used." Moreover, Price (1980) reports a number of spoken utterance syllable counting experiments with adult native speaker-hearers where the latter are not unanimous, consistent or confident in their judgements. It appears that despite the intuitive simplicity of syllabification, there are no uniform phonetic and/or phonological criteria for defining the syllable (Eek, 2008, p. 47-48). Hall (2006, p. 394) explains the difficulties of understanding syllabification by saying that "A syllable is an abstract phonological unit that is visible to phonological patterns such as stress assignment, minimal word requirements, allomorph selection, etc. [...] If syllabicity is a construct of the native speaker's mind, the presence of a syllable cannot be verified through strictly phonetic means, nor by a linguist's ear, contrary to assumptions that crop up frequently in the literature."

2.1 Allomorphs as indicators

How can we read what the native speaker's mind thinks about syllables? We should look for traces of the mind at action, e.g. at the way different words get inflected. The current paper shares Hall's view and relies on allomorph selection as a criterion when choosing between alternative ways to syllabify Estonian words.

Estonian is a flective language and words belong to different inflectional classes, based on their phonological and derivational structures, including the syllabic one. For a speaker, it is important to determine this syllabic structure in order to choose the right allomorphs for inflecting words correctly with regard to the language norm.

For example, it is the number of syllables that decides which vowel will be appended when forming the singular genitive case form of names ending in *-ing*: two syllables - choose *u* (*Tering - Teringu*), otherwise - choose *i* (*Tereping - Terepingi*). Now, one may turn this simple example around: having met the form *Teringu*, a linguist may conclude that the speaker counted two syllables in *Tering*.

Pronunciation determines inflectional type, so one can look at the way certain words are inflected, and deduce how the words must have been syllabified. When trying to determine whether two vowels form a diphthong or have a hiatus between them, one may look at how words like *video*, *Ikea* or *kinoa* are inflected. If there is a hiatus, then

the word has three syllables, and the allomorph for singular partitive case is *t*; if the word ends with a diphthong, then it has two syllables, stress being on the last, and the allomorph is *d*. A search for forms with the alternative allomorphs in etTenTen corpus[2] reveals that *video* has three syllables, *kinoa* has two (although some speakers think it has three), and *Ikea* seems to have three, although some speakers think it has two. Checking other similar words will eventually indicate that certain vowels tend to have a hiatus between them and certain vowels do not. The resulting knowledge is described in more detail in section 5.2

3 Estonian syllabification principles

Syllables are defined via speech-related terms, but syllabification software is typically used for operating on written texts, and this requires a few words on Estonian orthography which is guided by phonemic principle that each grapheme corresponds to one phoneme. Rule of thumb is that a short consonant or vowel is written as one letter, a long (or extra long) one is written as double letters; thus a long phoneme is considered to be a sum of two short ones. Diphthongs and consonant clusters are written as combinations of single letters, regardless of whether they are pronounced short or long. Stress is not marked in writing.

This convention makes it easier to formulate the syllabification rules; they are similar to Finnish (which has the same orthographic principles as Estonian), and may be considered much simpler than those of English. According to a grammar handbook (Erelt et al., 2007, p. 58), lexicographers (Lind and Viks, 1994, p. 59-60) and the software they use (*EKI syllabification*, n.d.), the rules are the following:

1. Mark the boundary between the components of a compound word (*vale-start*, *eba-usk*), as well as after the stem followed by a vowel-initial derivational suffix (*ego-ism*): a syllable may not extend over the boundaries separating these.
2. A syllable boundary is in front of the consonant that is followed by a vowel, provided the consonant is not part of a word-initial consonant cluster. E.g. *trans-port*, *tul-la*.
3. A syllable boundary may also be between vowels.
   - 3.1 In a sequence of three vowels, the boundary is before the last vowel, unless the last two form a double vowel; in that case, the boundary is in front of these. (A double vowel also indicates that the syllable bears stress.)
   - 3.2 If there is a sequence of two different vowels, then the syllable boundary is between them in case these vowels do not form a diphthong.
4. Foreign loanwords are syllabified the same way as the native Estonian ones. If a foreign loanword contains a common suffix (e.g. *graaf*) or prefix (e.g. *des*), then a syllable boundary should preferably be at this junction, e.g. *foto-graaf*, *des-in-fekt-si-oon*.
5. Foreign names retain their original orthography in Estonian texts, violating the Estonian grapheme-to-phoneme correspondence. When marking syllable boundaries in these names one should try not to separate graphemes that collectively stand for one phoneme, e.g. *Shakes-peare* (not *Sha-kes-pea-re*).

[2] `https://www.keeleveeb.ee/`

4 Related work

Since 1990ies, lexicographers at the Institute of the Estonian Language (IEL) have been using computational tools to generate information on inflectional types in lexicon entries, and syllabification has been part of their tool set (Lind and Viks, 1994). IEL's syllabifying software source code (C++, Pascal) is also available online[3], as well as a list of 221328 words[4], syllabified with it. The software follows the standard syllabification principles listed in section 3, and sets additional restrictions on diphthongs: sometimes a vowel pair forms a diphtong in a stressed syllable, but not in a syllable following a stressed one. The list and treatment of these pairs, however, differ slightly from the current treatment in section 5.2.

The standard syllabification recipe (section 3) refers to diphthongs, so in order to syllabify, one should be aware of potential and impossible diphthongs. Their treatment is geared towards nativity (core vocabulary vs loans) in books on Estonian phonology (Viitso, 2003, p. 21-22), (Asu-Garcia et al., 2016, p. 55-61). Omission of *öu* from the list of possible diphthongs by both sources is likely a mistake, as evidenced by the existence of a native Estonian verb *möurama* (yell, roar).

One of the mark-up layers of the Phonetic Corpus of Estonian Spontaneous Speech (Lippus et al., 2021) contains syllable boundaries. The PRAAT scripts[5] used for setting them are dependent on the format of the corpus. The scripts follow the standard recipe of section 3, but complex consonant clusters are occasionally divided between syllables differently, being more like the output of the rules in section 5.3.

Finite State Transducers (FST) present a convenient way of formalizing phonological and morphological regularities into practical software. The rich experience in computational linguistics with FST has been condensed in (Beesley and Karttunen, 2003) which may serve also as a course book. The FST approach of the present paper got inspiration from previous work on Finnish prosody by L. Karttunen (2006) and English syllabification by M. Hulden (2005).

5 Current solution (SYLL)

After having marked word boundaries in compound words with the morphological analyser by Filosoft[6], a FST script[7] will mark the syllable boundaries with dots ".". A word receives exactly one syllabification, e.g. *vi.de.o*, *i.ke.a*, *ki.noa*. The script is a collection of FST calculus expressions in Xerox finite-state morphology formalism (Beesley and Karttunen, 2003), compiled into a transducer by hfst-xfst, a tool belonging to the set of Helsinki Finite-State Technology – HFST (Lindén et al., 2011) tools. The resulting transducer is in HFST format, and hfst-fst2fst tool can transform it into other formats: OpenFST, Stuttgart FST, or Foma.

[3] `https://www.eki.ee/tarkvara/silbitus/#algoritm`
[4] `https://www.eki.ee/tarkvara/wordlist/silbitus.dic`
[5] `https://gitlab.keeleressursid.ee/partel/plugin_PhonCorpTools`
[6] `https://github.com/Filosoft/vabamorf`
[7] `https://cl.ut.ee/korpused/silbikorpus/silbita.xfscript`

SYLL assumes that the input conforms to standard Estonian orthography. A foreign word following the orthographic conventions of its original language will most likely be syllabified incorrectly. In case the input originates from chatrooms, the set of vowels has to include also *2*, *6*, and *8* for *ä*, *õ* and *ö*.

The main rule of Estonian (as well as Finnish) syllabification inserts a boundary in front of the last consonant followed by a vowel (Beesley and Karttunen, 2003, p. 74):

```
define V [a|e|i|o|u|õ|ä|ö|ü];
define C [b|c|d|f|g|h|j|k|l|m|n|p|q|r|s|š|z|ž|t|v|w|x|y];
define MarkCVSyll [ C* V+ C* @-> ... "." || _ C V ];
```

5.1 Vowel sequences

Estonian has nine vowels, positioned on the vowel chart as depicted on figure 1.

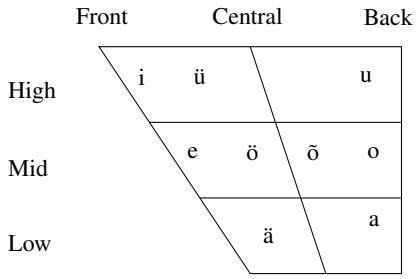

Fig. 1. Estonian vowel chart

The following vowel classes are used for handling vowel sequences.

```
define Vh [i|u|ü];   # high vowels
define Vr [u|ü|o|ö]; # round vowels
# long vowels
define V2 [a^2 | e^2 | i^2 | o^2 | u^2 | õ^2 | ä^2 | ö^2 | ü^2];
```

Inserting syllable boundaries into vowel sequences longer than two is simple: long vowels are written as double vowels, and always form a syllable nucleus. The context condition is matched on the output side, so when the transducer checks the context to decide whether it should insert a boundary symbol, it is aware of the boundary symbols it has inserted already. A double vowel may start or end a three-vowel sequence:

```
define MarkV2plusV [[. .] -> "." \/ V2 _ V] ;
define MarkVplusV2 [[. .] -> "." \/ V _ V2] ;
```

High vowels are also called closed vowels. They are either between two syllable nuclei, or act as the latter part of a diphthong. The last rule inserting a syllable boundary after the first two vowels is an elsewhere rule.

```
define MarkVVhplusV [[. .] -> "." \/ V Vh _ V] ;
define MarkVplusVVh [[. .] -> "." \/ V _ V Vh] ;
define MarkVVplusV  [[. .] -> "." \/ V V _ V] ;
```

Composition of these rules yields a transducer that syllabifies *uu.e*, *li.aan*, *lau.a*, *ge.oid*:

```
define MarkVVV [MarkV2plusV .o. MarkVplusV2 .o.
               MarkVVhplusV .o. MarkVplusVVh .o. MarkVVplusV] ;
```

5.2 Diphthongs

It is not quite clear how to decide that two vowels form a diphthong, i.e. are pronounced as a single unit without a hiatus between. It is known that *õ*, *ä*, *ö* and *ü* cannot be a second vowel of a diphthong, and that a combination of a front vowel *ä*, *ö* or *ü* with its back-vowel counterpart *a*, *o* or *u* cannot form a diphthong either (thus *äa*, *öo* and *üu* cannot be diphthongs) (Asu-Garcia et al., 2016, p. 55).

(*EKI syllabification*, n.d.) notes that if a potential diphthong is positioned after a stressed syllable, then its second vowel might be pronounced with an hiatus, i.e. the vowel pair does not form a diphthong after all, e.g. *vi-de-o*.

Table 1 lists the 37 Estonian diphthongs, positioning them so as to mimic the position of each initial vowel in the vowel chart, and ordering the second vowels according to their positions in that chart (starting from high front and ending with low back vowels). The second vowels that do not require a hiatus in front of them in a stress-following syllable are printed in bold. The proposal here is that the potential of a vowel to form a diphthong in a stress-following syllable is related to its position in the vowel chart: a non-high rounded or low vowel (*ö*, *õ*, *o*, *ä*, *a*) may append any possible vowel, while the appending potential of others is limited to *i*, and high front vowel *i* itself cannot append any vowel at all. (Diphthong *öu* is blocked by a separate rule that prohibits round vowels from following *ö*; see below.)

Table 1. Estonian diphthongs

i + u e o a	**ü** + **i** e o a		**u** + **i** e o a
e + **i** u o a	**ö** + **i u e a**	**õ** + **i u e o a**	**o** + **i u e a**
	ä + **i u e o**		**a** + **i u e o**

The way FST expressions are used for defining non-diphthongs is by telling them to insert syllable boundaries between vowels. Some of the non-diphthongs are covered by several definitions. For example, *iä* cannot be a diphthong in a stress-following syllable, because *ä* cannot be the final part of any diphthong, and also because a high vowel cannot be followed by a non-high vowel in that context. The definitions attempt to follow general phonological principles, and a special instance may well be covered by several of them.

Prohibit an accented letter to be the final vowel of a diphthong:

```
define NonDiphVõäöü [[. .] -> "."
  || [V - õ] _ õ , [V - ä] _ ä , [V - ö] _ ö , [V - ü] _ ü];
```

Prohibit a diphthong of front vowel *ä*, *ö* or *ü* followed by its back-vowel counterpart *a*, *o* or *u*:

```
define NonDiphäaöoüu [[. .] -> "." || ä _ a , ö _ o , ü _ u] ;
```

In a non-initial consonant-following syllable (which we take as a proxy for a stress-following syllable), prohibit diphthong of high vowel followed by a non-high vowel:

```
define NonDiphHigh [[. .] -> "." || "." C+ Vh _ [V - Vh]];
```

In view of table 1, prohibiting a front vowel *i* or *e* followed by a non-front vowel would be:

```
define NonDiphFr [[. .] -> "."
         || "." C+ e _ [V - e - i] ,
            "." C+ i _ [V - i]];
```

However, given that *eu* occurs frequently in non-initial stressed syllables of loanwords, it is tempting from an engineering point of view to exclude it from a general definition of non-diphthongs (by restricting the context condition):

```
define NonDiphFr [[. .] -> "."
         || "." C+ e _ [V - e - Vh] ,
            "." C+ i _ [V - i]];
```

and add separate *ad hoc* definitions for frequent foreign word endings, e.g. *-eum* or *-eus*, while excluding frequent words with altered pronunciation:

```
define MarkEU [[. .] -> "." || "." C+ e _ u (".") [m|s]] .o.
               ["."  -> ""  ||  m u u (".") s e _ u ("."),
                     p e t ("."") r o o ("."") l e _ u ("."") m];
```

An additional improvement comes from the observation that the third syllable of a word, being the first syllable of the second foot, might carry stress, and thus the condition prohibiting a diphthong be revoked; this concerns loanwords from Latin:

```
define DiphtongInFinal ["." -> "" ||
  V "." C+ V "." C e _ u ("."") [m|s] ,  V "." C+ V "." C e _ a];
```

Front round vowel *ö* followed by a round vowel is prohibited:

```
define NonDiphRound [[. .] -> "." ||  "." C+ ö _ [Vr - ö]];
```

To sum it up, non-diphthongs get marked by a composition of the individual rules:

```
define MarkNonDi [NonDiphVõäöü .o. NonDiphäaöoüu .o. NonDiphHigh
 .o. NonDiphFr .o. MarkEU .o. DiphtongInFinal .o. NonDiphRound];
```

5.3 Consonant clusters

Although the main rule for Estonian and Finnish syllabification states that the syllable boundary is in front of the last consonant of a consonant cluster preceding a vowel, it is hard to conceive how this could be possible in case of *ekst-ra* and *abst-rak-ti*: the sole phonologically plausible syllabification seems to be *eks-tra* and *abs-trak-ti*, as noted by Erelt et al. (1995, p. 111), Hint (1998, p. 107-108) and Hulden (2005).

The plausible syllabification here conforms to the sonority principle which states that the syllable boundary is at the sound with the vocal tract more closed than at the neighbouring sounds. The fricative *s* is more sonorous than a voiceless stop (depicted as *k*, *p*, *t*, *g*, *b* or *d* in Estonian orthography), so in a sequence of stop-fricative-stop the boundary has to be either before or after the fricative. Guided by the principle that potential word-internal syllable onsets should not be different from observable word-initial onsets, and word-internal syllable codas not different from what we see at word ends, the preferable syllable boundary would be after the fricative: native words ending with *ks*, *ps* or *ts* are common in Estonian, while only loanwords start with *sk*, *sp* or *st*.

The more a consonant cluster following and preceding a vowel contains consonants, and the more of them are voiceless obstruents (i.e. stops or fricatives), the more likely it is that the syllable boundary should be marked in front of a stop which is not the last consonant in the cluster. Currently, a syllable boundary is inserted between two different stops, and in front of a stop following *s*, provided the context:

```
define Stop [g|b|d|k|p|t] ;
define ComplexConsCluster [[. .] -> "."
        || V C*   [Stop - k] _ k r    C* V ,
           V C*   [Stop - p] _ p r    C* V ,
           V C*   [Stop - t] _ t r    C* V ,
           V C*            x _ Stop   C* V ,
           V C*       Stop s _ Stop   C* V ,
           V C*            s _ Stop r C* V ,
           V C* [C - Stop] s _ Stop s C* V ];
```

This results in *abs-trakt*, *ex-press*, *dok-triin*.

5.4 Loanword orthography conventions

5.4.1 *i* as a glide Loanwords often retain their original written form, but the pronunciation deviates from it. A recurring example is *i* following a consonant and preceding a vowel in a syllable that either follows an extra long syllable, or precedes it (*staadion*, *sotsioloog*). According to the rules prohibiting diphthongs, and *io* being in a post-stressed syllable, it should be split by a hiatus, i.e. contain a syllable boundary. However, when bordering with an extra long syllable, *i* becomes a glide *j*, with *o* remaining as the sole vowel of the syllable nucleus. The same process of *i* becoming a glide happens when a loanword becomes part of core vocabulary, and is adjusted to the Estonian way of speaking. Children learning to write also often make an error in such contexts, writing *j* instead of *i*. In order to syllabify correctly, the FST script must have additional rules for removing the syllable boundaries, previously inserted by the

diphthong-prohibiting rules. The rules would have rather complex context conditions, sometimes referring to vowel classes, sometimes to typical suffixes or word endings. Abbreviated versions of the rules that would result in *sti.pen.dium*, *sot.sio.loog* and *a.kor.dion*, would be:

```
define FixIU      ["." -> "" || V V "." C+ i _ u [\V | .#.]];
define FixIOC     ["." -> "" ||     "." C+ i _ o C* "." C V V];
define FixIOCWord ["." -> "" || a "." k o r "." d i _ o n];
define FixI [FixIU .o. FixIOC .o. FixIOCWord];
```

5.4.2 ***y* and *w*** Letters *y* and *w* do not belong in the Estonian alphabet, but they still appear in foreign words and in chatroom texts. They may denote vowels or consonants, depending on their context: if they are alone between two vowels, then they are likely consonants, and the syllabification would be accordingly:

```
define MarkCVSyllYW [[. .] -> "."
                  || V _ w V ,  V _ y V ] .o.
    [[. .] -> "." || V C* _ C y [\V | .#.] , y C* _ C [V | y] ] ;
```

5.4.3 Stress There are loanwords where the stress falls on a vowel following another vowel, thus a syllable boundary must be inserted in front of the stressed one. A short example list is:

```
define NonDiphtongInLoanWords [[. .] -> "."
         ||  d e m i _ u r g , p a _ e l l a , s e _ a n s s];
```

There are also words where the stress pattern overrules the assumption that a certain syllable is not stressed, or orthography hides the fact that a syllable is extra long. It means that for a list of words, a syllable boundary separating vowels must be removed:

```
define DiphtongInLoanWords ["." -> ""
    || m a "." r i _ o "." n e (".") t , k o "." r e _ a];
```

5.5 Final composition

The final composition of the syllable boundary marking transducers is the following:

```
define Syllabify [MarkCVSyllYW .o. NonDiphtongInLoanWords .o.
            ComplexConsCluster .o.MarkCVSyll .o. MarkVVV .o.
            MarkNonDi .o. FixI .o. DiphtongInLoanWords];
```

6 Evaluation

6.1 Word list

SYLL was developed and tested, using a list of 221328 syllabified words, provided by the Institute of the Estonian Language[8]. For 96.3% of these, i.e. 213184 words, SYLL

[8] `https://www.eki.ee/tarkvara/wordlist/silbitus.dic`

inserted word component and syllable boundary marks in the same places where the original list had them. For 2941 words, SYLL inserted only some word component boundary differently. This leaves 5203 words with different syllabification. (It is important to note that the authors of the list are modest enough not to state that it is a gold standard of syllabification.)

An important feature in the original word list is that 1019 words have a boundary marked as unsure. Two of these are words that have dual interpretation (*teist* (from you) vs *te-ist* (theist)), the rest being cases of potential diphthongs *eu*, *ia*, *io*, *iu* in an unstressed syllable, as described in section 5.2. SYLL is mistakenly unaware of *te-ist*, but manages to make decisions for all the potential diphthongs. However, at the moment the correctness of these decisions remains untested.

SYLL's consonant cluster rules (section 5.3) result in 1457 words where SYLL marks boundaries differently than the original list.

1026 words contain a sequence of *i* and a vowel that could be classified as an instance of *i* becoming a glide, as described in section 5.4.1.

The remaining list of differences contains 1701 words. The reasons are varied. In some cases, the division of a compound word into its components is different, i.e. a morphological analyzer has made an error; in some cases, the difference is in how to interpret non-Estonian orthography. The most interesting differences are in diphthongs vs vowels with hiatus. Currently, they are waiting to be described in more detail and with explanations.

6.2 Corpus

SYLL was used to syllabify a corpus[9] that contains five different subcorpora. The subcorpora were chosen to represent different types of language: written vs oral (fiction, newspapers and chatrooms vs conversations and child caregiver speech), edited vs unedited (fiction and newspapers vs conversations, chatrooms and child caregiver speech), toddler-oriented vs grown-up oriented (child caregiver speech vs all the other corpora). The aim was to safeguard against bias when estimating the performance of SYLL, and when making Estonian language-specific statistics on syllables.

The corpora provide different challenges for the syllabification algorithm, as they contain tokens that one would not consider to be normal Estonian words in standard orthography. Written genres contain abbreviations and acronyms, as well as foreign names (and other words) written according to their source language orthographic norms, i.e. tokens not following the Estonian spelling, and thus breaking the algorithm's assumptions of grapheme-to-phoneme correspondences. Unedited genres contain letter sequences denoting onomatopoetic sounds in the form of long sequences of vowels or consonants, violating the assumption of how letters normally form a syllable. Chatrooms also contain Estonian words written with letters and symbols which are not part of the standard Estonian orthography, e.g. *y* for *ü*. Finally, the transcribers of conversations and child caretaker speech have sometimes tried to indicate the deviations and unclear pronunciation of real-life speech, thus creating word forms in an alternative orthography.

[9] `https://cl.ut.ee/korpused/silbikorpus/`

For evaluation, 100 randomly selected word types from each corpus were checked manually. The encountered errors (7 altogether) were caused by either the morphological analyser marking word boundaries incorrectly, or because the word was a foreign name. Specifically, two instances involved a missing space between words: *o.lek_uss* (status snake) instead of *o.le kuss* (shut up), and *möö.dun.daas.ta* (lastyear) instead of *möö.dund aas.ta* (last year); one instance was an innovative compound based on English words *land artist*: *lan.dar.tis.ti.le* instead of *land_ar.tis.ti.le*, and one instance was a mis-analysis of an Estonian family name as a compound word: *järv_a.la* (lake area) instead of *jär.va.la*. The remaining three instances were English words *i.ce*, *pa.la.ce* and *spe.ci.al*. Their pronunciation requires syllabification as *ice*, *pa.lace* and *spe.cial*. All the seven incorrectly syllabified word forms occurred only once in the corpus, except *järvala* which occurred twice.

Table 2 gives some details about the corpora: size in word tokens and word types, and an estimation of correctly syllabified word types.

Table 2. Syllabified corpora

Text class	Tokens	Types	Correct types, %
Fiction (Estonian authors)	104,000	26,200	100
Newspaper texts	111,000	33,400	97
Conversations	100,000	12,800	99
Chatrooms	94,000	17,100	97
Child caregiver language	400,000	21,200	100

7 Conclusion and future work

The paper presented a detailed description of syllabification rules of Estonian. Much of the details are due to the various phenomena one encounters in real word forms, which have not been systematically described in grammars. It is possible that these phenomena are uninteresting from a theoretical point of view. However, one can make an informed decision on them only once they have been brought to light. This is what the paper tries to do.

It is possible that an alternative syllabification script would rely on more fine-grained phoneme classes, follow the phonological principles more closely, or implement a new theory. The new implementation could be compared against the current one, the comparison being not limited to existing corpora and word lists, but could be applied also to previously unseen data.

The treatment of stress is naive in the paper, and this is a serious deficiency that should definitely be repaired in the future.

Acknowledgements

This work has been supported by the Centre of Excellence in Estonian Studies (CEES, TK-145).

References

Asu-Garcia, E. L., Lippus, P., Pajusalu, K., Teras, P. (2016). *Eesti keele hääldus*, Tartu Ülikooli Kirjastus.

Beesley, K., Karttunen, L. (2003). *Finite State Morphology*, CSLI studies in computational linguistics: Center for the Study of Language and Information, CSLI Publications.

Eek, A. (2008). *Eesti keele foneetika I*, Tallinna Tehnikaülikooli Kirjastus, Tallinn.

EKI syllabification (n.d.). `https://www.eki.ee/tarkvara/silbitus/` [Accessed: May 2022].

Erelt, M., Erelt, T., Ross, K. (2007). *Eesti keele käsiraamat*, EKI, Tallinn :.

Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. (1995). *Eesti keele grammatika I. Morfoloogia. Sõnamoodustus.*, Eesti Teaduste Akadeemia Eesti Keele Instituut, Tallinn.

Hall, N. (2006). Cross-linguistic patterns of vowel intrusion, *Phonology* **23**(3), 387–429.

Hint, M. (1998). *Häälikutest sõnadeni: eesti keele häälikusüsteem üldkeeleteaduslikul taustal*, Eesti Keele Sihtasutus.

Hulden, M. (2005). Finite-state syllabification, *International Workshop on Finite-State Methods and Natural Language Processing*, Springer, pp. 86–96.

Karttunen, L. (2006). A finite-state approximation of optimality theory: The case of finnish prosody, *Proceedings of the 5th International Conference on Advances in Natural Language Processing*, FinTAL'06, Springer-Verlag, Berlin, Heidelberg, p. 4–15.
`https://doi.org/10.1007/11816508_3`

Lind, P., Viks, Ü. (1994). Mall - the tool of a linguist, *in* Viks, Ü. (ed.), *Automatic Morphology of Estonian*, Estonian Academy of Sciences, Institute of the Estonian Language, pp. 49–64.

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T. A., Silfverberg, M. (2011). Hfst—framework for compiling and applying morphologies, *International Workshop on Systems and Frameworks for Computational Morphology*, Springer, pp. 67–85.

Lippus, P., Aare, K., Malmi, A., Tuisk, T., Teras, P. (2021). Phonetic Corpus of Estonian Spontaneous Speech v1.2.
`https://datadoi.ee/handle/33/351`

Price, P. J. (1980). Sonority and syllabicity: Acoustic correlates of perception, *Phonetica* **37**, 327–343.

Räsänen, O., Doyle, G., Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units, *Cognition* **171**, 130–150.
`https://www.sciencedirect.com/science/article/pii/S0010027717302901`

Viitso, T.-R. (2003). Phonology, morphology and word formation, *in* Erelt, M. (ed.), *Estonian Language*, Vol. 1 of *Linguistica Uralica. Supplementary Series*, Estonian Academy Publishers, pp. 9–92.

Received August 15, 2022 , accepted August 16, 2022