Parallel Forms in Estonian Finite State Morphology

Heiki-Jaan Kaalep
University of Tartu
Department of Language Technology

Heiki-Jaan.Kaalep@ut.ee

Abstract

Parallel forms are two or more synonymous forms that convey an identical set of morpho-syntactic categories in a paradigm cell of a word. They deserve attention from a theoretical linguistic, as well as from a computational point of view. How do humans know which form to choose, and how should this preference be modelled computationally? The paper gives an overview of parallel forms in Estonian and discusses reasons for surface form variation. A considerable part of the article is dedicated to a simplified, but still technically detailed example of handling parallel plural partitive forms, one of which is more common, and the other a rarer form. An example is used to explicate the proposed method of handling parallel forms in finite state morphology, coupled with considerations of their preferencial choice. The method involves using a combination of two-level rules as a way of controlling the combinatorial explosion of continuation lexicons. The design has been implemented to fully cover the inflectional morphology of Estonian.

Kokkuvõte

Rööpvormid on ühe sõna erinevad muutevormid, millel on sama grammatiline tähendus. Nad väärivad tähelepanu nii teoreetilise kui ka arvutilingvistika poolt. Kuidas teavad inimesed, millist vormi valida, ja kuidas seda teadmist modelleerida? Artikkel annab ülevaate rööpvormidest eesti keeles ja arutleb nende olemasolu põhjuste üle. Suur osa artiklist on pühendatud lihtsustatud, kuid siiski

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details:
http://creativecommons.org/licenses/by/4.0/

Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018), page
Helsinki, Finland, January 8–9, 2018. ©2018 Association for Computational Linguistics

tehniliselt detailsele näitele, kuidas käsitleda mitmuse osastava rööpvorme, millest seejuures üks on tavaline ja teine haruldane. Näite abil selgitatakse, kuidas rööpvorme ja nende valikueelistusi saaks lõplike muundurite morfoloogias käsitleda. Pakutav meetod, mida on kasutatud kogu eesti keele sõnamuutuse modelleerimiseks, hõlmab kahetasemeliste reeglite kasutamist, millega piiratakse jätkuleksikonide kombinatoorset plahvatust.

1 Introduction

It is possible that a word paradigm cell is filled by two or more synonymous forms that realise the same set of morpho-syntactic categories. These alternative realizations are called *doublets* in the German linguistic tradition (Mörth and Dressler, 2014), *parallel forms* (Raadik, 2013) in the Estonian tradition (which will be followed in the current paper), and *overabundant forms* (Thornton, 2012).

Computationally, parallel forms pose no problem for analysis, as different surface forms are mapped to the same bundle of morpho-syntactic categories. However, per synthesis the situation is reversed, as one must decide which form to generate, and how does one choose in the case of one-to-many mapping? Traditionally, analysis has been the main application, be it in the context of spell-checking or information retrieval, thus instances when one surface form maps to several lemmas and/or morpho-syntactic categories have received much attention as the problem of ambiguity. This tradition may explain why parallel forms have received less attention, despite them posing a challenge in terms of explaining human language and the generation of rule-based machine translation.

2 Reasons for surface form variation

Parallel forms may occur due to different dialects than the written norm is based upon, or due to language change that is currently happening

The first scenario implies that some speakers have different intuitions regarding word inflection, because they have different dialectal backgrounds, e.g. the Standard German *Park-s* and the Swiss-German *Pärke* (parks). The parallel forms are the result of defining the norm in a liberal manner, thus allowing the inflectional systems of different dialects to co-exist under a common normative umbrella.

The second scenario implies that the speakers have different intuitions regarding word inflections, because although they all share a language variety (dialect), it changes over time. The existence of parallel forms in this case is at odds with the

principles of communication: the speaker can only choose one of the parallel forms at a given time, and one would expect the other forms to eventually fall into disuse. The existence of parallel forms in this scenario would indicate that the language has not finished its particular process of change.

It is not easy to decide which scenario is applicable to a given language: co-existence of dialects or language change. The question may be answered by investigating how new and rare words are inflected.

In the case of vigorous dialects, when speakers see a new word, they will first designate it to several inflectional classes (based on their dialect-specific intuition), and later, after mutual communication, may arrive at an agreement on a single acceptable, norm-adhering inflectional class.

In the case of language change, the speakers' initial intuitions about inflecting a new word are uniform. It is also expected that speakers abandon the previous inflectional classes of old words so that they join new ones.

In Estonia, the influence of dialects on morphology has become marginal.

When confronted with a rare or previously unseen word, speakers of Estonian immediately exhibit a remarkable consensus about what is the generally accepted (i.e. normal, correct) way of forming its inflectional forms. Their lack of disagreement is noteworthy, because in terms of an unseen word, the speakers could not have discussed its inflectional class beforehand.

Moreover, instances of actual negotiations about the inflectional class of a rare or new word (e.g. a foreign name) are virtually absent in everyday communication.

Evidence from a 270 million token corpus collected in 2013 from the internet called etTenTen [1] shows that there is actually no need for such negotiations: there is almost no variety in the choice of the inflectional class of a new word; typing errors account for a much larger amount of variation in word forms than misclassifications into alternative inflectional classes.

The vocabulary used in the corpus comprises 7.5 million wordforms. An Estonian morphological analyser [2], relying only on its lexicon and algorithm of productive derivation and compounding (thus using no guessing) classified 2.7 million of these wordforms as unknown. Manual check of 2,700 (0.1%) of the unknowns revealed that only 10 wordforms were inflected incorrectly (resulting from inflectional misclassification by the writer), while 300 contained a typing error. The rest of the unknown words were actually instances of incorrect punctuation (typically resulting in concatenating several words), proper names, foreign language words, and true neologisms, not showing variation in inflectional classes.

[1] `http://downloads.sketchengine.co.uk/ettenten13.processed.prevert.xz`, searchable at `http://www.keeleveeb.ee`

[2] `https://github.com/Filosoft/vabamorf`

However, at closer inspection one may observe a few telling instances when Estonians do have problems in deciding what the correct inflectional class of a word is. Those instances fall into two scenarios. In the first scenario, the choice is between an exceptional, old, unproductive inflectional class versus a regular, productive one, and involves cases when an old word has become rare and thus its exceptional inflectional behaviour cannot be remembered by everyone, or when an Estonian family name coincides with a common noun belonging to an exceptional inflectional class (e.g. in English the plural of the family name *Foreman* is *Foremans*, not *Foremen*). In the second scenario, the choice is between two productive classes, and happens when a new word or proper noun has extra-morphological properties that belong to orthogonal categories (e.g. phonetic properties and wordiness) that incidentally predict different productive class memberships. For example, *Breivik* is a foreign name that appeared in Estonian texts only in 2011. Being disyllabic and ending in *-ik*, it should phonetically belong to the class of *u*-ending singular genitives (*Breiviku*). Being a new and foreign word, it is an out-of-vocabulary word w.r.t. conventional Estonian, and thus should belong to the class of *i*-ending singular genitives (*Breiviki*). According to et-TenTen, 75% of the 400 mentions are *Breiviki*, and 25% *Breiviku*. The other possible stem vowels, *a* and *e*, are never used.

3 Parallel forms and finite state morphology

(Beesley and Karttunen, 2003, p. 300–310) used English plurals as a convenient example of parallel forms and showed elegant ways of dealing with them. They differentiated *overriding plurals*, i.e. situations when an irregular form substitutes the regular one, and *extra plurals*, i.e. situations when the regular forms also remain in use, without attempting to differentiate the forms according to their usage preference. Indeed, adding an extra lexical tag (e.g. *Use/Rare*) to the lexical string via an appropriate continuation lexicon would be very simple. The sole formal issue would be the multiplication of continuation lexicons, as this extra tag would effectively make otherwise similar inflectional classes formally different.

In the case of many continuation lexicons (each embodying an inflectional class), and some slots in the paradigm having multiple realisations, independent of the inflectional class, the number of lexicons explodes, as noted by (Beesley and Karttunen, 2003, p. 302–304).

Nevertheless, it is this approach that has been adopted when describing Finnish and Sami [3], in both cases it was assumed that parallel forms are due to dialects. In each language, non-preferred forms are marked with +Use/NG ("no generation"). The

[3]See /fin and /sme in `https://victorio.uit.no/langtech/trunk/langs/`

motivation came from rule-based machine translation, where one needs to synthesise only one wordform per one paradigm cell of a word (Antonsen et al., 2016).

By default, an inflectional type (represented via a cascade of continuation lexicons) is such that every paradigm slot has exactly one surface realisation. If a slot has possibly two realisations, then this inflectional type bifurcates: in addition to the old one with a unique realisation, there will be a new one with one parallel form. Upon adding preference information, this new inflectional type will bifurcate again: one version will be with form A preferred, the other with form B preferred. One inflectional type became three. If there happens to be one more paradigm cell that can have two realisations, then each of the previous inflectional types will have three more variants: thus a single inflectional type has become nine.

This attempt to make the description more informative has resulted in an inflated and less general description, and this is not a desirable result.

An alternative method would be to use a filtering mechanism that first defines a separate list of individual word forms and then use *Priority union*, following (Beesley and Karttunen, 2003, p. 306-309), or (Pruulmann-Vengerfeldt, 2010) for Estonian. The downside of this is that some surface forms of a word become decoupled from the dictionary headword in the stem lexicon, which in turn creates description difficulties when the headwords are homographs with different inflection patterns; there are more than 400 homographic words of this type in the Estonian lexicon in the Giellatekno repository [4].

4 Estonian

Estonian is a Fenno-Ugric language. It is closely related to Finnish, although a speaker of only Estonian and one of only Finnish are unlikely to understand each other. Finnish has retained its original nature more than Estonian, which has lost vowel harmony and moved from an agglutinative language towards a flective one. A specific regular difference is that Estonian has lost the last phone from many words it shares with Finnish, which has resulted in it losing some of the regularities of the Finnish inflectional system, and in the adding of innovations as a substitute.

Estonian has 14 cases, both in the singular and plural. Table 1 lists the possible affix variants and stem vowels that must be concatenated to a consonant-ending stem of the declinable word paradigm (stem gradation patterns are not presented). Note that the vowels *a*, *e*, *i*, and *u* are traditionally classified as theme vowels, associated with the stem, and thus not counted as affixes.

[4] `https://victorio.uit.no/langtech/trunk/experiment-langs/est/`

	singular	plural
grammatical cases		
nominative	∅	[∅ a e i u] d
genitive	∅, a, e, i, u	e, [∅ a e i u] te, [∅ a e i u] de
partitive	∅, d, [∅ a e i u] t	id, [∅ a e i u] sid, e, i, u
semantic cases		
illative	∅, a, e, i, u, de, [∅ a e i u] sse	[te de e i u] sse
inessive	[∅ a e i u] s	[te de e i u] s
elative	[∅ a e i u] st	[te de e i u] st
allative	[∅ a e i u] le	[te de e i u] le
adessive	[∅ a e i u] l	[te de e i u] l
ablative	[∅ a e i u] lt	[te de e i u] lt
translative	[∅ a e i u] ks	[te de e i u] ks
terminative	[∅ a e i u] ni	[te de e i u] ni
essive	[∅ a e i u] na	[te de e i u] na
abessive	[∅ a e i u] ta	[te de e i u] ta
comitative	[∅ a e i u] ga	[te de e i u] ga

Table 1: Declinable word affixes and added stem vowels

There are two general implicational patterns (or rules of referral) in the paradigm. The singular genitive stem serves as the base for forming singular semantic cases (except illative ending in ∅ or *-de*), plus the plural nominative; and the plural genitive stem serves as the base for all plural semantic cases (this even applies to otherwise very exceptional words).

5 Parallel forms in Estonian

Normative Estonian linguists favour parallel forms, accepting various realisations of the number and case category in declinable words. According to a normative dictionary (Raadik, 2013) that categorizes all declinations into 26 inflectional types, there is not a single type that does not contain words with several possible realisations per some categories, with two paradigm slots especially likely to have two parallel forms: the singular illative (Sg Ill) and plural partitive (Pl Par). Thirteen types exhibit type-internal parallelism per all plural semantic cases, and seven types exhibit the same parallelism per some words. Eight inflectional types have parallel forms per Pl Par, seven types exhibit total and exhaustive parallelism per Sg Ill, and twelve types exhibit the same parallelism per at least some of their words. Two more types have parallel

haplology forms per Sg Ill. In addition, a word may simultaneously belong to more than one inflectional class, which results in another set of slots with parallel forms. Of the 26, only seven inflectional types (six of them, with a total of 300 words, may be characterized as unproductive inflectional classes) contain no words that also belong to some other inflectional type.

Thus, the language norm assumes that it is very common for a word to have multiple ways of forming a surface form per some morpho-syntactic category bundle.

Usage-wise, the distribution of such forms is very skewed, though. For example, consider the frequency counts of words and their Pl Par forms in etTenTen per *taim* (plant), *luu* (bone) and *kamp* (gang): *taim* 42665, Pl Par *taimi*/*taimesid* 6868/9, *luu* 7060, Pl Par *luid*/*luusid* 699/13, *kamp* 6271, Pl Par *kampu*/*kampasid* 7/22. This example shows a universal trend in languages: if a case form is infrequent, its forming tends to be regular (*kampasid* in this example).

At a very rough approximation, Finnish could be regarded as a previous form of Estonian: some morphological features of Finnish have disappeared from Estonian. There are many words that are similar in both languages (save for the final lost phones in Estonian), and some of these are partly inflected in a way that resembles Finnish. For example, consider the Estonian *taim* (Finnish *taimi* (plant)). The Finnish Pl Par is *taimia*, while the Estonian irregular form is *taimi* (the regular form being *taimesid*). Or consider *luu* (Finnish *luu* (bone)). The Finnish Pl Par is *luita*, while the Estonian irregular form is *luid* (the regular form being *luusid*). Both these words have rather frequent Pl Par forms, which are lagging behind in their journey from the past, not yet caught up by the change. It is these types of words — otherwise regular, but having some exceptional frequent form that is a remnant from the past — that the following treatment addresses.

6 Parallel forms in Estonian FSM

When a word form is analysed or generated, usage info is written out with the grammatical categories of an inflectional form (Figure 1).

Conceptually, usage information is unrelated to the morphological description (morphotactics and morphonology). However, if one wishes to encode it in the lexicon, then — being a characteristic of individual words — it should be attached to the stem entries, and must be propagated from the stem entry to the final wordform.

For inflectional classes where alternative affixes are common, but still applicable only to a subset of individual words, one needs rules for selecting:

1. the inflectional affix (the traditional task of morphology)
2. a tag that indicates whether the word form is rare or common (not a traditional

```
taim+N+Pl+Par            taimi
taim+N+Pl+Par+Use/Rare   taimesid
kamp+N+Pl+Par            kampasid
kamp+N+Pl+Par+Use/Rare   kampu
```

Figure 1: Lexical and surface sides of Pl Par of *taim* (plant) and *kamp* (gang)

task of morphology, but similar to adding usage notes such as *archaic* or *colloquial* to forms in a paradigm table or traditional dictionary)

The affixes concerned are:

1. Per Pl Par: 1.1. *-sid* or ∅ with stem vowel change; 1.2. *-sid* or *-id*
2. Per Sg Ill: *-sse* or ∅ with stem grade strengthening.

The task is to generate forms. We know that per Pl Par, *-sid* is always possible, and per Sg Ill, *-sse* is always possible, and that the alternative form is possible only if a tag indicates so in the stem lexicon. If an alternative form exists, it is also the more common one, unless it has some tag in the stem lexicon indicating otherwise.

The full implementation covering Estonian inflection is available in the Giellatekno repository [5].

In the Giellatekno infrastructure, the standard way of building a transducer begins with modelling morphotactics by concatenating stems and continuation lexicons. The next step is modelling morphophonology by applying two-level rules (Koskenniemi, 1983) to the output of the previous transducer. This step substitutes and removes some symbols, typically abstract phonemes and functional symbols that were previously introduced to provide meaningful context to the rules.

Incidentally, a two-level rule can be used as a filter to prune some paths from a transducer. Every two-level rule says what output side symbol corresponds to which input side symbol, given a certain context. Imagine that the input symbol has only one potential corresponding output symbol. Now if the context does not match, then it is impossible to have this symbol correspondence in the path, and a path cannot be built. In a similar vein, one can define pruning contexts for input symbols that have more than one corresponding output symbol.

The following sections show how lexicons and two-level rules interact to arrive at the results on Figure 1. The exemplary alternation of *-sid* vs ∅ per Pl Par is used to explain the process.

[5] `https://victorio.uit.no/langtech/trunk/experiment-langs/est/`

```
LEXICON NOUNS
          taim+N:taim%>%{pl.i%} EIT ;            ! plant
          kamp+N:kamp%>%{pl.u%}%{rare%} PIIM ;   ! gang

LEXICON EIT        ! 1C with stem vowel e
                  :%{sg.e%} 1C ;

LEXICON PIIM       ! 1C with stem vowel a
                  :%{sg.a%} 1C ;

LEXICON 1C         ! monosyllabic consonant-ending word
                  :%>%{s%}%{i%}%{d%} PL_PAR_VARIANT ;

LEXICON PL_PAR_VARIANT   ! pl partitive may have parallel forms
   +Use/Rare:%{rare%} PL_PARTITIVE ;    ! less used form
            : PL_PARTITIVE ;            ! default form

LEXICON PL_PARTITIVE
          +Pl: PARTITIVE ;

LEXICON PARTITIVE
         +Par: # ;
```

Figure 2: Relevant extracts from `lexc`-lexicons to build pre-`twol` representations per Pl Par of *taim* and *kamp*

```
taim+N         +Pl+Par:taim >{pl.i}      {sg.e}>{s}{i}{d}
taim+N+Use/Rare+Pl+Par:taim >{pl.i}      {sg.e}>{s}{i}{d}{rare}
kamp+N         +Pl+Par:kamp >{pl.u}{rare}{sg.a}>{s}{i}{d}
kamp+N+Use/Rare+Pl+Par:kamp >{pl.u}{rare}{sg.a}>{s}{i}{d}{rare}
```

Figure 3: Pre-`twol` representations per the plural partitive of *taim* and *kamp*

7 Lexicons for parallel forms

Figure 2 gives `lexc` examples per *taim* and *kamp*.

The final desired lexical and surface strings per the parallel forms of these words would be as in Figure 1; the outcome from the `lexc` lexicons are on Figure 3.

If a word has an exceptional ∅-ending short form in addition to the default *sid*-ending Pl Par, i.e. one that is formed by substituting the stem vowel with a different one, then the vowel is explicitly given at the stem entry as `{pl.i}` or `{pl.u}` (see Figure 2).

The symbol `>` denotes a morpheme border; `{s}{i}{d}` constitute *-sid* and the ∅ alternation (see Figure 4).

A speaker of Estonian knows that they can often choose between alternative ways of generating a Pl Par wordform. This knowledge is expressed by the continuation lexicon PL_PAR_VARIANT with two entries: one is the default, and the other the less used form. Which is which, depends on the word and is marked in the stem lexicon: if the parallel form is actually less used that the default one, then the lexicon entry contains the tag `{rare}`. Note that the less used form has `+Use/Rare` on the lexical side, and a corresponding tag `{rare}` (which will not be visible in the final form) on the surface side.

The lexical and surface sides of the transducer path of a single wordform are assembled piece-wise when compiling the lexicon, starting from a stem lexicon and proceeding via a cascade of continuation classes. PL_PAR_VARIANT bifurcates the path, one of them containing `{rare}`. If this path contains `{rare}` somewhere upstream (originating from the stem lexicon), then this is the less preferred wordform and thus should contain `+Use/Rare` on its lexical side. In essence, one should check for the parity of `{rare}` tags.

8 Two-level rules for affixes

Figure 3 shows the input side of the two-level rules[6]. The strings are underspecified and redundant at the same time, so the rules must specify the output symbols, as well as prune some paths.

The input symbols with curly braces have the following possible output (surface) realisations: `{s}:s {s}:0 {i}:i {i}:0 {d}:d {d}:0 {rare}:0`, Pl Par stem vowels `{pl.i}:i {pl.i}:0 {pl.u}:u {pl.u}:0`, stem vowels `{sg.e}:e {sg.e}:0 {sg.a}:a {sg.a}:0`. `PlSV` per two-level rules is a set of plural stem vowels `{pl.i} {pl.u}`, and `stemV` stands for all stem vowels `a e i u`. Notice singular and plural stem vowels

[6]The formalism is described in `https://web.stanford.edu/~laurik/.book2software/twolc.pdf`

```
%{i%}:i <=>  :s _ :d ;   ! sid

%{d%}:d <=>  :s :i _ ;   ! sid

%{s%}:s  =>  _ :i :d ;   ! sid
```

Figure 4: Rules for alternating *-sid* with ∅

```
%{s%}:0 =>  PlSV:stemV   :* _ %{i%}:0 %{d%}:0;

Vx:0   <=>  %> PlSV:stemV :0* _ :0* %{s%}:0 ;
     where Vx in ( %{sg.a%} %{sg.e%} %{sg.u%} %{sg.i%} );

Vx:Vy  <=>  :Consonant :0* %> _ :0+ %{s%}:0 ;
     where Vx in ( %{pl.e%} %{pl.i%} %{pl.u%} )
           Vy in (     e        i        u    )
     matched ;
```

Figure 5: Rules per affixes *-sid* and ∅

have been defined in a way that makes it possible to allow them to surface only in certain contexts.

Conceptually, it is not only individual symbols that may be underspecified at this stage, but whole multi-character units such as *-sid* vs ∅. One may view these parallel affixes as different values of a single variable that depend on some context factors, and one can also treat an affix as a trigger to filter some context; in reality, it is enough to use only `{s}` (which happens to occur in all the relevant contexts) as a trigger. The affixes *-sid* and ∅ must be described symbol-by-symbol via the rules on Figure 4.

The first rule on Figure 5 states that the ∅-affix form may occur only if there is an appropriate tag in the stem lexicon entry.

The second rule says that a singular stem vowel cannot surface (i.e. it must be realized as `0`), if there is already a plural stem vowel and a ∅-affix .

The third rule says that a plural stem vowel must surface, if the form has a ∅-affix.

9 Two-level rules for usage tags

Two-level formalism is (mis)used to prune spurious paths that emerge from continuation lexicons. The key is to remember that an allowed path may contain either zero or two `{rare}` tags: the first `{rare}` originates from the stem lexicon, the second from a continuation class lexicon. Remember also that the path contains the surface symbols that define the alternative forms and which can also be used as context conditions.

The first rule on Figure 6 (with multiple contexts) defines all the contexts where the `{rare}` tag may occur. First, it may occur immediately after the parallel form tag in the stem lexicon (`{pl.i}` or `{pl.u}`); this is where the lexicon writer has put it. Second, it may occur immediately after an inflectional ending, but only in a certain context (if it gets there via the continuation lexicon that pairs the lexical side `+Use/Rare` with the surface side `{rare}`). Note that although the parallel form tag indicates that the word has an alternative form, its existence in this pre-final surface form alone is not sufficient to determine the final form. The correct way of reading the rule contexts should be backwards from the end: a rare Pl Par form ends with ∅, if the path contains `{rare}` after the plural stem vowel `{pl.i}` or `{pl.u}`, which gets realised as the surface vowel (defined here via sets `PlSV` and `stemV`); a rare Pl Par ends with *-sid*, if there is no `{rare}` after the plural stem vowel (`PlSV`). Notice that in this context this vowel must surface as ∅, because it is the singular stem vowel that goes with *-sid*; this choice of the correct vowel is achieved by the two-level rules on Figure 5.

The first rule on Figure 6 connects `{rare}` with lexicon tags and spelled-out inflectional affixes. The affixes, in turn, should also be connected to the `{rare}` and lexicon tags (embodied by the set `{PlSV}`). This is what the second and third rules do. Notice that `{s}` is used as the crucial symbol to define the rest of the affix, thus these rules really relate affixes to context, not just the symbol itself.

The second rule states that the Pl Par surface form with *-sid* cannot be common (i.e. not rare, `\{rare}`) if the word already has a common short Pl Par form.

The third rule on Figure 6 states that a short Pl Par cannot be common if the lexicon tag says it is rare.

Figure 7 shows the result of pruning two-level strings of the Pl Par forms *taimi* / *taimesid* (plant), where *taimesid* is the less-used form. Pairing the non-failed surface side strings with the lexical side from Figure 3, and removing the morpheme border symbol `>`, will yield the desired result of Figure 1.

```
%{rare%}:0 => PlSV: _ ;
              PlSV:stemV %{rare%}: :* %{s%}:0 %{i%}:0 %{d%}:0 _ ;
              PlSV:0    \%{rare%}: :* %{s%}:s %{i%}:i %{d%}:d _ ;

%{s%}:s /<= PlSV:0    \%{rare%}: :* _ %{i%}:i %{d%}:d \%{rare%}: ;

%{s%}:0 /<= PlSV:stemV %{rare%}: :* _ %{i%}:0 %{d%}:0 \%{rare%}: ;
```

Figure 6: Rules to prune paths with the tag {rare}

```
lexical: t a i m > {pl.i} {sg.e} > {s} {i} {d} .#.
surface: t a i m >    i      0   >  0   0   0        OK
surface: t a i m >    0      e   >  s   i   d        FAIL

lexical: t a i m > {pl.i} {sg.e} > {s} {i} {d} {rare}
surface: t a i m >    i      0   >  0   0   0     0  FAIL
surface: t a i m >    0      e   >  s   i   d     0  OK
```

Figure 7: Surface strings after applying the two-level rules

Conclusion

Parallel forms, i.e. two or more synonymous forms that realise the same set of morpho-syntactic categories in a paradigm cell of a word, deserve attention from a linguistic theory, as well as from a computational point of view. The paper presented examples of technical solutions for handling parallel forms in Estonian. The proposed method involves using two-level rules as a way of controlling the combinatorial explosion of continuation lexicons. The simplified, but still technically detailed example consisted of only one paradigm slot and a single usage tag. In a full description of a language, there are likely more paradigm slots with parallel forms and/or more usage tags. Currently, the full treatment of Estonian also includes Sg Ill and +Use/NotNorm per word forms that are "incorrect" according to normative dictionaries.

Acknowledgments

This work has been supported by the Estonian Ministry of Education and Research grant IUT 20-56 "Eesti keele arvutimudelid (Computational Models for Estonian)", by the Norwegian-Estonian Research Cooperation Programme grant EMP160 "SAMEST — Sami-Estonian language technology cooperation — similar languages, same technologies", and by the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies).

References

Lene Antonsen, Trond Trosterud, and Francis M. Tyers. 2016. A north saami to south saami machine translation prototype. *North European Journal of Language Technology* 4:11–27. http://www.nejlt.ep.liu.se/2016/v4/a02/.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications.

Kimmo Koskenniemi. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, University of Helsinki. http://www.ling.helsinki.fi/ koskenni/doc/Two-LevelMorphology.pdf.

Karlheinz Mörth and Wolfgang U. Dressler. 2014. German plural doublets with and without meaning differentiation. In Franz Rainer, Francesco Gardani, Hans Christian Luschützky, and Wolfgang U. Dressler, editors, *Morphology and Meaning*, John Benjamins, Amsterdam, Current Issues in Linguistic Theory 327, pages 249–258.

Jaak Pruulmann-Vengerfeldt. 2010. *Praktiline lõplikel automaatidel põhinev eesti keele morfoloogiakirjeldus*. Master's thesis, Tartu Ülikool. https://cyber.ee/uploads/2013/04/Pruulmann-Vengerfeldt_msc.pdf.

Maire Raadik, editor. 2013. *Eesti õigekeelsussõnaraamat*. Eesti Keele Sihtasutus, Tallinn.

Anna M. Thornton. 2012. *Reduction and maintenance of overabundance. A case study on Italian verb paradigms*, Edinburgh University Press, volume 5, pages 183–207.