Oahpa! Õpi! Opiq!
Developing free online programs for learning Estonian and Võro

Heli Uibo
University of Tartu
UiT The Arctic University of Norway
heli.uibo@ut.ee

Jaak Pruulmann-Vengerfeldt
University of Tartu
Cybernetica AS
jjpp@cyber.ee

Jack Rueter
University of Helsinki
rueter.jack@gmail.com

Sulev Iva
University of Tartu
sulev.iva@ut.ee

Abstract

This paper describes porting Oahpa, a set of advanced interactive language learning programs, to two new languages both of which spoken in Estonia – Estonian and Võro. Our programs offer a platform where the user can practice vocabulary and the generation of morphologically complex forms both in isolation and within sentential contexts. An overview of the Oahpa system and its two important building blocks – the morphological finite state transducer and the pedagogical lexicon – is given. The development of morphological finite state transducers for Estonian and Võro, as well as tailoring the specific transducers for pedagogical purposes are described. The adaptation of both Estonian and Võro Oahpa to the target user groups is also discussed.

1 Introduction

1.1 The languages

Estonian is the second largest Baltic Finnic language with approximately 1.2 million native speakers. It has several morphological features common in agglutinative languages. Estonian, however, has had a lot of influence from Swedish, German and Russian, as such it has lost vowel harmony and is shifting towards becoming a fusional language.

Estonian is the only official language in Estonia, and, in many professions, high-level Estonian language skills are required. Free online programs for learning Estonian grammar would contribute to better Estonian language proficiency among the people with other mother tongues in Estonia (31.3% of the whole population in 2011). The motivation to learn the Estonian language is generally high among students and working-age people. Estonian morphology and the use of correct cases are the most difficult things for people with non-Uralic languages as their mother tongues. Therefore, a morphology-aware ICALL system would be a helpful tool for Estonian language learners of all ages.

Recently, a couple of free online language learning environments for Estonian have appeared that are not commercial but require the creation of a user account: keeleklikk.ee and eestikeel.ee. These programs, however, have slightly different foci and target groups compared to Oahpa. They are not very well suited to the needs of university students.

The Võro language belongs to the same branch of the Uralic language family as Estonian and Finnish. Traditionally it has been considered a subset of the South Estonian dialect group of the Estonian language, but nowadays it has its own literary language and the activists of Võro are applying for the recognition of Võro as a regional official language in Estonian. The population of Võro speakers is estimated at 74,400, most of them reside in southeastern Estonia.

At the end of the 1980s a revival of South Estonian varieties started. A new standard of the Võro language was developed by native speakers and activists, linguists and non-linguists alike. The standardisation led to the publication of a bilingual Võro-Estonian dictionary in 2002, containing 15,000 entries, and the Estonian-Võro dictionary in 2014, with 20,000 entries.

A course in the Võro language and local (cultural) history, was taught in 19 schools in the language area in 2012/2013. The Võro language is

Heli Uibo, Jaak Pruulmann-Vengerfeldt, Jack Rueter and Sulev Iva 2015. Oahpa! Õpi! Opiq! Developing free online programs for learning Estonian and Võro. *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015*. NEALT Proceedings Series 26 / Linköping Electronic Conference Proceedings 114: 51–64.

taught mostly in primary school, in most cases as an extracurricular activity, but as an elective in nine schools (Koreinik, 2013).

Most teaching materials for the Võro language have been created, published and provided by the Võro Institute. The materials include a reader/textbook (Võrokiilne lugõmik, 1996), a primer (ABC kiräoppus, 1998), a song collection (Tsirr-virr lõokõnõ, 1999), a workbook for the primer, a workbook for the audiotape, a book of local cultural history (Võromaa kodolugu, 2004), an illustrated vocabulary (Piltsynastu, 2004), and a variety of audio and (audio-)visual materials. In addition, there are many texts which can and are being used for teaching: fiction, poetry, a travelogue, print media and an annual series of the children's own creation (Mino Võromaa, since 1987). (Koreinik, 2013)

Since 1996 the Võro language as a subject can be studied at the University of Tartu. Since 2003 the subject has been called "South Estonian I" for beginners, and "South Estonian II" for advanced students. Since 2004 there have been two series of lectures: "Modern Southern Estonian Literature" and "History of the South Estonian literary language". The language of instruction of all these courses is Võro. Some theses and dissertations have also been defended in the Võro language. In 2006 and 2012 it was also taught at the University of Helsinki.

A free online language learning system is very important for the survival of the Võro language. It will be integrated into the curriculum at University of Tartu. At the same time we aim to design the system in a way that would make it usable for individual internet-based learning. This is the only way to learn the Võru literary language for many people because most of the Võro speakers have never learned the language at school; there are still few possibilities for traditional learning and also the literary language is relatively new.

1.2 Oahpa

The ICALL system Oahpa (Antonsen et al, 2009) has been developed at Giellatekno, the centre for Saami language technology at UiT The Arctic University of Norway. The intended target group of Oahpa are adult language learners and it is primarily meant as a supporting tool for learning vocabulary and grammar for a students attending respective language courses.

The pedagogical motivation behind Oahpa was to develop a language tutoring system which

- has free-form dialogues and sophisticated error analysis
- gives immediate error feedback and advice to the user
- is flexible
- is easily integrated into instruction at schools and universities
- enables the choice of main dialect and metalanguage
- is freely accessible via the Internet

Oahpa consists of six games: a vocabulary quiz (Leksa) which is based solely on a semantically enriched electronic dictionary, a numeral quiz (Numra) based on a small finite state transducer that generates and recognises numbers, date and time expressions, the morphology drill games Morfa-S (isolated word forms) and Morfa-C (word forms in sentential contexts) that require a morphological finite state transducer, a question-answer drill (Vasta) and a dialogue game (Sahka). The last two games require morphological disambiguation and syntactic analysis on top of the morphological analysis.

The first and so far the only instance of Oahpa that incorporates all the six modules – North Saami Oahpa – can be tried out on the URL http://oahpa.no/davvi/. For some other languages a version of Oahpa with four modules exist – Leksa, Numra, Morfa-S and Morfa-C. We are planning to create Võro Oahpa in the same scale. For Estonian our purpose is to go a step further and also implement the fifth module, Vasta, that assumes morphological disambiguation.

Thanks to a cooperation project between the Universities of Tartu and Tromsø, we can make use of the powerful language technology development infrastructure (Moshagen et al, 2014) that has been set up at Giellatekno, and among other things reuse their technologies of creating ICALL applications.

This paper presents work in progress. The described systems are in the stage of development and most of the modules of both Estonian and Võro Oahpa are still incomplete.

2 The prerequisites for creating Oahpa

In order to set up the above mentioned modules of Oahpa the minimal set of language resources consists of

- a morphology engine, e.g. a morphological finite state transducer (FST),
- a pedagogical lexicon that is enriched with semantic categories and other information that is used in Oahpa.

2.1 Morphology engine

We have chosen finite state transducers as a model for formalising Estonian and Võro, partly because this technology is supported by the Giellatekno infrastructure but also considering its theoretical and performance-related pros and cons.

Most modern natural language processing (NLP) applications perform their tasks using statistical language models. At the same time, for morphologically rich languages, estimation of the language models is problematic due to the high number of compound words and inflected word forms. Thus, rule-based models are better suitable for describing the morphology of highly inflected languages. Another argument for choosing the rule-based methods is the relatively limited amount of electronically available texts for languages such as Estonian with its 1.2 million speakers, and even more, Võro, as its literary language is new.

The attractiveness of the finite-state technology for natural language processing stems from four sources: modularity of the design; the compact representation that is achieved through minimization; efficiency, which is a result of linear recognition time with finite-state devices; and reversibility, resulting from the declarative nature of such devices. (Wintner, 2008)

Moreover, given the pedagogical applications in sight, we were not only interested in automatic morphological segmentation but in a system that would be able to generate the complete and correct morphological paradigm for each lemma in the lexicon. That is, for our application correctness was more important than coverage. The resources of an educational application must be manually revised, otherwise such an application would not make any sense.

2.2 Morphological FST of Estonian

North Saami and Estonian stand out among the Uralic languages as the ones deviating most from the agglutinative type. The net outcome of this is a system of non-concatenative morphology (consonant gradation, diphthong simplification) combined with a small set of reusable affixes. This requires concatenative and suprasegmental transducers being composed as serial transducers in order to represent the morphology in an adequate way (for an analysis of Saami and Estonian see (Trosterud and Uibo, 2005)).

2.2.1 Existing implementations

There are at least three implementations of computer morphology of Estonian but they all share one common basis that was described in the lexicon and grammar parts of Concise Morphological Dictionary (CMD) (Viks, 1992). On one hand, CMD was created in cooperation with computational linguists and is quite formal and easy to implement. On the other hand, CMD deals mostly with morphology of simple words and with some derivational processes but ignores completely compounding which gives approximately 10..20% of the words in Estonian texts. Also, its base dictionary is an outdated normative dictionary which has a lot of old words and words that are used only in some dialects. There are words that no one knows what they mean or where they come from. That means that there are some problems with using this system for modern Estonian both in rules and vocabulary.

The best implementation of Estonian morphology is Estmorf (Kaalep and Vaino, 2000) by Filosoft, with roots in CMD, the lexicon has been heavily edited, rules have been adjusted and whole new compounding mechanism is added so that Estmorf would be suitable for using as a spelling check engine and analyser for real Estonian texts.

Another implementation has been created at the Institute of Estonian Language, based on the principles of open morphology (Viks, 2000). It is mostly an implementation of CMD with an added mechanism to allow analysis of compounds.

The third system is an FST-implementation of CMD that started its life as an experiment of describing Estonian with two-level morphology (Koskenniemi, 1983) in Heli Uibo's master's thesis (Uibo, 1999). It was then gradually extended with descriptions of some derivational processes

by Heli Uibo (Uibo, 2005) and with complete dictionary of stems from CMD by Jaak Pruulmann-Vengerfeldt in his master's thesis (Pruulmann-Vengerfeldt, 2010). Also, some compounding rules were added and the whole FST was compiled of multiple smaller, specialized FSTs – there was a FST that described generation of simple word forms, another for simple-word exceptions that would override regular forms, a FST that described which of all possible concatenations of simple word forms are allowed as compounds etc. All those smaller FSTs were combined to a large final FST, that was able to generate and analyze word forms. There were a number of unsolved problems like the need to revise the dictionary similarly to what has been done for Estmorf, overgeneration because of weak compounding rules etc.

2.2.2 Adaptation

Oahpa is built on the Giellatekno infrastucture and so far all the morphology systems that have been in Oahpa have been FST-based. Thus, it was quite natural to try and adapt the existing FST-based system for Oahpa by integrating the existing FST into the Giellatekno infrastructure. This was useful for the other parts of the cooperation project that deal with machine translation as well. Also, the wider context of cooperation project motivates some of the decisions we made about FST.

For most languages, FST-s are described using a large lexicon FST (usually as Xerox lexc (Karttunen, 1993) source or at least something that is compiled to become a lexc source) and another FST to describe phonological processes using two-level rules. The Giellatekno infrastructure is well suited for such a structure and offers a comfortable set of supporting scripts and filters to generate a lot of specialized FSTs from the same source, if one follows some conventions. It is also worth mentioning that most active languages whose FST description is developed in Giellatekno infrastructure are close relatives to Estonian – multiple Saami languages, Finnish and now also Võro.

Our FST started out with a two-FST model. For various reasons, it was developed into a much more complicated system of FSTs. The source code of FSTs consisted of regular sources for automata and custom made build scripts that generated full source files from smaller parts, compiled binary FSTs from source and then combined those automata to get a final lexical transducer. In order to build our FST in the Giellatekno infrastructure, our first step was to reorganize our sources. Some of the reorganization meant that we precompiled some of the sources that were previously generated dynamically. Those build steps that combined small FSTs were merged into the Giellatekno infrastructure. The Giellatekno infrastructure is under active development to cater better to the needs of languages and applications that use language descriptions. The maintainers of Giellatekno infrastructure added necessary hooks, so that we could do some specialised processing between regular build steps of the Giellatekno infrastructure.

After we had managed to build our sources using the Giellatekno infrastructure and get a FST that worked more or less identically to what we had had before, the next step was to adapt our source. Mostly, this meant converting the tag set that was in use in the original FST to use the conventions used in Giellatekno. Tag adaptation had two aspects – most of the conversion was simple relabeling but in some instances the tag sets were not compatible or there were other reasons to consider bigger changes. Our existing tag system was mostly inspired by the structure that was dictated by CMD. Specific labels were chosen so that it would be as compatible with an existing constraint grammar syntax description of Estonian as possible. We suspect that Giellatekno's tag system has similar roots. The tag system is based on the first supported language descriptions and it has been extended and improved upon with the addition of new languages with somewhat different requirements. Most of the infrastructure and applications that depend on it have some adaptation to the existing Giellatekno tag set. We were also aware of the fact that the constraint grammar tools we were originally trying to interface with were about to be integrated into the Giellatekno infrastructure as well. So, it was decided that we would change tags at source level in all our rules and in our morphological lexicon. In addition to simple relabeling where the same thing was expressed with different tags (e.g. +in vs +Ine for inessive, +nom vs +Nom for nominative) there were some minor differences in the meaning of tags. For instance in our system we had separate tags for number (e.g. +pl, +sg) and person (e.g. +ps1, +ps2, +ps3) that were combined where needed as Giellatekno has

precombined tags (e.g. +Sg1, +Pl2 for the first person singular and the second person plural, respectively).

One smaller part of tag relabeling was to convert uppercase letters that were used in two level rules to multichar tags so that uppercase letters could be used as a part of regular alphabet.

The sequence of tags in FSTs is as important as sequence of letters in a word. That means that in order to generate a specific word form with a generating FST, one usually has to know the lemma and the exact sequence of grammatical tags. The simplest form of rule-based machine translation would take a word form in source language, analyze it, replace the stem using translation dictionary and then try to generate the word using the same grammatical tags as the original analysis returned. Of course, the real languages are not that easy to translate and there are numerous more complicated rules but having the compatibility at that level is still a desirable property. As Oahpa needs to generate word forms as well, similar tag ordering rules for different languages are useful for developers and linguists who need to deal with multiple versions of Oahpa in parallel.

Our team members have studied languages that we have been prioritized for a machine translation subproject, that is North Saami and Finnish. Comparing different languages we realized that there is a lot of tradition involved in the ordering of tags in language descriptions. This means that even if there were a generic ordering that could be used for all languages involved, for historical reasons, it would be hard to enforce.

As a result of the analysis of different tagging conventions, the most notable change made to Estonian system was the restructuring of the tags for verb forms by Heiki-Jaan Kaalep theoretical foundations of which are described in (Kaalep, 2015). The aim of restucturing was to have a better match between grammatical meanings and specific surface forms that are used nowadays.

Another important difference between Giellatekno tradition and our previous FST was that in our original FST we automatically and dynamically generated regular derivations as if they were regular lemmas in the dictionary. For example, there are productive rules that derive name of action and actor from a verb (e.g. ujuma 'to swim' gives ujumine 'swimming' and ujuja 'swimmer'). Generating lemmas is not quite trivial as some of such derivations are based on weak-grade stem (e.g. lugema 'to read' and loetu 'something that has been read'). Some of our original more complicated system of FSTs dealt exactly with those derivations. However, Giellatekno infrastructure does not do such things but rather adds derivation-tags to mark that this word was derived from some base lemma using some specific derivation (loetu would be analyzed as *lugema+V+Der/tu+N+Sg+Nom* instead of *loetu+N+Sg+Nom* as before). It appeared that for disambiguation rules that kind of information is useful and so, in the current version we analyze (and generate) such derived words with both the synthetic lemma and original lemma with derivational tags. One of the future tasks is to analyze whether the synthetic lemma is really useful for any application or whether we could drop them and simplify our build system.

During the adaptation and testing of our FST with Oahpa, it appeared that our system did not have a good way to differentiate (partial) homonyms. There are quite a few paradigms that have exactly the same written form for nominative case forms, which is traditional the dictionary form for nouns and adjectives (e.g. sokk (nominative), soki (genitive) 'sock' vs sokk (nominative), soku (genitive) 'male goat'). This is often due to the loss of the final vowel in the nominative and such words actually inflect differently. For any application that needs to generate word forms by knowing the lemma and the grammatical information, that is of course a problem. So, to differentiate paradigms with overlapping nominative we used homonymy tags the Giellatekno tag set contains. This means that applications using our FSTs have to be aware of those tags as well, usually in the form of translation dictionaries having mapping not to the usual nominative but to the lemma with an additional identifier (e.g. sokk+Hom1 and sokk+Hom2).

The generic conclusion from the last two problems is that the lemma in the morphological module is actually an identifier of paradigm and as with other aspects of morphology module – what is useful and what makes sense depends somewhat on the intended users of the module.

The big problem of our current system is the overgeneration. The problem is largely a result of the mechanism of compounding. The current system combines the automaton of simple words with

itself and then applies the filters that should only allow proper compounds and simple words. Such a structure enables us to add compounding relatively easily and deal with certain other problems that were hard to solve within our system of multiple automata. The current rules of compounding are too generic and there is a lot of allowed forms that actually are not used. Inclusion of some short words like names of musical notes in the lexicon makes this problem worse. Using a regular unweighted FST makes it also hard to prefer simple word analysis over compound word analysis.

The usual way to implement compounding rules in the Giellatekno infra would use diacritic flags as described in (Beesley and Karttunen, 2003) and cycles in lexicon description. Converting our system to use compounding mechanism that would be more in line with other language descriptions that use the Giellatekno infra is something that we have considered but have postponed so far. The most important reason for this is that there is no clear and formal description of compounding rules for Estonian. The gap between existing formal rules (e.g. a noun can be added to the nominative or genitive form of another noun) and what really is used (which of those two forms is preferred or if some combinations are used at all) is quite big and sometimes explained only by tradition.

As an experiment with flag diacritics we implemented rules for the lowercasing of proper name derivations. We found that the current system of filters makes the creative use of extra flags quite hard as the flags have to be precisely described in filters to show where exactly they can appear. Also, as flag diacritics are special and nontrivial to implement, they tend to cause problems that are hard to debug in both the filtering and composition of FSTs. One issue, for instance, is whether the negation of the whole alphabet contain flag diacritics in different implementations of FST tools or even in different operations in the same tool? What happens when we apply priority union [1] to FSTs with and without flag diacritics?

The other aspect of overgeneration is parallel forms. There are alternative forms of illative (long illative that uses regular morpheme and short illative that usually uses stem alteration, e.g. majasse vs majja 'into a house') with some preference rules for different inflection classes ('majja' is the preferred form but 'majasse' is understood as well). Some inflection classes also allow multiple forms for some plural cases (regular plural with morpheme vs plural stem, e.g. õpikutele vs õpikuile 'onto the textbooks'). Knowing all those forms is necessary for analysis but for generation in machine translation and for educational purposes it would be good to have only the preferred forms generated. Giellatekno infrastructure offers the tag +Use/NG to denote forms that are used but should not be generated for such purposes. We do have some experimental use of that flag but we still need to check and tag a lot of parallel forms either by word class or, in some cases, actually word by word.

[1] an operation on FSTs which allows us to declare and combine a large regular FST and a smaller one with some exceptions that override regular rules with much shorter descriptions than lexc-only descriptions would allow

2.3 Morphological FST of Võro

Võro, as is the situation with many of the other Uralic languages, has an abundance of regular morphology in both the nominal and verbal parts of speech. As such, ready solutions for many of the morphological challenges in Võro might be sought out in previous work done on the open-source, Saami language technology infrastructure "Giellatekno" in Tromsø, Norway. Morphophonological work at Giellatekno on the Saami languages has dealt with stem-internal vowel and consonant change as well as orthographical word compounding.

In the initialization of a new language at "Giellatekno", there are a number of default files for the development of two-level model and lexc descriptions. Concepts useful in the development morphophonological strategies and present in the default files include triggers and allophonic variables. Typically, triggers might be used in coordinating gradation in the stems, whereas allophonic variables might be utilized in progressive vowel harmony. There are, however, a few advanced languages to follow in development at Giellatekno, namely, Northern Saami, Southern Saami and Finnish. When in doubt these are the ones to quote and question because they are the scenes of most active development.

The morphophonological characteristics of Võro, at first glance, appear to be reminiscent of those attributed to Finnish. In addition to a parallel of the gradation system found in Estonian and Northern Saami, where changes in stem quantity and quality can be attested without apparent

surface-level motivation, Võro possesses progressive front-back vowel harmony. This said it was easy to find parallels on the Giellatekno infrastructure.

2.3.1 Initial approach to finite state description

Classification of the Võro language is available from Sulev Iva's dissertation (Iva, 2007) and online in the Võro-Estonian-Võro dictionary site `http://synaq.org`. Sulev Iva provided a digital copy of his word type classification lists, and work could be commenced.

In a parallel to previous lexc work on the Giellatekno infrastructure, inflection type names are simply words representative of the given type. Thus the inflection type name "VÕROKÕNÕ" 'a person from Võro' is used to distinguish nominals sharing its declension characteristics (cf. North and South Saami and Finnish). Since subsequent work often includes syntactic disambiguation, a part-of-speech indicator is prefixed onto the inflection type, which renders A_VÕROKÕNÕ, N_VÕROKÕNÕ and PROP_VÕROKÕNÕ continuation lexica that can be further directed to a mutual nominal lexicon in NMN_VÕROKÕNÕ (it is done similarly in the morphological FSTs of other Uralic languages such as Finnish, Livonian, Moksha, Hill Mari, Livvi, Skolt Saami and Nenets).

According to the initial description used for Võro, there are approximately 50 different declension groups (47 vs 53) and 40 conjugation groups (40 vs 36). The inflection groups contain words representative of both front and back vowel harmony, and therefore work was immediately begun on the description of progressive front-back vowel harmony.

In accordance with previous work on Mari, Erzya and certain Balto-Finnic languages a ready solution was reached for front-back vowel harmony. Two-level rules can deal with vowel harmony through the definition of vowel sets and contexts. In Võro this was initially accomplished through the definition of back, front and neutral vowels (1), and subsequent contexts (2).

```
(1)
VowBack = a o u \~{o};
VowFront = \"a \"o \"u ;
VowNeutral = e i ;
```

As is the case in Finnish, there is a set of neutral vowels that do on block vowel harmony, in Võro these vowels are *e* and *i*. Since work with two-level rules is something that continues over a certain period of time, it is always useful to provide illustrative example contexts for the individual rules. These examples will also help in future development since the original writer will not always remember or be there to explain them.

Back vowel context, as illustrated in (2.1) can be broken into four increments. The first increment is a required word boundary followed by zero or more consonants. The second increment is the optional insertion of one or more neutral vowels followed by zero or more consonants. The required third increment is the presence of an underlying or surface back vowel, which is followed by a fourth increment that cannot contain a word boundary or front vowels, be they underlying of surface.

```
(2.1)
Back vowel context

BT = # Cns* ([VowNeutral]+ Cns:*)
[VowBack: | :VowBack]
[# | VowFront: | :VowFront]* ;

(2.1.1)
!€# viska%^WGStem%>%{A\"a%}q
!€0 vis0a00aq
```

The example in (2.1.1) shows a combination of a trigger ^ WGStem (weak grade stem) and an allophonic target {Aä} – front-back harmony for low unrounded vowels *a* and *ä*, where the resulting vowel harmony is back-harmony *ä*.

```
(2.1.2)
!€# füüsiga%^StrGStem%^VowRM%>i%>d%{ÕE%}
!€0 füüsik0000i0de

(2.1.3)
!€# füüsiga%>l%{ÕE%}
!€0 füüsiga0lõ
```

Problematic contexts with mixed harmony can be observed in examples (2.1.2) and (2.1.3), where the word "füüsiga" contains both front and back vowels. Here irregularity in just a few stems compromises the simplicity sought in two-level rules. One possibility, of course, is to list these irregularities as exceptions. A second possibility, it will be noted, is to classify stems on the basis of front-back harmony for all inflecting word classes. This is what the OMorFi description of Finnish does, no two-level rules are given for progressive vowel harmony.

The continuation lexica in the OMorFi Finnish description explicitly indicate both harmony and

gradation, and virtually leave the two-level rules unused. In practice this utilization of lexc doubles the number of inflection type lexica for those with vowel harmony targets, which, in the case of Finnish nouns, would comprise seven instances out of twelve. The illative in Finnish provides its own challenge, since it entails a duplication of the stem-final vowel, i.e. eight vowels. Gradation in Finnish centers on the plosives “k”, “t” and “p”, which is a small number in comparison to what is attested in Võro. These combinations are augmented by the need for expressing pluralia tantum, and the result is upward of 550 noun types (25.03.2015).

This is a good time to ask whether such a solution might be used in Võro, and whether it would be useful. First we have to ask ourselves what the transducers will be used for. If we are interested in ICALL, then we want intelligent feedback for our language learners.

Intelligent feedback can be written directly in the descriptions accompanying each inflection type. By establishing vowel harmony in an inflection type, we are providing the computer the necessary information needed to prompt the learner with regard to vowel harmony issues. We are looking into making information available at the lexicon level for computer reading. It is hoped that the information might be automatically added to individual words directed through a given lexicon, see (3).

```
(3)
(3.1) LEXICON N_PEREH
(3.2) ! pereh:perre
(3.3) ! vowel_harmony: front
(3.4) ! gradation: yes
(3.5) !! * Yaml: __N-pereh_gt-norm.yaml__
(3.6) :%^VowRM%>i FRONT_PL-GEN_de ;
(3.7) +Use/NG:%^WGStem%> h FRONT_PL-GEN_de ;
```

The declarations for vowel harmony at (3.1.3), and gradation at (3.1.4) are information bits that can be transferred to the ICALL infrastructure. In fact, it is the initial continuation lexicon associated with a given lemma and stem pair that makes reference to all this information, see “N_PEREH” in (3.1).

Hence the continuation lexicon “N_PEREH” in (4) is associated with the definition at LEXICON N_PEREH in (3.1), and attribute values can be added to the “<l>” element in (5). At the same time there is contemplation going on with regard to the use of well-documented triggers, such that the information from a single input line could be utilized by the computer for meaningful feedback. Such feedback in (3.1) might include “vowel loss” derived from the trigger “%^ VowRM”. The trigger “%^ WGStem” would indicate “weak grade stem”. Both would be associated with plural genitive in “de”, as indicated expicitly in the continuation lexicon “FRONT_PL-GEN_de” at (3.1.6) and (3.1.7). The presence of the tag “+Use/NG” in (3.1.7) implies that the tagged sequence will be accepted by the analyzer but not generated.

```
(4)
pereh:perre N\_PEREH ;
```

Here the attribute values in the “<l>” element are hoped to be applied to the morphological games.

```
(5)
<e>
  <lg>
    <l gradation="yes" pos="N"
        vowel_harmony="front">
        pereh
    </l>
  </lg>
...
</e>
```

2.3.2 Brief assessment of progress

A happy medium is being sought for the intermingling of lexc and two-level rules strategies. Gradual progress is being made towards a lexc solution to vowel harmony, whereby word stems are classified for front-back harmony, so as to allow for immediate vowel-harmony error generation. Changes in the stem, however, are being worked on with triggers paralleling morphophonological rules such that wrong triggers can be applied in order to produce erroneous forms, according to strategies already developed for Northern Saami.

2.4 Pedagogical lexicons

We have used different approaches when composing the lexicons for Estonian and Võro Oahpa. The lexicon of Estonian Oahpa has been created on the basis of the word list of the Estonian textbook for beginners ”E nagu Eesti” (Pesti and Ahi, 2011). The word list divided into the twenty-five chapters is given in the end of the textbook. It is a list of ca 1500 Estonian words and expressions with translations to English, Finnish, Russian and German.

We can bring out the following advantages and disadvantages of using a textbook's dictionary as a basis of Oahpa lexicon:

+ Book and chapter information are given, the lexicon can be easily used for additional grammar training in courses that base on that book. Translations of words to four most common languages of learners of Estonian exist.

- Information about part-of-speech and semantics had to be added manually or semi-automatically.

The creation of the Võro Oahpa lexicon started out with a small lexicon that had just one word for each inflection type. We have chosen this approach because we think it is important that the morphological exercises cover all the inflection types. We have tagged the representative words of inflection types so that exclusively these words can be chosen for the morphology drill exercise Morfa-S. Otherwise, the words are chosen from the full lexicon where some of the inflection types are less frequent and as the words are randomly selected there is a possibility that the user does not get a chance to practise some inflection types. After that we added ca 2500 words from the lexicon of North Saami Oahpa that incorporated translations to North Saami, Finnish, English, Norwegian (bokmål), Swedish and German.

Advantages and disadvantages of using an Oahpa lexicon of another language as a basis were the following:

+ Part-of-speech and "Oahpa-style" semantic information were already there, as well as translations to Finnish and English.

- The word list does not match the word list of any textbook of Võro, therefore this information must be added afterwards. Because it was a North Saami lexicon, it contained many words that were irrelevant for Võro (words about Saami handicraft, reindeering, also too strong focus on the topic "Christmas").

- Translations to Võro and Estonian had to be added.

3 Creation of the Oahpa applications for Estonian and Võro

Oahpa is a web application developed in Django framework. Django is a powerful open-source framework for creating web applications supporting the model-view-controller (MVC) design.

3.1 Setting up the Django application

As there has already been set up a number of instances of Oahpa for different languages in the Giellatekno infrastructure the process of creating Oahpa for Estonian and Võro was quite routine and did not require much effort. Obviously, each Oahpa instance has different settings – paths to linguistic tools, database access data, list of supported languages etc. We are aiming at having almost all the language-specific information in the settings file and in the database, rather than in the Python code. However, the Python code is still not entirely language-independent. A few adjustments were needed in the lists of grammatical and semantic categories, in the set of word attributes that come in from the lexicon, in the list of language pairs in the vocabulary drill game Leksa, in the list of localisation languages, in the initial settings of the games and in the spell-relax function.

3.2 Creation of the database

The complete database for an instance of Oahpa that incorporates Leksa, Morfa-S and Morfa-C (note that the program Numra is solely based on the finite state transductors converting the numerical, time and date expressions into textual form and vice versa) contains

- words, their translations and semantic classes
- tags used in the morphological analysis and their possible sequences (paradigms)
- word forms for Morfa-S
- question templates for Morfa-C linked to the word forms that can replace the variables in the templates
- morphological feedback information for each word form (this is combined from different characteristics and features of the word that gives hints about how to inflect the particular word)

We use a morphological generator to make the paradigms automatically. During the generation of the word forms and saving them into the database an error log is written to a file. So the database generation process also serves as testing of the morphological FST.

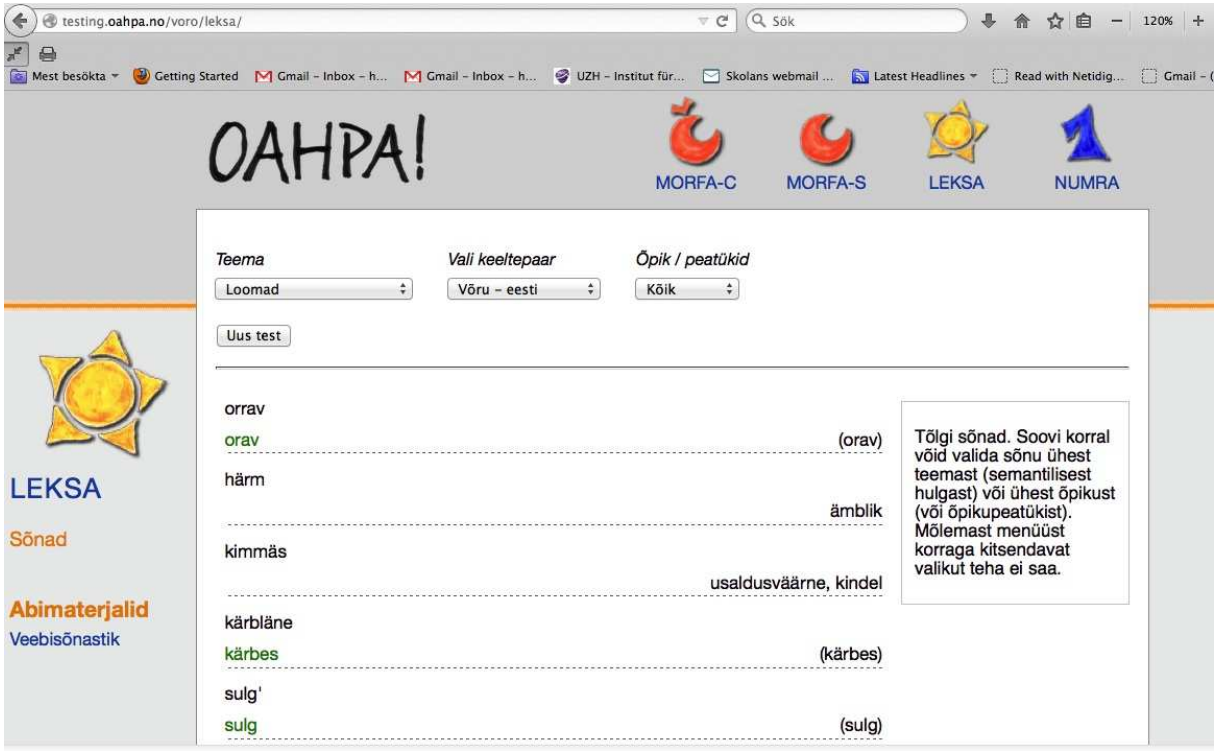

Figure 1: Screenshot of the vocabulary drill program Leksa in Võro Oahpa

So far we have set up Leksa, Morfa-S (substantives) and Morfa-C (substantives) for Estonian. The programs have been tested by the developers of FSTs and Oahpa and demonstrated to the teachers of Estonian at Tartu and Uppsala universities and at Estonian School in Stockholm for getting some feedback.

The working modules of Võro Oahpa are Numra, Leksa and Morfa-S (substantives and verbs). The user interfaces of both Estonian and Võro Oahpa have been translated to Estonian.

The user interface of Leksa in Võro Oahpa can be seen on Figure 1. There are three menus for specifying the exercise – *teema* ('topic', i.e. semantic category), *keeltepaar* ('language pair') and *Õpik / peatükid* ('book / chapters'). The first menu makes it possible to constrain the set of words offered to the user by semantics. On Figure 1 the words are chosen from the category *Loomad* ('animals'), for example *orrav* ('squirrel'), *härm* ('spider'), *kärbläne* ('fly'), *sulg'* ('feather'). There are 19 semantic categories in the list, among others family, food/drink, time, body, clothes, buildings/rooms, work/economy/tools. From the second pull-down menu the user can choose the language pair. The default is from Võro to the language of the user interface (in the given case – Estonian). Other translation languages in the list are Finnish, English, German, North Saami, Swedish and Norwegian. The correct answers of the user are displayed in green. In the second column the correct answers are presented by the system. The correct answer is given in parentheses if the user's answer is correct.

The current version of Estonian Oahpa can be tried out at the address `http://testing.oahpa.no/eesti` and Võro Oahpa at `http://testing.oahpa.no/voro`. The programs are free to use for everyone and do not require any registration.

3.3 Some problems and their solutions

3.3.1 Spell-relax

Spell-relax means that the program accepts different variants of typing for some characters or sequences of characters. This feature had been previously implemented in Oahpa in order to make Oahpa usable for users who do not have access to a keyboard (either virtual or real) with the layout of the language in focus.

We have not implemented any spell-relax in the Estonian Oahpa. We could perhaps consider accepting 'sh' instead of 'š' and 'zh' instead of 'ž' because it might be that everybody has not

installed the Estonian keyboard but probably it would be a better idea to have a link to installation instructions of an Estonian keyboard. The written Estonian is highly normative and it would not be pedagogical to accept wrong spellings where for example the letters ä, ö, ü, õ are replaced by corresponding letters without diacritics. It is also important for the learner to capitalise proper names etc. where needed. Uppercase/lowercase mistakes are not tolerated.

The situation is quite different for the Võro language. We have implemented spell-relax in Võro Oahpa because the written language of Võro is relatively new and there is a big variety for how some phenomena are expressed. The things that are being spelled in various ways are not usual phonemes but rather symbols that mark a slightly different pronunciation:

1. palatalisation (conventionally denoted by modifier apostrophe, but all the other apostrophe-like characters are also accepted)

2. glottal stop (conventionally denoted by the letter 'q' but the use of 'q' is not consequent in the texts that are being published in the Võro language)

3.3.2 What is a correct word form?

As a developing language with multiple accepted forms, Võro may prove overwhelming for the beginner. For solely pedagogical purposes, it may prove necessary to limit the number of forms generated by the computer prompter to one given standard while allowing students the liberty of writing all possible forms. To this end the tag "+Use/NG", which has been used in MT previously at Giellatekno, can be used. Its use will provide form preference, something parallel to word preference already marked in the oahpa xml dictionaries with the "stat" attribute value "pref".

Should we accept some forms that are not normative but widely used? We might do it for Võro as the standarisation of this language is not finalised yet. The program should, however, suggest the normative form as the correct answer after it has accepted a widely used but non-normative form.

4 User groups of Estonian and Võro Oahpa and adaptation issues

The primary target group when designing Oahpa framework (that is, when developing the first version of the North Saami Oahpa in 2009-2011) were university students and other adult language learners who were learning North Saami as L2. The North Saami Oahpa has been integrated into the university courses at UiT. There are course pages with different kinds of materials for learning North Saami – texts with reading comprehension questions, recorded dialogues, grammar explanations, lexicons. When taking a university course in North Saami the students are working with Oahpa in the logged in mode that makes it possible for students to see their progress and for teachers and researchers to track the activity of the students, also they can see which topics in the course seem to be most difficult and hence should be given more attention. From the lessons there are direct links to appropriate drill exercises in Oahpa.

On the course pages `http://kursa.oahpa.no` and in North Saami Oahpa the scientific linguistic terminology is used and the grammar is explained on a level that is appropriate for its primary target group – adult learners. It should be noted, however, that some primary and secondary schools have also expressed an interest in using Oahpa. Since Oahpa is freely available on the Internet, it should be adapted to the wide user group – there should be possibilities to ask for help about difficult terms etc. That is why the developers of the North Saami Oahpa have introduced additional tooltip explanations of terms in Vasta and Sahka that pop up when pointing on a term in the error feedback. We plan to implement such help tooltips for Estonian Oahpa as well.

The target group of Võro Oahpa will be the students at the Võro language courses at University of Tartu. These are typically students of the Estonian language, thus they usually have a solid linguistic background.

The Estonian Oahpa will have three or four quite different user groups. The first group are the foreign students who have come to study at the University of Tartu and are taking a course in Estonian as L2. They will use Oahpa parallel to traditional language lessons in the classroom. The students have different mother tongues and are learning Estonian on the basis of English, Finnish or Russian. The University of Tartu Language Centre is organising these kinds of courses and plenty of students sign up for them each term.

The second group are students who take the

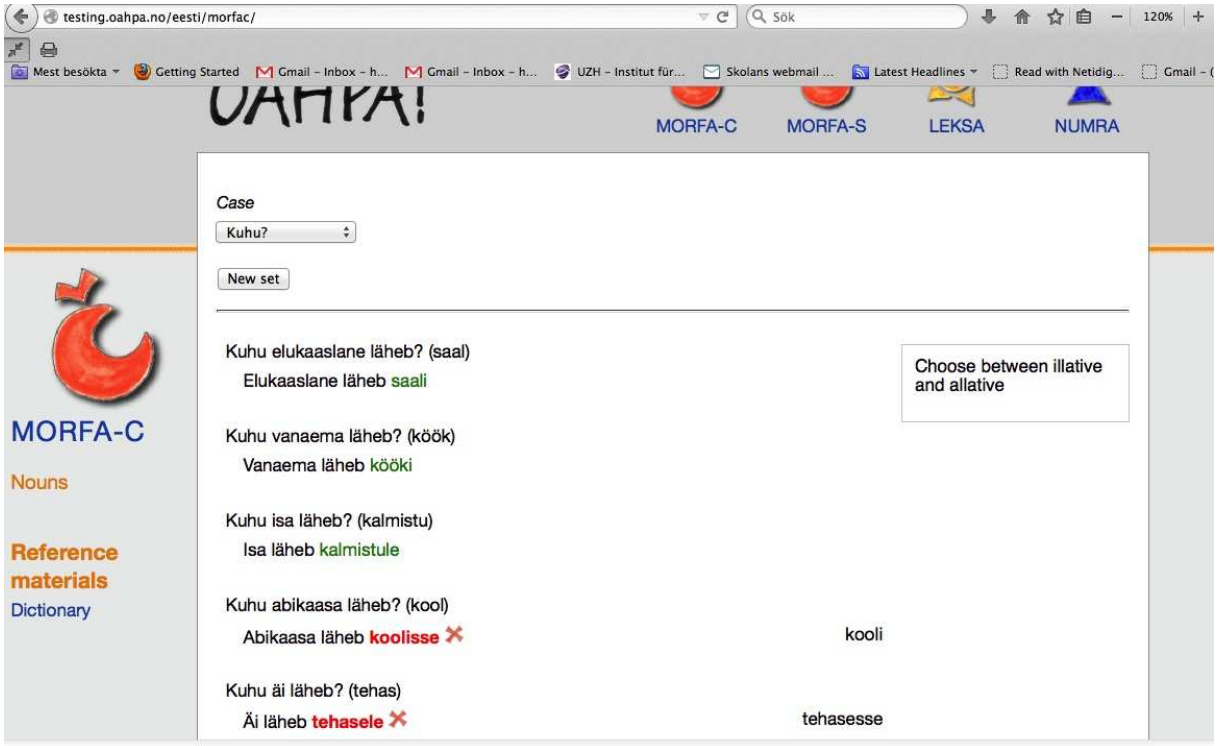

Figure 2: Screenshot of a more advanced Morfa-C exercise

web-based course in Estonian at Uppsala University in Sweden. The course that has been created by prof. Raimo Raag (Raag, 2010) is totally internet-based, the teachers meet their students only at video conferences. The course authors and teachers estimated Estonian Oahpa potentially useful for their students in achieving their vocabulary and grammar learning goals, given that some more exercise types will be implemented and links set from the lessons in the course materials to the relevant exercises in Oahpa.

The third group are the pupils at Estonian School in Stockholm (ESS). This is the only school in Sweden where Estonian language and culture are taught. The pupils in this school have different language backgrounds. Part of them have Estonian as their mother tongue and have recently moved from Estonia to Sweden, another part has lived in Sweden for a longer time and grown up in the Swedish language environment. Some pupils are grandchildren of Estonians who moved to Sweden in the 1940s. Another part of the pupils has no connection to the Estonian language at all.

Considering these different user groups we definitely have to make some adaptations, in particular for the young learners.

We have translated the lexicon of Estonian Oahpa to Swedish, for making Leksa usable for pupils at ESS. Leksa has been tested by the second grade pupils at this school and the feedback was positive. The teachers see the use of Leksa for learning both Estonian and Swedish. For young children with Estonian as the mother tongue and for Estonian as L2 learners Leksa may be used for training spelling of Estonian words. Estonian children who have recently moved to Sweden can also use Leksa for learning Swedish words.

Instead of international (Latin) case names that are generally not known to primary and secondary school students and because the school grammar books use Estonian case names, we are using case questions (e.g. *kelle?* ”whose” *Omastav* ”Genitive” instead of *Genitiiv*) and Estonian case names.

Led by the feedback of university teachers we have deviated from the standard setup of case list in Estonian Morfa-C. Instead of always explicitly giving the case we have implemented some exercises which include a choice between two grammatical forms. For example, in the exercise “Kuhu?” ’Where (to)?’ the student must choose between illative and allative. An example screenshot of this exercise is presented on Figure 2.

The Estonian language teachers at ESS have also given some other ideas for Morfa-C exer-

cises that are on the waiting list of implementation: choosing the object case, choosing the correct infinitive form (there are two infinitives in Estonian – da-infinitive and ma-supine – the usage of which in a given context is difficult for non-natives) and more. These are examples of exercises that are well supported by the Oahpa framework and easy to implement.

Another possibility to adapt the same instance of Oahpa to different user groups is to have the choice between different sources of words. Both Leksa and Morfa-S have the corresponding menu 'book' in their user interface. Lexicons, word lists of textbooks, single chapters or groups of book chapters can be listed in this menu.

For teaching the university course of Estonian as a foreign language the textbook "E nagu Eesti" (Pesti and Ahi, 2011) is used both at the University of Tartu and Uppsala University.

We are also planning to add the dictionaries of the Estonian textbooks used at the Estonian language courses at ESS into the Oahpa lexicon. The same textbooks are also used at Russian schools in Estonia and the Estonian school in Riga.

One of the initial ideas when designing Oahpa was that only "the known" words (words that occur in the textbook's word list and also in the vocabulary drill program Leksa) will be used in grammar exercises. We will make it more fine-grained. According to feedback from the teachers of Estonian it is important that beginners' grammar exercises would not contain too advanced vocabulary. Thus, there is a new detail in the Morfa-C question frames for Estonian – not only the semantic class but also the book chapter where the word is introduced is determined when selecting words for a particular grammar exercise.

Some of the vocabulary can also be unknown because of cultural differences. For example food differs quite a lot even between otherwise culturally quite similar countries Estonia and Sweden. People who have not grown up in Estonia may wonder what *ühepajatoit* (a typical Estonian late summer / autumn hot pot usually made of pork, carrot, turnip and cabbage) or *rosolje* (a Russian beet root salad) is. We still think that learning a language cannot be separated from getting acquainted with the culture. Probably, these words are not appropriate in the exercises meant for absolute beginners but they could come a bit later.

5 Conclusions and future work

The use of FSTs for morphological analysis and generation and standardised XML formats to store lexicon and exercise frames makes it possible to effectively create a variety of morphological drills for learners of morphologically complex languages.

Our experiments with setting up language learning system for two new languages – Estonian and Võro – prove that the method that has been worked out at Giellatekno research group in Tromsø is efficient and makes it possible to create the first prototypes of vocabulary and morphology drill modules with a relatively small effort. The obligatory prerequisites for creating such a system are a lexicon that can just be a word list of the course textbook in the pdf format and the morphological FST that will be used for generating all the inflection forms of the words in the lexicon.

The major work that has to be done is the work in developing the morphological FST. At the same time, the FST in itself is a multi-purpose building block that can be used in a variety of applications as for example spelling check, machine translation and an intelligent dictionary.

In our case, we had to create the Võro FST from scratch. Despite that, we could start developing the language learning system in parallel with that and very soon come up with a prototype of the morphology drill program that just contained one representative from each inflection type.

There existed a "beta version" of FST for Estonian that had to be restructured somewhat in order to accommodate it in the Giellatekno infrastructure.

The pedagogical applications also set some additional constraints on the FST. It is usual that some of the parallel forms and infrequent words have to be excluded from the pedagogical FST. Here, once again, the Giellatekno infrastructure already had a solution that we could apply.

The described systems are in the middle of development but as the feedback from teachers has been positive we feel optimistic about continuing the work on both Estonian and Võro finite state transducers and the respective Oahpa instances where with first of all plan to complete Morfa-S and Morfa-C with other inflectable word classes.

It is also important to add more guidelines and error feedback to the systems. We are going to use the same approach as in North Saami where

the feedback of morphological forms is combined from pieces of information that characterise the inflection type of the given word. This approach is described in (Antonsen, 2012) and (Antonsen et al., 2013).

References

Lene Antonsen, Saara Huhmarniemi, and Trond Trosterud. 2009. *Interactive pedagogical programs based on constraint grammar*. Proceedings of the 17th Nordic Conference of Computational Linguistics. Nealt Proceedings Series 4.

Lene Antonsen. 2012. *Improving feedback on L2 misspellings – an FST approach*. Proceedings of the SLTC 2012 workshop on NLP for CALL, Lund, 25th October, 2012. Linköping Electronic Conference Proceedings 80: 1-10.

Lene Antonsen, Ryan Johnson, Trond Trosterud, and Heli Uibo. 2013. *Generating modular grammar exercises with finite-state transducers*. Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013, May 22-24, Oslo, Norway. NEALT Proceedings Series 17: 27-38.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology.* CSLI publications in Computational Linguistics, USA.

Sulev Iva. 2007. *Võru kirjakeele sõnamuutmissüsteem 'Inflectional Morphology in the Võro Literary Language'*. PhD Thesis. Tartu Ülikooli Kirjastus, Tartu, Estonia.

Heiki-Jaan Kaalep. 2015. Eesti verbi vormistik. 'Estonian verb paradigm' *Keel ja Kirjandus 1/2015:* 1–16. Eesti Teaduste Akadeemia ja Eesti Kirjanike Liidu ajakiri. SA Kultuurileht, Tallinn, Estonia.

Heiki-Jaan Kaalep and Tarmo Vaino. 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. 'The complete morphological analysis of the text in the toolbox of a linguist.' *Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised:* 87–99. Tartu Ülikooli Kirjastus, Tartu, Estonia.

Lauri Karttunen. 1993. *Finite-State Lexicon Compiler.* Technical Report. ISTL-NLTT-1993-04-02. April 1993. Xerox Palo Alto Research Center. Palo Alto, California.

Kadri Koreinik. 2013. The Võro language in Estonia. ELDIA Case-Specific Report. *Studies in European Language Diversity 23.* (Ed.) Johanna Laakso Research consortium ELDIA c/o Prof. Dr. Anneli Sarhimaa, Northern European and Baltic Languages and Cultures (SNEB), Johannes Gutenberg-Universität Mainz.

Kimmo Koskenniemi. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD Thesis. University of Helsinki.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 2014. Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages *Proceedings of CCURL (Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era) workshop 2014 organised with LREC2014:* 71–77 European Language Resources Association (ELRA).

Mall Pesti, and Helve Ahi. 2011. *E nagu Eesti. Eesti keele õpik algajaile 'E as Estonia. Estonian for beginners'*. TEA Kirjastus, Tallinn, Estonia.

Jaak Pruulmann-Vengerfeldt. 2010. *Praktiline lõplikel automaatidel põhinev eesti keele morfoloogiakirjeldus 'Practical Finite State Morphology of Estonian'*. M.Sc. Thesis. Tartu Ülikool, Tartu, Estonia.

Raimo Raag. 2010. Den språkliga mångfalden – småspråkens renässans. (Ed.) Jenny Lee *Kunskapens nya världar, Uppsala: Uppsala Learning Lab, Uppsala universitet:* 211–221.

Trond Trosterud and Heli Uibo. 2005. Consonant Gradation in Estonian and Sami: Two-Level Solution. (Eds) Antti Arppe et al. *Inquiries into Words, Constraints and Contexts:* 136–150.

Heli Uibo. 1999. *Eesti keele sõnavormide arvutianalüüs ja -süntees kahetasemelist morfoloogiamudelit rakendades. 'The computerized analysis and synthesis of Estonian word forms using the two-level morphology model'*. M.Sc. Thesis. Tartu Ülikool, Tartu, Estonia.

Heli Uibo. 2005. Finite-State Morphology of Estonian: Two-Levelness Extended. (Ed R. Mitkov) *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP) 2005:* 580–584. Borovets.

Ülle Viks. 1992. *A concise morphological dictionary of Estonian: introduction & grammar.* Estonian Academy of sciences, Institute of language and literature.

Ülle Viks. 2000. Eesti keele avatud morfoloogiamudel. 'An open morphology model of Estonian' *Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised:* 9–36. Tartu Ülikooli Kirjastus, Tartu, Estonia.

Shuly Wintner. 2008. Strengths and weaknesses of finite-state technology: a case study in morphological grammar development. *Natural Language Engineering 14(4):*457-469. Cambridge University Press.