Building and using language resources and infrastructure to develop e-learning programs for a minority language

Heli Uibo
University of Tartu
UiT The Arctic University of Norway
Landstormsvägen 12
68534 Torsby, Sweden
heli1401@gmail.com

Jack Rueter
University of Helsinki
Keinutie 11 M 72
FIN-00940 Helsinki
rueter.jack@gmail.com

Sulev Iva
University of Tartu
Ülikooli 18, 50090 Tartu, Estonia
Võro Institute
sulev.iva@ut.ee

Abstract

We will demonstrate Võro Oahpa (http://oahpa.no/voro), a set of language learning programs for Võro, a minority language in Estonia. When setting up and developing the system, we have made use of the infrastructure developed at the Saami language technology centre Giellatekno, UiT the Arctic University of Norway and the Võro language resources and tools – the online electronic dictionary synaq.org that also includes pronunciations; Võro speech synthesis; the morphological finite state transducer that is being developed as a part of the same project and a multilingual word list from North Saami Oahpa. Võro Oahpa consists of four language learning programs: Leksa – a vocabulary quiz, Numra – a program for practicing numerals and time expressions, Morfa-S – morphology drill and Morfa-C – morphology exercises formulated as question-answer pairs. The development is still in progress but the programs have already used within the Võro language course at the University of Tartu. We discuss the issues specific for Võro and show how combining the existing infrastructure, resources and experiences can make the development of a learning system for a language with limited resources easier and give extra values to the system.

1 Introduction

The Võro language is a South Estonian language with ca 70 000 speakers. (The Estonian written language is based on the dialects of Northern Estonia). Võro organisations (Võro Institute etc.) want to recognise Võro officially as a separate language which has been discussed twice in Estonian parlament Riigikogu, however, without positive decision. That is why Võro is officially considered a dialect of Estonian even up to now. Despite of this Estonian government supports the maintenance and development of Võro by financing Võro Institute – a state institute dealing with Võro language and culture. Võro also has its own official ISO language code 'vro'.

The Võro language is taught in ca 20 kindergartens (in so-called language nests) and about the same number of schools in South-Eastern Estonia. Altogether 450 primary and secondary school students are learning Võro language and culture or participating in other classes where the language of instruction is Võro (Opetajate leht, 2017). The kindergartens and schools are located in the area where children's parents or grandparents also might speak Võro but it is not necessarily the case. On the other hand, thousands of Võro speakers or people interested in learning Võro language live in other parts of Estonia and Võro Institute has got queries about distant courses of Võro for adults. Since 1996 the Võro language as a subject can be studied at the University of Tartu. The Võro language course is given every term and has a form of a traditional language course with auditorial lessons.

We are aware of only one other online program for Võro language learning that existed before Oahpa – the game "Mein Zimmer" (http://edlv.planet.ee/meinZimmer/) that has among others been adapted to the Võro language. It is a nice "find-a-key" game but it is focused on one particular topic and thus very limited.

When learning a Uralic language the most difficult thing is morphology. Although there are a lot of language learning programs available, most of them deal with vocabulary learning.

Thus, there was and is a need for free online language learning tools for Võro that would cover

Heli Uibo, Jack Rueter and Sulev Iva 2017. Building and using language resources and infrastructure to develop e-learning programs for a minority language. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017*. Linköping Electronic Conference Proceedings 134: 61–67.

the basic vocabulary for a learner and, most importantly, the basic grammar. In 2013, as a part of a cooperation project between the language technology researchers at the University of Tartu and the Saami language technology centre Giellatekno at UiT the Arctic University of Norway, we started to adapt Oahpa, a set of language learning programs, initially created for North Saami and by now implemented for more than 20 languages, to the Võro language. The ICALL system Oahpa (Antonsen et al, 2009) is primarily meant as a supporting tool for learning vocabulary and grammar for adult students attending respective language courses. But as the usage statistic shows, a lot of people who do not attend any course, also use the system for learning North Saami because it is freely available on the internet. During a 6 months period there were 3,676 unique visitors of North and South Saami Oahpa pages (Antonsen et al., 2013) while the number of people who were taking the respective language courses was about ten times smaller.

So, Oahpa should be a good choice for the intended users of our language learning programs – the participants of the Võro language course at University of Tartu and all other Võro language learners whereever in the world, with possibly no or little contact with the spoken Võro and no access to Võro language courses. When designing the content of Võro Oahpa we are trying to meet the needs of both user groups. Our programs can mostly be used to support the students' individual training of vocabulary and grammar of the Võro language.

Other grammar learning programs we are aware of are e.g. Killerfiller (Bick, 2005) and ESPRIT (Koller, 2005). These are text-based ICALL systems where sentences are extracted from a corpus. In the system VIEW (Meurers et al, 2010), any webpage that is in the right language can turned into a grammar exercise. This is a fantastic system but concerning the Võro language, however, the material on the web is still quite small.

2 Existing resources and infrastructure

Thanks to the cooperation project we could make use of the Giellatekno and Divvun infrastructure (Moshagen et al, 2014) – a development infrastructure created to make it easier for people working on languages with limited textual resources to build language technology applications. The general idea is that (computational) linguists compose formalised grammar descriptions and lexicons, and the intrastructure makes it possible to use the lexicons and grammar as the basis for NLP tools (e.g. morphological and syntactic analyser) and end user tools such as proofing tools and electronic dictionaries. We got easily used to Giellatekno infrastructure that has standard places for language data (word lists, source code for the morphological transducer, Oahpa source files, documentation files, etc.) and standard procedures for the production of language technology tools and end-user programs out of these. The infrastructure is well suited for morphologically rich languages.

As one of our goals was to provide pronunciations for the people who live in the environment where they do not hear spoken Võro we decided to make use of the existing audio and text-to-speech resources.

One important Võro language resource is the online dictionary `http://synaq.org` that includes 15 000 entries in the direction Võro-Estonian and 20 000 entries in the direction Estonian-Võro. The dictionary also includes high quality audio files for Võro words. The audio files have been produced in cooperation of the Võro Institute with the Center of South Estonian Language and Culture and Laboratory of Phonetics, University of Tartu.

During the development of Võro Oahpa a prototype of Võro speech synthesis was developed at the Institute of Estonian Language. There are two voices to choose between: a middle aged man and a 11-years-old girl. The quality of the synthesized speech is good, very close to natural speech. The demo of the speech synthesis is available at the following URL: `www.eki.ee/~indrek/voru/index.php` and the software can be downloaded from here: `github.com/ikiissel/syntht s_vr`.

3 Our work: Võro Oahpa – a set of language learning programs

We have a previous experience of setting up Oahpa for a number of languages. Although the overall procedure of setting up a new instance of Oahpa is similar, each language has some specific issues that need to be dealt with. For the Võro language these issues were:

- extensive spell-relax
- many parallel forms

Spell-relax means that the program accepts different variants of typing for some characters. Checking of the correct answers must not be too strict because the written language is quite new (from 1990s) and there is no consensus on how to mark e.g. glottal stop and palatalisation; some letters of the Võro alphabet are missing from the keyboard layouts (there is no special Võro keyboard layout yet).

The illative and inessive plural of some nouns may attest to as many as 6-9 forms, e.g. the word *pereh* "family":
pereh+N+Pl+Ill: [*perrihe*, *perriihe*, *perride*, *perriide*, *perehtehe*, *perehtede*]
pereh+N+Pl+Ine: [*perrin*, *perriin*, *perrih*, *perriih*, *perrihn*, *perriihn*, *perehten*, *perehteh*, *perehtehn*], whereas the second person singular can attest to 3 if not 6 forms, e.g. the word *ehitelemä* "to decorate" *ehitelemä+V+Act+Ind+Prt+Sg2*: [*ehitelit*, *ehiteliq*, *ehitelideq*, *ehitellit*, *ehitelliq*, *ehitellideq*]

In the Oahpa exercises we need to decide which forms to accept as possible forms and which ones to display as correct answers. Whereas the parallel forms issue has to do with the morphology exercises Morfa-S and Morfa-C the relaxed spelling applies to all four games implemented in Võro Oahpa.

3.1 Multi-purpose side product – morphological finite state transducer of Võro

A finite state transducer (FST) incorporates both a morphological analyser and a generator. It defines correspondences between tag strings and word forms of a language. There exists a powerful FST development environment in the Divvun ad Giellatekno infrastructure. Using the standard file and tag names and other conventions makes it possible for a FST developer (linguist) to use the automatic build process that is taken care of by a number of filters and scripts. The compiled transducers can be used in several applications as language learning programs, online dictionaries, spelling checkers and machine translation tools. The Võro morphological transducer has so far been used in Oahpa and in the morphology-aware dictionary `http://sonad.uit.no`.

While building the morphological FST we have made use of the experience of developing morphology descriptions for other Uralic languages as the Saami languages, Erzya, Hill Mari a.o.

The problems we tackled when modeling Võro morphology were the following:

- Vowel harmony is not always predictable from the nominative or genitive singular forms, variation between singular and plural stem harmony, e.g. ("host") *esäk – esäku* genitive singular but *esäkidegaq* comitative plural.
- Consonant gradation [2] – as many as 4 grades: *häbü*, *häu*, *häpü* and *häppü* ("shame" nominative, genitive, partitive, illative singular).
- Many inflection types. Even if it seems that the word belongs to the same type there might be some forms in the paradigm that are different. The classification of nouns and adjective stem types has uncovered further irregularities, that might be dealt with through geographic/dialect classification.
- Parallel forms. For pedagogical purposes, it should be desirable that the preferred parallel forms are tagged differently from the non-preferred ones. Therefore, we have tagged all the non-preferred parallel forms with the tag *+Use/NG*. The non-preferred forms are accepted when the user enters those but not shown as correct answers.

We have applied the systematic error correction procedure of the FST:

1. All the simple words, i.e. derived and compound words excluded, have been generated by the FST as a large table.
2. A testing person has marked the errors in the table.
3. The errors have been corrected in the FST.

New subtypes of the inflection types for both nouns and verbs have been described in the FST as a result of this systematic work. For example, the noun types where singular nominative ends with a consonant but the stem vowel appears in genitive and other cases have been split by stem vowel to 3-4 separate types. That was implemented by introducing new continuation lexica.

[2] Consonant gradation is a type of consonant mutation where during the inflection either the length of a consonant is changing, a consonant is replaced by another consonant or a consonant is disappearing. E.g. *supi : suppi* ("soup" genitive singular vs partitive singular), *anda : anna* ("to give" infinitive vs connegative)

Currently all the 13260 yaml tests pass, i.e. the morphological FST generates correctly all the forms that are given in the tests.

The FST has also been tested on the running text (Võro wikipedia and children's book "Suur must koer"). The current testing results are presented in the table Table 1.

	Total	Missing	Missing %
All tokens	82 390	294 335	28%
Unique tokens	30 695	50 142	61%

Table 1: Evaluation results of the Võro FST.

For Oahpa the lexical coverage is good enough, as long as all the words that are in the Oahpa lexicon are in the FST. The most important thing is, however, that all the generated forms are correct. But in the longer perspective, of course, we aim at much better lexical coverage that would facilitate morphological analysis and spelling check of running Võro texts.

3.2 Online language learning tools (Oahpa games)

3.2.1 Numra – program for training numerals and date and time expressions

Numra is probably the simplest game that a beginner might start with. The easiest setup of the Cardinals game presents numerals 1-10 as the sets of five and the user's task is to guess which number corresponds to which word.

Three special finite state transducers were created to enable these exercises – a transducer of cardinal and ordinal numerals, a transducer of time expressions and a transducer of date expressions. The transducers define correspondences between numerical and textual representations of numbers, time points and dates.

3.2.2 Leksa – a vocabulary training program

Leksa is a classical vocabulary test where the user has to translate isolated words or everyday expressions from Võro to a metalanguage or vice versa. The drop-down menus enable the selection of words by topic (semantic category, sometimes in a broader sense): human, animal, food/drink, time, body, clothes, school, nature, work/economy, etc.

There are several metalanguages – Estonian, Finnish, English, German, North Saami, Norwegian, Swedish. This makes it possible for people with different language backgrounds to learn Võro vocabulary. To make Võro Oahpa more accessible we have also localised the whole user interface to Estonian, Finnish, English and Võro. The lexicon size of Leksa is ca 1300 words. The core of the lexicon comes from North Saami Oahpa (therefore we also have translations to North Saami and Norwegian for most of the words). But we have adapted the lexicon to our needs – removed some words that belong to Saami cultural space and added lists of frequently used Võro words with translations to some semantic classes (alltogether ca 300 words).

We have also added audio to Võro Leksa – a possibility to listen to the pronunciations of the Võro words. The pronunciations have been integrated from the sound database of the Võro-Estonian-Võro electronic dictionary *synaq.org*. The words have been read in by native speakers of Võro.

3.2.3 Morfa-S – a morphology drill program

Given the primary form (nominative singular for nouns and infinitive for verbs), the task is to build a specific inflected form. For nouns all the 14 cases in singular and plural can be practiced (except for essive that does not have separate singular/plural forms). For verbs there are exercises on indicative mood personal mode present and past tense first till third person in singular and plural, including negation forms. For adjectives we have exercises on positive and comparative grade. It is possible to practice their declination in all cases in singular and plural. Morfa-S exercises are based on isolated words.

3.2.4 Morfa-C – morphology exercises in the context

The Morfa-C game is based on question-answer templates and the word form database that also includes semantic information. Each exercise consists of a question and an answer where one word is replaced by a blank that the user has to fill with a word in the appropriate inflected form. The semantic tags are used to build semantically plausible sentences. Despite of that, the sentences sometimes come out funny or inappropriate. Is it okay to present a grammar exercise where the policeman steals (vro: *politsei varastas*) or a priest drinks vodka (vro: *keriguopõtaja juu viina*)? For more advanced students the humor can be on its place whereas it can be confusing for beginners (also unpedagogical for adolecents). Our solution was a very fine-grained semantic classification. For example, we have picked only the action verbs suitable for Morfa-C present and past tense verb inflection

exercises and added ca 50 verbs to this list. At the moment we have 151 semantic categories defined but the number will probably increase as we add new Morfa-C question-answer templates. Some semantic categories that we are using are listed in Table 2, together with the number of words in each category.

Semantic category	Nr of words
ANIMAL	71
BODYPART	41
FOOD_DISH	38
FOOD_GROCERY	36
CLOTHES	36
PROFESSION	20
FAMILY	20
WEATHER	10
SCHOOL	6

Table 2: Examples of semantic categories used in Võro Oahpa.

Another example. The question-answer pairs that are about buying and eating things require distinction between the food that can be bought from the grocery shop and the food that can be eaten as a meal. Often the food and drink words belong to both categories but not always. We also needed a special category for the food words that are natural to use in plural (things that we normally eat a plenty of, not only one, e.g. peas, berries, nuts). The lists of words denoting foods and drinks have also been extended with more foods and drinks that are common in Estonia or specific to South-East Estonia.

There is also a specific exercise for practicing back negation. Back negation has got a special attention because it exist neither in Estonian nor in Finnish. In Estonian, Finnish (and also in Võro parallel to back negation) the front negation is used where the negation word precedes the verb (e.g. *ei olõq* = "not is"). In back negation, the negation appears as a suffix that is added to the verb (e.g. *olõ-õiq* = "is-not"). There are more examples of back negation on Figure 1.

Morfa-C game in Võro Oahpa has a new feature that does not exist in any of the other implementations of Oahpa. Namely, the computer will read aloud the sentences (questions) using Võro synthetic voice (of a 11-years-old girl) when the user clicks on the loudspeaker icon.

A problem we have discovered was repetition of the identical exercises. This is partly due to the small number of words in some semantic sets but can still be avoided by improving the algorithm. There are three types of repetitions that we would like to eliminate:

1. Identical exercises within an exercise set consisting of three or five question-answer pairs should be prohibited.

2. It would also be good to avoid repetitions in the subsequent exercise sets. That is, if the user presses the button "New set" then the task words she had in the previous set should not occur in the new set of exercises, or even better – the words that she answered correctly should not occur but the words where she made a mistake could be presented again. But this idea is difficult or impossible to implement until we have not implemented the authentication of users and binding the usage data to specific users.

3. Avoid presenting the negatively loaded words (e.g. *ossõndama* "to vomit", *varastama* "to steal", *pelgämä* "to be frightened", *ullitama* "to act the fool") too often. That presumes a modification of the exercise creation process: weights should be assigned to the words (low weights to the words that should appear rarely) and these weights should be taken into account in the word selection algorithm.

3.3 Discussion

The most important question is: Would Võro Oahpa meet the users' needs?

We assume that most of the users are speakers of Estonian or Finnish. Therefore we need to focus on features of the Võro language that are different from Estonian:

- vowel harmony
- partially different case endings
- using of illative (the case corresponding to the English preposition "into") vs allative (the case corresponding to "onto"), inessive ("in") vs adessive ("on"), elative ("out of") vs ablative ("off"), particularly in connection with place names (there are place names that are used with different cases in Estonian and Võro)

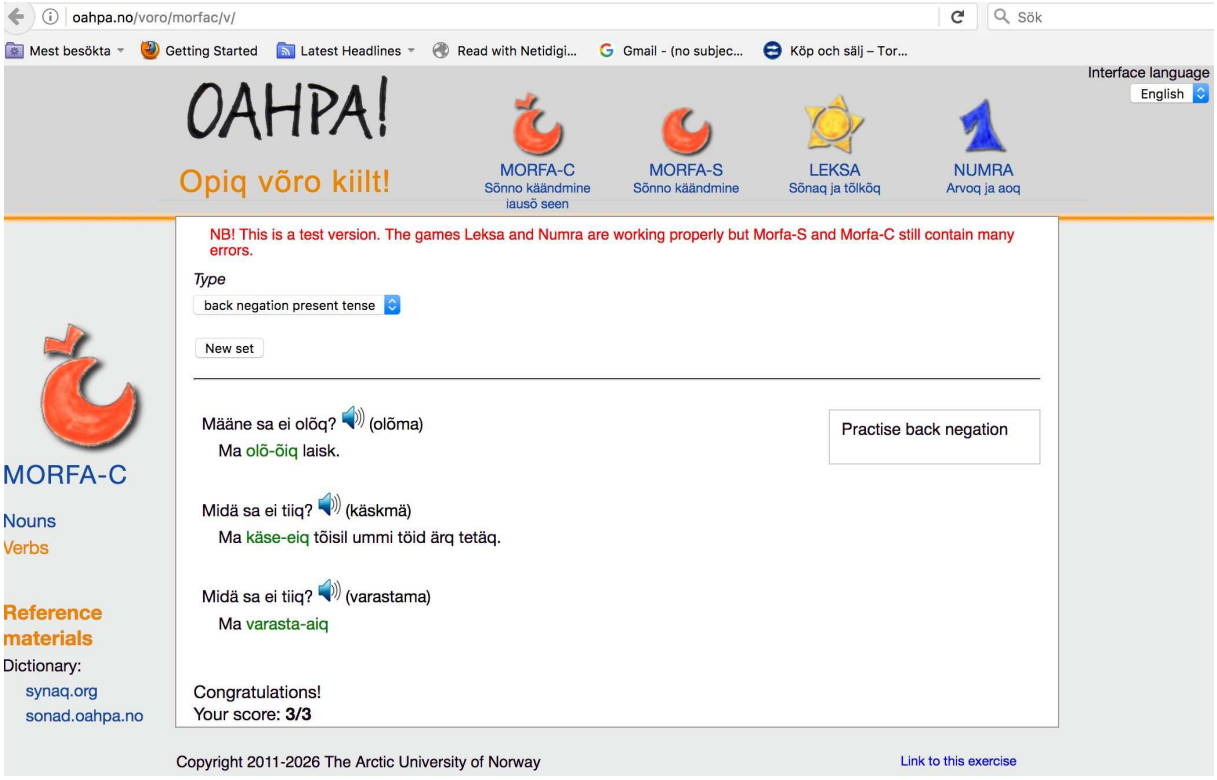

Figure 1: Screenshot of the Morfa-C verb back negation exercise

- two different ways of building negation: front negation (*ei olõq*) and back negation (*olõ-õiq*)
- different negation word in present vs past tense (*ei olõq* = "is not", *es olõq* = "was not")
- palatalisation mark in the written language
- more extensive use of diminutive
- pronouncation (especially important for the people who live outside of South-East Estonia)

All of the above, except for vowel harmony, also holds for Finnish speakers.

We also have to think about users with other mother tongues. Features that might be difficult for people with non-Uralic mother tongue:

- many morphological forms
- vowel harmony
- pronouncation
- usage of all the cases

All the listed topics are in fact included in Oahpa exercises in either implicit or explicit way but we need to create more specific exercises to make the learner pay attention to the particular features of Võro. For example, we have specific exercises in Morfa-C for practicing back negation and using the correct negation word (*ei* or *es*) but we should also create some special exercises on difficult inflection types, vowel harmony rules and diminutive building.

Võro Oahpa is free to use for everybody on the URL `http://oahpa.no/voro`. The authors will be grateful for any feedback about the system.

4 Conclusion

In this article we have presented our work on Võro language learning programs. This is the first freely available program for Võro that gives the users the possibility to train the basic 1300 words vocabulary, date and time expressions and morphology. While setting up and developing the programs we have made use of the Divvun and Giellatekno infrastructure as well as Võro language resources that were either created externally (online Võro-Estonian-Võro dictionary `synaq.org` where we got the pronuncations of the Võro words from and software for Võro speech synthesis) or within the same project (Võro morphological transducer). We

can confirm that the infrastructure was helpful for our work. The biggest challenge is modeling the Võro morphology – covering all the inflection types, marking the preferred and non-preferred parallel forms and handling the different ways of spelling. Adding the audio dimension adds extra value to Võro Oahpa as many of the program's prospective users are not exposed to spoken Võro. Reading aloud the Morfa-C questions is the feature that is totally new – it has not been implemented in any of the previous instances of Oahpa. The work on Võro Oahpa is continuing to enable practicing of larger vocabulary and more of the grammar.

Acknowledgments

The work has been funded by The Research Council of Norway project EMP160 "Saami – Estonian language technology cooperation: similar languages, same technologies".

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, 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.

Eckhard Bick 2005. *Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL*. Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004: 171–185 Museum Tusculanums Forlag.

Thomas Koller 2005. *Development of web-based plurilingual learning software for French, Spanish and Italian*. Studies in Contrastive Linguistics. Proceedings of the 4th International Contrastive Linguistics Conference (ICLC4). University of Santiago de Compostela Press.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, Niels Ott 2010. *Enhancing authentic web pages for language learners*. Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications: 10–18

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).

Parijõgi M. 2017. Kool on kodukeele viimane kants. (School is the last stronghold of the home language) *Opetajate Leht*, 10.03.2017.