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

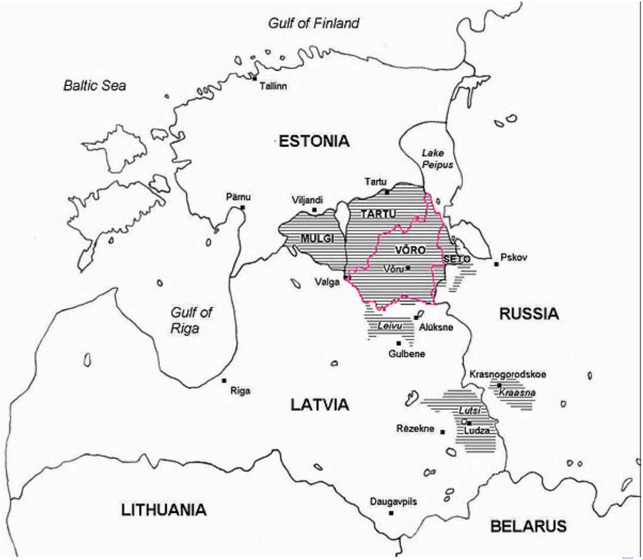
Heli Uibo (University of Tartu & UiT The Arctic University of Norway)
Jack Rueter (University of Helsinki)
Sulev Iva (University of Tartu & Võro Institute)

May 22, 2017

The Võro language – a minority language in Estonia

- Võro is a descendant of the old South Estonian regional language which is least influenced by Standard Estonian (which is based on Northern Estonian dialects).
- The Võro enthusiasts are fighting for giving Võro the status of a language, instead of a dialect.
- The language has its own ISO code **vro**.

Võro language area



The Võro language

The Võro language is closer to Finnish than standard Estonian.

- Finnish and Võro are more conservative than Estonian.
- Vowel harmony
- i-plural and i-past tense
- The old Baltic Finnic vocabulary has been preserved.

The Võro language – 2

Other features distinct from Estonian

- Post-negation, vro: *olõ-i* et: *ei ole* ("is not")
- Different negation word (*ei* vs *es*) in present and past tense, *"ei opiq, es opiq"* ("does not learn, did not learn")
- Nouns have 13 cases (not 14 as in Estonian), the essive case does not exist.
- Inflectional endings differ from both Estonian and Finnish.
- Võro has some phonemes that do not exist in Estonian: both strong and weak africate **ts** (*tsiga* pig) and **ds** (*küdsä* "(he/she) bakes"), glottal stop, denoted by **q** (*lauluq* "songs").
- Diftong **ai** instead of **ei**. et: *seisma*, vro: *saisma* ("to stand").
- Palatalisation – a larger number of palatalised consonants compared to Estonian, and the palatalisation is marked in the written form.

Our leading idea: Reuse the existing language resources and infrastructures to build language learning tools

Võro language resources

- Võro-Estonian-Võro online dictionary (15 000 Võro-Estonian entries and 20 000 Estonian-Võro entries)
- Võro morphological finite state transducer (the development started in 2013 within the current project)
- Võro text-to-speech software (Institute of the Estonian Language, 2016)

Reuse of existing infrastructure

- Divvun and Giellatekno infrastructure at UiT the Arctic University of Norway was created to facilitate the development of language technology tools for under-resourced languages.
- The development of both building blocks, e.g. morphological analysers, and end user products as spelling checkers, electronic dictionaries and language learning tools is supported.
- The infrastructure is well suited for morphologically rich languages.
- Samest – a collaboration project between UiT and U Tartu.

Oahpa!

Oahpa is a set of interactive language learning programs where exercises are generated automatically:

- Leksa (vocabulary drill)
- Numra (spelling out numbers and time expressions)
- Morfa-S (building of grammatical forms)
- Morfa-C (building of grammatical forms in the sentential context)
- Vasta (writing grammatically correct answers to the questions generated by the program)
- Sahka (dialogue game)

All the six modules are implemented in North Saami Oahpa `oahpa.no/davvi`.

Oahpa instances that consist of 2-4 modules exist for 20+ languages.

Võro morphological finite state transducer (FST)

- At the outset of the project we had a formal description of the morphology (Iva, 2007) but no computerised implementation.
- The development of the FST started in parallel with the development of Oahpa.
- We have made use of the experience of developing morphology descriptions for other Uralic languages as the Saami languages, Erzya, Hill Mari a.o.
- Divvun and Giellatekno infrastructure is used for the development.
- Thanks to the infrastructure the compiled transducers can easily be used in end user applications: Oahpa and the morphology-aware dictionary `http://sonad.uit.no`. The first prototype of the spelling checker has also been created.

Võro morphological FST: difficulties

- many inflection types
- partly irregular vowel harmony
- consonant gradation
- many parallel forms

Võro FST: difficulties: consonant gradation

Up to four degrees in the consonant gradation (the usual number in North Saami and Estonian is two).
E.g. the word *häbü* ("shame"):
häbü - sg nominative
häu - sg genitive (**b** has disappeared)
häpü - sg partitive
häppü - sg illative

Võro FST: difficulties: parallel forms

- Example: The illative and inessive plural may have 6-9 parallel forms, e.g. the word *pereh* "family":
  pereh+N+Pl+Ill: [*perrihe*, *perriihe*, *perride*, *perriide*, *perehtehe*, *perehtede*] ("into the families")
  pereh+N+Pl+Ine: [*perrin*, *perriin*, *perrih*, *perriih*, *perrihn*, *perriihn*, *perehten*, *perehteh*, *perehtehn*] ("in the families")
- The tag +Use/NG is used to mark the forms that should be accepted as possible forms but not displayed as correct answers.

Võro FST: testing results

The FST has been tested on Võro Wikipedia (Accessed 2016-08-29).

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

Table: Evaluation results of the Võro FST.

Not good enough for spelling check or morphological analysis of the running text but sufficient for Oahpa because its vocabulary is limited to the learner's dictionary.

Oahpa: what was special about Võro?

- Need to handle many parallel forms, some of which are acceptable but not recommended.
- The learning programs should be tolerant to different spelling variants.

Handling of parallel forms

- The tag +Use/NG is used in the FST to mark the forms that should be accepted as possible forms but not displayed as correct answers.
- This is taken into account when generating the Oahpa database.

Tolerance to several spelling variants

The Võro written language is relatively new, no standard orthography yet.

- Palatalisation mark – the standard is modifier letter apostrophe but also other apostrophe-like characters are accepted. *pall’o, palĺo, pall’o, pall‘o* (”much, a lot of”)
- Glottal stop – conventionally denoted by the letter q but omitted by many writers of Võro. E.g. both *poisiq* and *poisi* (”boys”) are accepted.

Use of audio in Oahpa

Goal: to provide pronunciations for the people who live in the environment where they do not hear spoken Võro.

- Leksa – isolated words, the sounds from the database of the online dictionary `http://synaq.org`. Recorded by the native speakers.
- Morfa-C – speech synthesis is used for reading aloud the questions.
- Speech synthesis has not been used in any other instance of Oahpa before.

Leksa demo

`http://oahpa.no/voro/leksa`

Morfa-C demo

http://oahpa.no/voro/morfac

Usage

Now:

- Leksa and Numra have been used within the introductory course of the Võro language at Uni of Tartu.
- Morfa-S and Morfa-C might still contain errors but the FST quality is improving all the time, so from the beginning of the next term Morfa-S and Morfa-C will also be used.

In the future:

- The language courses at Võro Institute.
- Kindergartens and schools in South-Eastern Estonia. (ca 450 children are learning Võro language and culture or participating in other classes where the language of instruction is Võro).
- Individual training of vocabulary and grammar.

Võro Oahpa is free to use for everybody, no registration required!

Conclusion and lessons learned

- Reusing the existing resources and infrastructures speeds up the process of building NLP-based language learning programs.
- The most time-consuming work is the construction of the morphological FST. This work is still in progress but we have almost covered the whole morphology but the lexicon needs to be extended to gain better coverage than 72%.
- It is very important to have a practicing language teacher permanently in the team.