Sami - Estonian language technology cooperation: similar languages, same technologies

Heiki-Jaan Kaalep[1] Trond Trosterud[2]

[1]TÜ

[2]Giellatekno, Centre for Saami Language Technology
http://giellatekno.uit.no/

September 20, 2017

Giellatekno

TARTU ÜLIKOOL 1632 UNIVERSITAS TARTUENSIS

norway grants

Introduction

- Starting point: shamelessly selfish
  - Estonians:
    - "have ling. software, need infrastructure for applications"
    - "have ideas, need infrastructure for implementing"
  - Norwegians: "have infrastructure for applications, open for more languages to add"
- Project idea: goals are compatible
- Outcome: "Languages added, applications online, + new applications made for this project reused for other languages as well"

What was the aim of the project?

- Robust modules for Estonian and Võro morphology
  - ... that can be included in software products
  - ... like Sámi has been already
- Machine translation (MT)
  - Finnish-North-Sámi
  - Finnish-Estonian
- Interactive language learning (iCALL)
  - Estonian, Võro
  - ... like Oahpa! for Sámi

What did we spend our time on?

- Computational linguistics
  - FST morphology (Estonian, Võro)
  - CG disambiguation (Estonian, Finnish)
  - Lexicon building (Finnish-Estonian)
- Integrating them in MT, iCALL
- Results are operational, but not perfect

iCALL: Oahpa!

- Vocabulary
- Inflection
- http://oahpa.no/voro/

Rule-based machine translation - why?

- Contrastive linguistics in practice
- Systematic similarities, differences
- Checking the linguistic claims (automatically)

Rule-based machine translation

- Solved
  - Word + grammatical categories $\leftrightarrow$ word form
    vaimoille $\rightarrow$ vaimo+Pl+All $\rightarrow$ naine+Pl+All $\rightarrow$ naistele
- Not solved
  - vocabulary is too small (15,000)
  - word choice:
    joka = iga kes mis
  - gramm. category conversion
- etc etc ...

Rule-based machine translation: example

echo 'Pitääkö vaimosi lyhyistä lomista ?'

Rule-based machine translation: example

```
echo 'Pitääkö vaimosi lyhyistä lomista ?'
| apertium -d . fin-est
Kas su naisele meeldivad lühikesed puhkused ?
```

Rule-based machine translation: example

echo 'Pitääkö vaimosi lyhyistä lomista ?'
| apertium -d . fin-est
Kas su naisele meeldivad lühikesed puhkused ?
Google Translate:
Kas sul on naine lühikeste vahedega ?

Future perspectives

- FST development – is going on, and Estonian gives new perspectives on FST manipulation
- ICALL – Tromsø now has a new programmer, Võro is going on + good prospects (is used in teaching)
- Core technology is continuing via the Linux Debian development,
- GT infra: new languages, new techniques (weighting, context, mobile..., grammar checking)
- MT – is continuing – langtech with an industry to back it up...
- MT - collaboration with Institute of Estonian Language
  - large Finnish-Estonian dictionary (70,000 entries)
  - for MT, lexicography

Conclusion

The project is still going on (refuses to die...)

- Bad: there are unfinished things
- Good: we are now better at solving them (and work together)