Morphological analyzers and other digital tools for Uralic languages

Jack Rueter

University of Helsinki, Digital Humanities

An outline

- Some Terminology
  - Word associations
  - Working out a methodology
- Morphological analyzers under development
- Where it started
- Setting up a rule-based description
- Tools
- Important players

Some terminology

- Open-source
- Language form
- Rule-based
- Finite-state morphology
- Disambiguation
- Multiple reusability
- Linguistics and Language technology
- Neural networks
- Transfer learning
- Active learning

Word associations for multi-reusability

- (1) Keyboards
- (2) Spellers
- (3) Dictionaries
- (4) ICALL
- (5) Translation
- (6) Text-to-speech
- (7) ….

Searching for a methodology for linguistics

- (1) Extract paradigms from grammars, readers and research to build an analyzer.
- (2) Extract words, part-of-speech information and definitions from existing dictionaries and research. Build on what has already been done (Dutch, French, German, Russian,…)
- (3) Test analysis coverage on written texts. Are the forms unrecognized proper words?
- (4) Disambiguate morphological analyses based on grammars and research. Point out gaps in descriptions
- (5) Test syntactic disambiguation on example sentences cited in grammatical descriptions of the language. And then retest on text corpora.
- (6) Make disambiguated sentences public, so others can test. One by-product of these golden standards are treebanks.
- (7) Use all phases to benefit the speaker and research community

Open-source morphological descriptions for Uralic languages

- First transducers of minority Uralic languages after Finnish 1983 (Kimmo Koskenniemi)
  - Meadow Mari ~1986 (Jorma Luutonen)
  - Komi-Zyrian 1996 (Jack Rueter)
- Giellatekno ~2000 begins work with Sami descriptions (Trond Trosterud et al)
  - Barents Sea languages, Circum Polar languages
  - ~2004-> other Uralic languages

Minority Uralic language forms with finite-state morphology development

- Balto-Finnic: fit = Meänkieli, fkv = Kveen, izh = Ingrian, krl = Karelian, liv = Livonian, olo = Olonets-Karelian aka Livvi, vep = Veps, vot = Votic, vro = Võro
- Sami: sjd = Kildin Sami, sje = Pite Sami, sma = South Sami, sme = Northern Sami, smj = Lule Sami, smn = Inari Sami, sms = Skolt Sami
- Mordvin: mdf = Moksha, myv = Erzya
- Mari: mhr = Meadow & Eastern Mari, mrj = Hill Mari aka Western Mari
- Permic: koi = Komi-Permyak, kpv = Komi-Zyrian, udm = Udmurt
- Ob Ugrian: kca = Khanty, mns = Mansi
- Samoyedic: nio = Nganasan , sel = Selkup, yrk = Nenets

Pertinent majority languages with finite-state morphology development

- Uralic languages in majority: est = Estonian, fin = Finnish, hun = Hungarian
- Auxilliary languages: deu = German, lav = Latvian, nob = Norwegian Bokmål, rus = Russian, tat = Tatar

Setting up a Morphological analyzer

- Find a source and use the known morphological information
- Find or build a lexicon to propogate this word type

Use known paradigmatic information, Ingrian (Junius 1936)

	Yksikkö.	Monikko.
Nom.	varis	varikse-**t**
Gen.	varikse-**n**	variks-**i(i)-n**
Akk.	varikse-**n**	varikse-**t**
Part.	varis-**t**a	variks-**i-a**
Illat.	varikse-**e**	variks-**i-i**
In.	varikse(e)-**s**	variks-**i(i)-s**
El.	varikse-**st**	variks-**i-st**
All.	varikse-**lle**	variks-**i-lle**

Make a test file

```
Noun - varis: # Noun 'crow'
   varis+N+Sg+Nom: varis
   varis+N+Sg+Gen: variksen
   varis+N+Sg+Acc: variksen
   varis+N+Sg+Par: varista
   varis+N+Sg+Ill: variksee
   varis+N+Sg+Ine: [variksees, varikses]
   varis+N+Sg+Ela: variksest
   varis+N+Sg+All: varikselle
   varis+N+Pl+Nom: varikset
   varis+N+Pl+Gen: [variksin, variksiin]
   varis+N+Pl+Acc: varikset
   varis+N+Pl+Par: variksia
   varis+N+Pl+Ill: variksii
   varis+N+Pl+Ine: [variksiis, variksis]
   varis+N+Pl+Ela: variksist
   varis+N+Pl+All: variksille
```

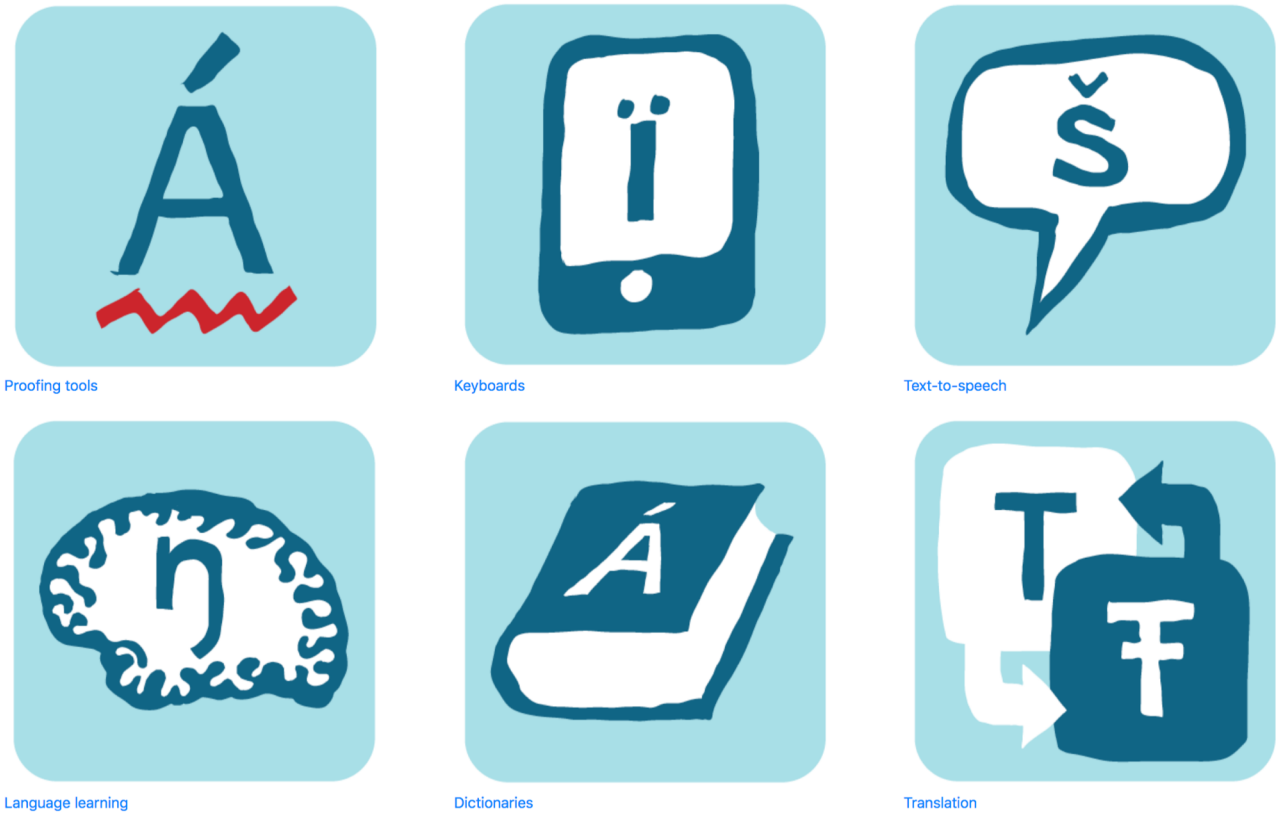
Generate an initial paradigm type

```
LEXICON N_VARIS ! varis:vari
+N+Sg+Nom:s
+N+Sg+Gen:ksen
+N+Sg+Acc:ksen
+N+Sg+Par:sta
+N+Sg+Ill:ksee
+N+Sg+Ine:ksees
+N+Sg+Ine:kses
+N+Sg+Ela:ksest
+N+Sg+All:kselle
+N+Pl+Nom:kset
+N+Pl+Gen:ksin
+N+Pl+Gen:ksiin
+N+Pl+Acc:kset
+N+Pl+Par:ksia
+N+Pl+Ill:ksii
+N+Pl+Ine:ksiis
+N+Pl+Ine:ksis
+N+Pl+Ela:ksist
+N+Pl+All:ksille
```

Propogate the lexicon

```
varis+N:vari N_VARIS  "" ;
ohjas+N:ohja N_VARIS  "" ;
toitus+N:toitu N_VARIS  "" ;
toohus+N:toohu N_VARIS  "" ;
toohus+N:tuohu N_VARIS  "" ;
toohus+N:tuuhu N_VARIS  "" ;
tutkimus+N:tutkimu N_VARIS  "" ;
toimitos+N:toimito N_VARIS  "" ;
lyhennös+N:lyhennö N_VARIS  "" ;
modus+N:modu N_VARIS  "" ;
parennos+N:parenno N_VARIS  "" ;
jätös+N:jätö N_VARIS  "" ;
peenennys+N:peenenny N_VARIS  "" ;
yhtehös+N:yhtehö N_VARIS  "" ;
vaihtiimistapahus+N:vaihtiimis#tapahu N_VARIS  "" ;
kysymys+N:kysymy N_VARIS  "" ;
painutos+N:painuto N_VARIS  "" ;
```

Divvun for language community and technology



Tools

- Keyboards Giellalt/
- Spell checkers: Hunspell, Voikko
- Click-in-text dictionaries
- Language learning
- Text-to-speech
- Translation

Giella Dictionaries

- Click-in-text dictionaries,
  - Giella:
    - https://sanit.oahpa.no/
    - https://saan.oahpa.no/ , https://sanat.oahpa.no/, https://valks.oahpa.no/, https://muter.oahpa.no/, https://kyv.oahpa.no/, https://vada.oahpa.no/
- Language internal and external links
  - Akusanat:
    - https://www.akusanat.com/,
    - https://www.akusanat.com/verdd
    - https://www.akusanat.com/semantics
- Material Collaborations: FU-Lab, University of Turku, University of Tartu, University, EKI, University of Vienna, Livones, Võro Instituut

Language learning

- ICALL at Giella (Giellatekno & Divvun)
- Northern Sami (Flag ship) http://oahpa.no/davvi

Translation

- Apertium: Shallow-transfer for closely related languages
- http://wiki.apertium.org/wiki/Main_Page

Text-to-speech

- Common Voice (Mozilla)
- https://voice.mozilla.org/en

Infrastructures with intense activity

- EKI in Tallinn, Estonia
  - http://portaal.eki.ee/sonaraamatud.html
- FU-Lab in Syktyvkar, Komi Republic
  - https://fu-lab.ru/fulabteam
- Giella in Tromsø, Norway
  - http://giellatekno.uit.no/
- Mari Research Institute for Language, Literature and History
  - http://marnii.ru/
- UL Livonian institute
  - http://www.livones.net/liv
- Võro Instituut
  - https://wi.ee/

Aitäh!

Thank you!