Morphological Descriptions of 5 Minority Uralic Langauges on Giellatekno 2013-2014.

Blumberga, Kuprina, Rantakaulio, Rueter, Salo, Silfverberg, Soosaar

Congressus Duodecimus Internationalis

Fenno-Ugristarum

Oulu, Finland, August 17th – 21st, 2015

Abstract

The «Morphological Analyzers for Minority Finno-Ugrian Languages» Project was conducted in the Giellatekno infrastructure at Tromsø, Norway, with funding from the Kone Foundation in Helsinki.

Researchers participated from Finland and Estonia, and there was synergy with colleagues from numerous university departments, as well as library projects.

Transducer descriptions and translation work was made available the first month of the project:

http://giellatekno.uit.no/all-lang.fin.html

https://victorio.uit.no/langtech/trunk/

Multi-usage of the morphological descriptions were experimented with: Web dictionaries, spell checkers, ICALL, OCR, rule-based translation,...

Goals

Extensive developmenty in Giellatekno Infrastructure

Achievements surpass initial expectations

5 languages 20,000 Finnish translations per language has varied results

Goals problems with measurement

Words difficult to count across languages

Workers spent extra time looking for second and third meanings.

Funding

Kone Foundation provided scholarships:

6 researchers for 2 years

Computers, programs

OCR seminar

Giellatekno provided extensive interaction on interface use and organization of seminars.

Dates

Weeks training January 2013 (Giellatekno)

Demonstrating 1st month at Tiit-Rein Viitso seminar (March 7.-8., 2013)

Regular expressions for describing Ingrian, Erzya, Vepsä, Moksha, Hill Mari, Meadow Mari (April 2013)

Machine translation à la Apertium with test materials for Olonets Karelian, Moksha, Livonian, Erzya, Hill Mari and Võro (May 2013)

Dates 2

Digi Day presentation with National Library of Finland Kindred Language Pilot (June 6, 2013)

Presnetation in Alcanena, Portugal: (*The development of finite-state transducers for small Uralic languages – the case of Tundra Nenets*) (October 17, 2013)

Presentation at Paris Inalco, (November 2013)

Two-directional web dictionaries for all 5 languages (January 2014)

The Livonian-Estonian-Latvian Dictionary as a Threshold to the Era of Language Technological Applications (May 2014)

Dates 3

Hill Mari taught in Hungary

LREC Conference Reykjavik extended work including Latvian (May-June 2014)

OCR course applying HFST (June 2014)

ICALL course (September 2014)

Direct contacts at Oma Mua in Petrozavodsk (December 2014)

Dates 4

Evaluation of Olonets Karelian project begun by professional writer (January, 2015)

OCRicola presentation in Tromsø, Norway (January, 2015)

Language technology seminar in Saransk, Mordovia. Spellcheckers introduced in editorial staffs and university departments (April, 2015)

Evaluation and enhancement of Moksha begins (April-October, 2015)

Evaluation of Livonian begins (June, 2015)

Following progress on the web

* Analyzers in the Giellatekno infrastructure:

http://giellatekno.uit.no/cgi/index.izh.fin.html

http://giellatekno.uit.no/cgi/index.liv.fin.html

http://giellatekno.uit.no/cgi/index.mdf.fin.html

http://giellatekno.uit.no/cgi/index.mrj.fin.html

http://giellatekno.uit.no/cgi/index.olo.fin.html

http://giellatekno.uit.no/cgi/index.yrk.fin.html

Vocabulary and morphology

* X-kielestä suomennettujen sanastojen tilastoista käy ilmi seuraavasta:

http://www.ling.helsinki.fi/~rueter/AKU/finnish-translations_stats_2015-05-20.html

* Sanamuotoja tekstintunnistusta varten

http://www.ling.helsinki.fi/~rueter/AKU/OCRWordForms/

* Säännöllisiä lausekkeita, suomalais-ugrilaisten vähemmistökielten kuvausmalleja:

http://www.ling.helsinki.fi/~rueter/AKU/OcrRegex/

Web dictionaries

http://sanat.oahpa.no/ Olonets Karelian

http://sonad.oahpa.no/ Livonian, Ingrian

http://valks.oahpa.no/ Moksha, Erzya

http://vada.oahpa.no/ Tundra Nenets

http://muter.oahpa.no/ Hill Mari, Meadow Mari

http://kyv.oahpa.no/ Komi, Udmurt

Voikko Spell checkers

http://www.ling.helsinki.fi/~rueter/AKU/OXT/

Open-source

Continued work in Livonian, Moksha, Erzya, Komi, Udmurt

Contributions to ICALL

http://testing.oahpa.no/livokel/Livonian

http://testing.oahpa.no/mdf_oahpa/ Moksha

http://testing.oahpa.no/mrj_oahpa/ Hill Mari

http://testing.oahpa.no/olo_oahpa/ Olonets Karelian

http://oahpa.no/yrkoahpa/ Tundra Nenets

http://testing.oahpa.no/bxr_oahpa/ Buryat

http://oahpa.no/erzya/ Erzya

http://oahpa.no/kpvoahpa/ Komi-Zyrian

http://oahpa.no/aanaar/ Inari Sami

http://testing.oahpa.no/udm_oahpa/ Udmurt

OCR experimentation

* OCRicola on vapaa tekstintunnistustyökalu morfologisesti monipuolisia kieliä varten.

http://sourceforge.net/p/ocricola/wiki/AddNewLanguage/

http://www.helsinki.fi/~mpsilfve/ocr_course/

Future

The floor boards have been laid

Morphological analyzers

X-Finnish translations!!

What can you do?

Use the open-source morphological analyzers, OAHPA, spellcheckers, web-dictionaries, OCR

Get involved: test, extend, find new uses

Advocate multilingual facilitation for these beautiful languages

Evidence of a Future

- Vienna makes use of Hill Mari morphology from Giellatekno
- Estonian Wikimedia is looking into incorporating Voikko spellcheckers in «Minority Translate»
- Work in Livonian is incorporating our dictionaries and morphology for initiated syntax
- Evaluation of Karelian Language Form affinity project expresses interest in what is being done for Olonets, Ludic and Karelian
- University of Turku Udmurt-Finnish dictionary will soon join the fold
- University of Turku Meadow Mari-Finnish dictionary available
- ICALL project begun with Erzya
- University of Szeged Udmurt-Hungarian dictionary collaboration

THANK YOU!

- Researchers: Renate Blumberga, Julia Kuprina, Timo Rantakaulio, Merja Salo, Sven-Erik Soosaar, Miikka Silfverberg
- Coworkers and developers: Giellatekno, HFST, Kindred Language Digitization, SUKI, Voikko
- Previous development: Academic work in Tartu, Turku, Syktyvkar, Saransk, Joshkar Ola, Izhevsk, Helsinki...
- Funding: Kone Foundation