THE LIVONIAN-ESTONIAN-LATVIAN DICTIONARY AS A THRESHOLD TO THE ERA OF LANGUAGE TECHNOLOGICAL APPLICATIONS

Jack Rueter
University of Helsinki

Abstract. This article outlines the multiple use of electronic source materials from the Livonian-Estonian-Latvian Dictionary of 2012 in a “Kone Foundation” funded project for developing finite-state morphological parsers. It provides an introduction to the project, the language-independent Giellatekno infrastructure at Tromsø, Norway, and the materials utilized in the electronic manuscript of the dictionary. The introduction is followed by an extensive description of what has been developed on the Giellatekno infrastructure with explicit indications of where parallel projects might be initiated.

Keywords: Livonian, Uralic languages, “Kone” language programme, open-source, language-independent infrastructure, HFST, Giellatekno, morphological parser, spellcheckers, morphology-savvy web dictionary, intelligent computer-assisted language learning

DOI: http://dx.doi.org/10.12697/jeful.2014.5.1.14

1. Introduction

Reuse of source materials intended for two-dimensional or actually any publications is something to strive for. When the “Livonian-Estonian-Latvian dictionary” appeared in 2012, one of the first ideas conceived was to extract as much of the knowledge contained in the publication for reuse in open-source language technological applications. The materials could be used to set up a finite-state description of the language, and the results would be available for use in further open-source application development. All of this presupposes a project, infrastructural work, and well developed dictionary material.

1.1. The “Morphological Parsers for Minority Finno-Ugrian Languages” project

In December 2012 the “Kone Foundation” in Finland granted scholarship funding for a language-technological development project which would provide open-source morphological parsers for five

Uralic languages. The languages to be documented were the Balto-Finnic Livonian and Olonets-Karelian; Moksha Mordvin; Hill Mari, and Tundra Nenets. The purpose of the project was to develop regular morphological descriptions to the extent of 20,000 lemmas per language over a two-year period (2013–2014). Each of the lemmas would be provided with Finnish-language glosses. The open-source nature of the project would promote further open-source development of the new resources. It was observed that where possible utilization of already existing electronic materials for these languages would greatly accelerate development.

1.2. A prominent language-independent infrastructure

Since Livonian has a relatively complex morphology, setting up a finite-state description of it might be done analogous of Finnish and Sámi descriptions. Like Sámi, however, research must consider seeking an infrastructure strategy where a little effort goes a long way. Such an infrastructure exists in Tromsø, Norway, which is known today as Giellatekno (GT). Giellatekno is the Sámi language technology center where work in the finite-state description of circumpolar and small indigenous languages has been carried out since the end of the last millennium.

The GT infrastructure is open-source and run by open-minded individual researchers. Morphology-wise much of the applied infrastructure is based on language technologies proven in work with languages exhibiting ample inflectional paradigms. In fact, some of the core open-source technologies originate across the border, in Helsinki, namely, Two-Level Description (Twol) by Kimmo Koskenniemi, and Helsinki Finite-State Technology (HFST), and the Constraint Grammar (CG) by Fred Karlsson.

In the beginning of 2013, Giellatekno offered room for transducer construction with technical support. Morphological analyzers and generators were available on the GT web site, and there was plenty of documentation openly available for other language projects. Reuse of transducer development at that time was partially directed at automatically generated spellcheckers; if you had an analyzer you basically had a spellchecker. There was also work directed at web-based morphology-savvy dictionaries for languages with transducers and language pair translations. In fact, the more you look at the GT infrastructure, the more you see: computer-assisted language learning environments, rule-based syntax, rule-based translation cooperation with Apertium in Spain, text-to-speech research and more.

Giellatekno is the infrastructure for introducing languages with both minimal and ample research resources. This is where The Kone Foundation funded “Morphological Parsers for Minority Finno-Ugrian Languages” project would commence work on transducers for Livonian, Olonets-Karelian; Moksha; Hill Mari, and Tundra Nenets. Workers at Giellatekno willingly helped in conducting an indoctrination course in the beginning of 2013 for project members and others interested in the GT infrastructure from Finland, Estonia and Norway.

1.3. The materials

The Livonian-Estonian-Latvian dictionary of 2012 is an erudite documentation of Livonian language study results. The most profound of the facets are: the extensive paradigmatic classifications of nominal and verbal stems; the alignment of language pair translations, as well as the electronic format of the manuscript. It would be unheard of to neglect the reuse of this document in electronic media.

The dictionary is accompanied by an extensive paradigmatic classification of Livonian nominals and verbs. Approximately 240 nominal, and 60 verbal stem types have been illustrated in the appendices. This classification work can be immediately applied to a finite-state description of the Livonian language and transducer testing.

The bulk of the dictionary consists of Livonian lemmas with translations of simple words, compound words, idiomatic phrases and contexts for illustrating word usage. Language-pair translations are Livonian-Estonian and Livonian-Latvian. There are also indications of synonymy and domain affiliations for some of the words.

The electronic format of the manuscript allows for extraction of simple lemmas, paradigmatic information, as well as other segments of macro entries.

Each macro entry is organized in such a way that the trilingual element can be readily segmented into child elements and language-aligned sub-elements. This makes it possible to extract only those elements which are pertinent to the construction of a transducer.

2. Work from Day One on

With the consent of the authors and copy of the electronic manuscript secured the “Kone Foundation” funded project in Helsinki commenced work on a finite-state transducer description of the Livonian language.

From day one there was documentation available online to indicate our progress, as well as an analyzer and generator that could be used by anyone.

Documentation:
http://giellatekno.uit.no/cgi/index.liv.eng.html

Here documentation several aspects of Livonian language documentation. The prominent elements of the documentation include: Two-level models for representing the orthographical interpretation of phonological rules governing morphological inflection in Livonian; Lemma and stem pairs with Finnish-language glosses; Affixation strategies for simultaneously tagging and inflecting words from various parts-of-speech, and yaml tests for generation and analysis of two different transducer models of Livonian.

Analyzer:
http://giellatekno.uit.no/cgi/d-liv.eng.html

Generator:
http://giellatekno.uit.no/cgi/p-liv.eng.html

Here the analyzer may prove highly valuable. Text can be dropped into the analyzer window of sizable quantities (1–195 word units), and the resulting morphological analysis is soon available. This could be utilized by researchers in search of ambiguous analyses. It would provide stimuli for the construction of a disambiguating Constraint Grammar for the language, which would corroborate the accuracy of grammar work, on the one hand, and be useful in future reuses of the transducer in text-to-speech or machine translation. Such CG disambiguators are available already for some of the GT infrastructure languages.

After a mere two months of work, day-and-night, a morphology-savvy web dictionary of Livonian was presented in early March of 2013. It was possible to demonstrate an operating system where at least some of the Livonian words on a site could be demonstrated. Development could also be assessed: If the non-headword morphology could be recognized but no translation was available, it meant there was more Finnish translation work to be done. If, however, the morphology was not recognized, the translation could still be assessed with the lemma.

Morphology-savvy web dictionary:
http://sanat.oahpa.no/liv/fin/

The developing online dictionary was available in two formats. There was a window for entering individual word forms for analysis and translation, on the one hand, and a bookmarklet that could be opened for use on individual pages. The bookmarklet provided word-by-word translation when the user held down the alt-key and double clicked a word on an html page. The html page could also be simply a text document with html beginning and end tags on the users home computer.

With the success of the Livonian-Finnish morphology-savvy online dictionary apparent, moves were quickly made to apply the same Livonian transducer to the already existing Livonian-Estonian and Livonian-Latvian translation pairs. These dictionaries were available by early April, 2013.

Additional morphology-savvy web dictionaries for Livonian:
http://sanat.oahpa.no/liv/est/
http://sanat.oahpa.no/liv/lav/

By late December of 2013 nearly ten thousand Finnish translations were available to gloss the Livonian lemmas of the morphological analyzer. The analyzer had also improved notably. Testing materials for transducer development had been directly derived from the electronic paradigm tables for nominal and verbal stems. Spell relaxes had also been set up to accommodate a variety of user needs.

Yaml tests:
https://victorio.uit.no/langtech/trunk/langs/liv/test/src/morphology/

Spell relaxes:
https://victorio.uit.no/langtech/trunk/langs/liv/src/orthography/

The online dictionary can be accessed by mobile devices that do not necessarily have keyboards for fulfilling the requirements of the orthography. To simplify usage and yet retain the accuracy of the transducer elsewhere, special filters were introduced for the so called social media. These filters made if possible to type the letter “a” and have it recognized as an array of different letters: “a”, “ā”, “ä” and “ǟ”. Similar filters had already been applied to the analyzer for allowing letters with cedilla marking where a combining comma was expected.

Social media variant:
http://sanat.oahpa.no/livM/fin/

In January, 2014 the web dictionaries were augmented in yet another direction; an automatically reversed dictionary Finnish-Livonian had arrived. Albeit, automatically reversed materials do not provide for high-quality precision or coverage, but they do provide an ample starting point for work in this direction. Reversed materials also help locate short-comings in the database, and this is another way to evaluate our work. This Finnish-Livonian rendition of the morphology-savvy was made possible by the existence of an open-source Finnish finite-state transducer (OMorFi). At present an analogous Estonian transducer is under construction, but there is nothing to report on a Latvian transducer, yet.

Automatically reversed dictionary:
http://sanat.oahpa.no/fin/liv/

On February 28th, 2014, Kalevala Day in Finland, a new dimension of our reusable transducer work was unveiled. A beta version was now being produced for proof-reading tools in LibreOffice for both Microsoft and Macintosh platforms that would allow Livonian text proof-reading. The necessary files can be downloaded at the following pages.

Beta proof-reading tools for LibreOffice:
http://divvun.no/libreofficeoxt.html
http://www.ling.helsinki.fi/~rueter/AKU/beta_eng.shtml

In anticipation of teaching in Livonian, minor work has also begun in parallel to the OAHPA learning environments for the Sámi languages[1]. A text-to-numeral generator, originally conceived for text-to-speech in Southern Sámi, has been constructed for Livonian as well. Forth-coming will be work involving semantic tagging of the Livonian-Finnish language pair which automatically applies to the existing Livonian-Estonian and Livonian-Latvian language pairs to boot.

[1] At Giellatekno a decision has been made to introduce infrastructure for a dozen new Oahpa-type learning environments for certain Uralic and North American minority languages in the first half of 2014. These environments will require financially supported developers representing language instruction for these languages.

Numeral generators:
http://giellatekno.uit.no/num.eng.html

The Northern Sámi Oahpa
http://oahpa.no/index.eng.html

Extensions made to the original paradigm structure

The Livonian transducer has been extended to provide paradigm coverage for all words introduced. This, it must be noted, entails overgeneration, and means the parser will analyze words and word forms which have not previously been attested in corpora. The down-side is that there will be need for work in syntax to solve the problems of disambiguation, hence more documentation and linguistic description of the language.

Derivation

There are now possibilities for automatic derivation analysis in the Livonian transducer. One of the furthest developed is deverbal derivation of nouns, adjectives (participles) and adverbs (gerunds).

There is still work to be done on adjective-to-noun derivation and elsewhere, to name but one feature.

Statistics on the Livonian transducer development and Finnish glossing of lemmas is best illustrated in comparison with other Uralic language projects in the GT infrastructure. At present statistics cover only those two facets of development with reference to several minority Uralic languages other than the Sámi languages: Olonets-Karelian, Moksha, Hill Mari, Tundra Nenets (Kone funded transducer and translation project); Ingrian, Kven, Veps, Võro; Erzya; Meadow Mari; Komi Zyrian (Kone funded Finnish translation), Udmurt; Northern Khanty; Nganasan.

Statistics:
http://www.ling.helsinki.fi/~rueter/AKU/statistics_eng.shtml

3. Summary

The open-source project funded by the “Kone Foundation” to produce a morphological analyzer for Livonian and Finnish-language glosses of the lemmas in the analyzer code, has provided the Livonian language with multiple facets for future research and development. This achievement would not have been possible without the concept of open-source, the presence of the Giellatekno language-independent infrastructure, and the reusable materials of the Livonian-Estonian-Latvian Dictionary manuscripts.

The original project is clearly on its way to the realization of a 20,000 word stem analyzer with Finnish-language glossing. This can be observed in open-source documentation, web-based parsers, morphological-savvy web dictionaries, morphological spellcheckers and budding language learning environments.

In late 2014 the project will have presented a host of directions for future development:

(1) Word stem and lemma documentation in the analyzer beyond 20,000
(2) Rule-based syntax for disambiguation of Livonian morphology
(3) Development of previously overlooked text resources
(4) Development of Livonian language learning environments
(5) Development of Finnish to Livonian dictionary work

Address:
Jack Rueter
Department of Modern Languages
PO Box 24
00014 University of Helsinki
Finland
E-mail: Jack.Rueter@helsinki.fi

Online resources

Beesley, Kenneth and Lauri Karttunen (2003) *Two level rule compiler*. Available online at <http://www.stanford.edu/~laurik/.book2software/twolc.pdf>. Accessed on 17.03.2014.

CG = *Constraint grammar*. Documentation available on line at <https://victorio.uit.no/langtech/trunk/langs/sme/src/syntax/> and <http://beta.visl.sdu.dk/constraint_grammar.html>. Accessed on 17.03.2014.

GT = *Giellatekno*. Available on line at: http://giellatekno.uit.no. Accessed on 17.03.2014.

HFST = *Helsinki finite-state transducer technology*. Available on line at <http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/>. Accessed on 17.03.2014.

Kone Foundation language programme (English) Available on line at <http://www.koneensaatio.fi/index.php/download_file/view/457/244/>. Accessed on 17.03.2014.

OMorFi = *Open morphology for Finnish*. Maintained by Tommi Pirinen. Available online at <http://www.ling.helsinki.fi/kieliteknologia/tutkimus/omor/>. Accessed on 17.03.2014.

TWOL = *Two-level model*. (cf. Beesley and Karttunen 2003.)

Viitso, Tiit-Rein and Valts Ernštreits (2012) *Līvõkīel-ēstikīel-leţkīel sõnārõntõz.* [Livonian-Estonian-Latvian dictionary.] Available online at <http://www.murre.ut.ee/liivi/>. Accessed on 17.03.2014.

Kokkuvõte. Jack Rueter: Liivi-eesti-läti sõnaraamat lävepakuna keeletehnoloogiliste rakenduste ajastusse. Artikkel annab ülevaate elektroonilise lähtematerjali “Liivi-eesti-läti sõnaraamat 2012” mitmekülgsest kasutamisest Kone fondi rahastatud projektis morfoloogiliste analüsaatorite arendamiseks. Artikli sissejuhatav osa esitab sissevaate projekti ning Tromsøs loodud keelest sõltumatusse Giellatekno taristusse; tutvustatakse ka sõnastiku elektroonilises käsikirjas kasutatud materjale. Seejärel kirjeldatakse Giellatekno tarkvara arendusega loodud võimalusi ning tuuakse näiteid sellest, kuidas saab sarnaseid projekte algatada.

Märksõnad: liivi keel, uurali keeled, Kone keeleprogramm, avatud lähtekood, keelest sõltumatu infrastruktuur, HFST, Giellatekno, morfoloogiline analüsaator, õigekirjakontroll, “Morphology-savvy” veebisõnastik, arvutipõhine keeleõpe

Kubbõvõttõks. Jack Rueter: Līvõkīel-ēstikīel-leţkīel sõnārǭntõz nemē kīeltehnolōgilizt kȭlbatimizt āiga kīndõks. Kēra āndab iļļõvaņtļimiz iļ “Līvõkīel-ēstikīel-leţkīel sõnārǭntõ” amāpūoļiz kȭlbatõmiz, laz kazāntõg loptõb morfolōgiliži analīzijidi Sīe projektrǭ āndaji um Kone fond. Kēra klīerõb īžpīlijizt Giellatekno infrastruktūrõ, mis um lūodõd Tromsøs Norvēgjis, ja sõnārǭntõ elektrōnilizõs kädkēras kȭlbatõd materiālidi. Nei īž kēra nīžõb iļ võimizt, mis tarmõb Giellatekno lūodõd programvīļa ja nägţõb, kui tämvītliži projektidi võib irgtõ.