

Apertium-fin-sme: machine translation between Finnish and Northern Sámi

Ryan Alexander Johnson

April 9, 2010

Name:

Ryan Alexander Johnson

E-mail address:

ryan.txanson@gmail.com

Other information that may be useful to contact you:

- **Cell phone:** +47 9417 7119
- **IRC:** pyry` on freenode
- **Jabber:** ryan.txanson@gmail.com
- **Sourceforge:** rtxanson

Why is it that you are interested in machine translation?

I have a strong background in linguistics, and have spent about two years working in web development and a fairly geeky lifetime learning all sorts of things relating to these things. Machine translation and language technology in general seems to be a sort of ideal marriage of both of my major skill-sets. I see machine translation as an ideal way to help aid minority languages in the production of new resources, be it books, dictionaries, spell-checkers or machine translators.

I have always had a latent interest in machine translation and language technology, and Google Summer of Code seems like the ideal way to challenge myself, and gain a strong footing on some of the recent developments in the field. When I was younger, I remember

programming morphological generators for Quenya, one of Tolkien's constructed languages for the elves, on my TI graphing calculator in TI-BASIC. Recently, the interest has come back, and I've begun teaching myself a little of finite-state morphologies with the Xerox Finite-State Tools¹. Now it's time to put these interests and skills to use.

Why is it that you are interested in the Apertium project?

I became acquainted with one of the project members, Francis Tyers, while attending a machine translation workshop at the University of Tromsø in Fall '09, and this was essentially my introduction to Apertium. Apertium captured my attention, because it is open source, and anyone can contribute; and Apertium strives to use open data formats and open data sources. Open source translation software seems to be an ideal way to address the needs of various minority language communities, where it may be more difficult to find top-down support from governments, but where support may be found either in the community in peoples' free time, or from projects such as Google Summer of Code that spring up from time to time to provide a little funding.

Apertium's mix of statistical and linguistic translation is particularly interesting, because linguistic translation methods are absolutely necessary for languages that are more morphologically complex. Statistical translation clearly works well with some of the languages available in something like Google Translate, but it is most successful when the target language is morphologically simple, and when there is a vast amount of text available in both of the languages in the pair being translated. Something like Inuktitut, or Northern Sámi would be much more difficult for statistic translation methods.

Which of the published tasks are you interested in? What do you plan to do?

I am interested in adopting a new language pair: Finnish and Northern Sámi, and begin working in the direction of Finnish to Northern Sámi. Getting a good start on this language pair would be great particularly for Northern Sámi as a language, and as a project within Apertium. In addition to benefiting Northern Sámi, it would provide a starting point for Finnish in Apertium, and hopefully in the long run involve more of the Finnish linguistics and computational linguistics community. With a foot in the door for Finnish, other translation pairs like Fin-Eng, or Fin-Swe could be started with less trouble. This project also builds on existing work in Apertium with Northern Sámi (specifically, the Northern Sámi/Norwegian Bokmål language pair).

There are plenty of resources that already exist to make Fin-Sme possible:

¹XFST: <http://www.cis.upenn.edu/~cis639/docs/xfst.html>

- Omorfi: Finnish morphological analyzer²
- Joukahainen: Finnish wordlists³
- Giellatekno's⁴ Northern Sámi morphological analyzer and generator, and Northern Sámi constraint grammar
- Apertium's Sme-nob language pair
- Fred Karlsson's Finnish Constraint Grammar (see below)

This project relies crucially on Fred Karlsson's Finnish Constraint Grammar, which was previously not available for public use, but Fred has agreed to open it up for use with Apertium. The other tools are already available for use: the Finnish morphological analyzer and, Northern Sámi language technology is completely comprised of open source tools available from Giellatekno. As such, beginning a project with Fin-Sme doesn't necessarily entail starting from the absolute basics. Most of the work within Apertium will thus be in disambiguation and transfer rules; significant lexicons and morphological analysis and generator tools already exist for both languages, and thanks to the work on Sme-Nob, some automated methods for improving lexicons already exist.

There are also additional bilingual text resources that exist, and organizations that may be approached to see what data they are willing to share, or data that is essentially already open via GPL or fair use:

- The New Testament in Finnish and Northern Sámi
- The Sámi Parliament's speaker's bilingual blog in Finnish and Northern Sámi⁵
- KOTUS⁶ maintains an etymological database of the Sámi languages⁷ with thousands of words and translations into Finnish and other Sámi languages. Although their data is not expressly available for use in projects such as this, it is available for free in a browsable format online. They would be worth contacting to find out about whether their whole database is accessible for other language technology projects.

Title

Apertium-fin-sme: machine translation between Finnish and Northern Sámi

²Omorfi - Open Morphology for the Finnish language: <http://home.gna.org/omorfi/omorfi/README.html>

³Joukahainen: <http://joukahainen.puimula.org/>

⁴Sámi Giellatekno - Sámi Language Technology Group: <http://giellatekno.uit.no/>

⁵Suoma Sámedikki Ságadoalli: <http://klemetti.blogspot.com/>

⁶The Research Institute for the Languages of Finland: <http://www.kotus.fi/>

⁷Álgu Database: <http://kaino.kotus.fi/algu/>

Why should Google and Apertium sponsor it?

Google and Apertium should sponsor this project for several reasons: the project is beneficial to Apertium, which gains a new starting point for several language pairs with Finnish as well as an implementation of a Finnish Constraint Grammar and Finnish morphological tools. Any benefit to Apertium results in benefits to machine translation as a whole, and may result in bringing in more enthusiastic linguists from Finland, or with background or specializations in Finnish to produce additional language pairs. There are also benefits to Google as well, with its interest in machine translation. Fin-sme also represents a benefit to the Northern Sámi language community, which will gain a new inlet for language resources.

This project is ideal for Apertium because it takes advantage of existing Apertium projects involved with Northern Sámi: some materials from the Sme-Nob project may be shared with Fin-Sme. This project could also rely on work produced by other Google Summer of Code projects with Apertium, such as the project involving integration of the Helsinki Finite-State Tools⁸. In addition to contributing to use of HFST instead of Ittoolbox, this project will contribute to the use of Constraint Grammar with Apertium, thereby further developing these tools. As a result, Apertium will offer several parallel tools for different levels of the linguistic analysis. These tools are relevant for morphologically complex languages, as well as languages with profound syntactic differences. Good results from this project will have real importance for other translation pairs wanting to use Constraint Grammar and HFST.

For Google, it will be relevant to look at the limitations of Google Translate. Traditionally, it is claimed that the technology behind Google Translate is restricted to morphologically poor languages. Still, Finnish is supported within Google Translate, and there is indeed literature⁹ about the merits and challenges of statistically-based machine translation for Finnish. Developing Finnish for Apertium will make it possible to give a broader comparison of the two approaches than what has so far been the case.

How and who will it benefit in society?

As a machine translation project that results in translation into a minority language, it is easy to say that there are many benefits to society. Machine-aided translation from Finnish to Northern Sámi will allow quicker translation of all sorts of things, textbooks, newspaper articles, literature. New translations into Northern Sámi could also benefit the Lule Sámi community as work progresses on the Northern Sámi->Lule Sámi (sme-smj) project.

⁸HFST: <http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/>

⁹Google Scholar: <http://scholar.google.com/scholar?q=statistical+machine+translation+finnish>

Work plan

There are three main steps to this project, which will begin by utilizing already existing resources for Finnish and Northern Sámi. The first step will be to work with Omorfi, the Finnish morphological analyzer and Fred Karlsson's (generously donated/freed) Finnish Constraint Grammar, in order to convert Karlsson's constraint grammar¹⁰ tags to work with Omorfi and VISL CG-3¹¹. The next step will be to proof the existing data set to get it up to par. An existing dataset is already checked in in Apertium's SVN repository on SourceForge¹² and a tagged and aligned lexicon with probabilistic entries has been compiled from versions of the New Testament in Northern Sámi and Finnish.

Some of the proofing work may be automated, so some time can be spent looking at developing a proofing tool that either does bilingual lookups in other languages, such as fin-eng, or fin-nob; or a tool that just speeds the human-end of proofing vast amounts of text. In the end, a data set of 6000-10000 entries will be necessary for it to be of any use, so some expansion of the lexicon will likely be necessary. Expansion and testing of the lexicon can be done following the ideas behind Testvoc¹³, but will need to be modified to take into account the fact that HFST will be used instead of ltoolbox.

Once there is a complete set of data that is parseable and available in the bilingual dictionary, work can begin with transfer rules. This step will involve looking at the differences between Northern Sámi and Finnish and writing a comparative grammar. Then, tests will be written to handle the divergence between these languages. This step will also involve managing derivations of the words and managing compound words/multiwords.

Miscellaneous tasks

Producing additional documentation for this project is a must, as it will use some new tools which are not in use in other Apertium language pairs. Documentation can be produced on the go.

¹⁰This constraint grammar has been recently opened for our use by Fred Karlsson, but is not yet available.

¹¹VISL CG-3: http://beta.visl.sdu.dk/constraint_grammar.html

¹²SourceForge repository: <https://apertium.svn.sourceforge.net/svnroot/apertium/incubator/apertium-sme-fin/>

¹³Testvoc: <http://wiki.apertium.org/wiki/Testvoc>

Week plan

Week 1	Convert Fred Karlsson's CG tag set to the Omorfi tag set
Week 2	Begin working with Karlsson's CG with VISL CG-3
Deliverable	<i>Converted Constraint Grammar</i>
Week 3	Create a Testvoc process using HFST analyzers for Fin and Sme
Week 4	Work with existing and new texts; expanding lexicon with Testvoc
Week 5	Continue adding new data, as necessary
Deliverable	<i>Initial bidirectional dictionaries</i>
Week 6	Beginning comparative grammar, and creating test sentences
Week 7	Comparative grammar investigation of noun phrases, writing transfer rules
Week 8	Comparative grammar investigation of verb phrases, writing transfer rules
Deliverable	<i>Transfer grammar</i>
Week 9	Begin writing transfer rules, and transfer tests
Week 10	Further work on transfer rules
Deliverable	<i>Transfer rules</i>
Week 11	Wrap-up revision: run testvoc process on lexicons, continue testing sentences
Week 12	Wrap-up week
Final Product	Tested language pair

List your skills and give evidence of your qualifications

I have a B.A. in Linguistics from the University of Minnesota, with a minor in Finnish. This minor in Finnish is the result of the completion of the Finnish language program at the University of Minnesota, and was preceded by a year of self-study and followed up with a study abroad program at the University of Helsinki in Finland. Coursework there supplied valuable coursework in Finnish language morphology and phonology, as well as general linguistics courses taught in Finnish; and some Finno-Ugrian studies courses in Karelian¹⁴ and Northern Sámi. As a result, I have a detailed knowledge of Finnish grammar, dialect variation and related Balto-Finnic languages as well as some general knowledge of the history of Finnic languages. My studies have given me many insights into the structure of Finnish that could assist me in this project.

Spending time in Tromsø is not complete without learning some Northern Sámi, and I have been doing just that. I consider my skills to be "conversational", so not on the level of Finnish; but, I am able to read texts, express myself in speech, and have knowledge of the overall grammatical structure of the language.

¹⁴http://en.wikipedia.org/wiki/Karelian_language

My knowledge of programming is mostly self-taught and comes from a life-time of general geekery, which was recently used in a professional setting working as a web developer and tech support analyst at the Office of Information Technology at the University of Minnesota for two years following graduation. I have knowledge of several scripting languages: Python, Perl, PHP and Bash; and experience managing linux servers. Since I worked in web development, I have experience with XML and XSLT, and although they're unrelated to machine translation, I have experience with all sorts of other things related to modern standards-oriented open source web development, SQL databases, XHTML and CSS; web security and usability. I consider myself to be a well-rounded web developer, although I have a definite preference to Python.

I have worked some with Apertium as well. I attended a machine translation workshop in Fall '09, where I gained an introduction to the basics, and have since been teaching myself with the tutorials available on the Apertium wiki, and asking for help occasionally in the #Apertium channel on IRC.

List any non-Summer-of-Code plans you have for the summer

Since I am currently living in Norway on a self-financed masters program, I need to spend some additional time in the summer working. The timeline for Google Summer of Code means that I will have plenty of time after projects are announced to plan a work schedule that ensures that I have 30 hours a week for GSoC. I anticipate that I will always be able to find this time; I'm quite the inspired free-time programmer, and like to spend my time learning and studying. Of course, GSoC wouldn't be "free-time" per se, but I generally spend about 30 hours a week on my own projects when there is nothing else to do; so GSoC will fit into this schedule perfectly.