

# Automatic translation from North to Lule Sámi

December 10, 2008

North and Lule Sámi are closely related languages. In a test project, we picked a part of the Lule Sámi and North Sámi lexicon and made use of our simple North-Lule Sámi bilingual dictionary to start a rule-based automatic machine translation program in Apertium <http://www.apertium.org>, which has been used for a number of Romance language such as Spanish, Catalan, Portuguese, French, Occitan, Galician, but also Basque and English.

We started out with a set of 14 testsentences, most of them taken from the North Sámi Wikipedia.



The screenshot shows the Apertium website interface. On the left, there is a navigation menu with links for Main Page, Community portal, Current events, Recent changes, Random page, Help, and Donations. The main content area displays the title "Northern Sami and Lule Sami/Regression tests" and a list of 14 test sentences in Northern Sami with their corresponding English translations. The sentences are:

- (sme) *Wikipedia dárbaša du veahki.* → Wikipedia dárbaš duv viehkev.
- (sme) *Sámegiela Wikipedia leat dál 2,717 artiikkala.* → Sámegiela Wikipedia leat dál 2,717 artiikkala.
- (sme) *Isak Saba 1875-1921 lei sápmelaš oahpaheaddji ja politiikkár.* → Isak Saba 1875-1921 lei sápmelaš oahpaheaddji ja politiikkár.
- (sme) *Son beroštii arkeologalaš bargguin, ja son čokkii dávviriid ja dávtiid boares hávdái.* → Son beroštii arkeologalaš bargguin, ja son čokkii dávviriid ja dávtiid boares hávdái.
- (sme) *Girjái Áillohaš lea čuovvun historjjálaš govaiddáid sámiin birra máilmmi.* → Girjái Áillohaš lea čuovvun historjjálaš govaiddáid sámiin birra máilmmi.
- (sme) *Sámiid dahjege sápmelašat ássat Ruoššas, Suomas, Ruotas ja Norggas.* → Sámiid dahjege sápmelašat ássat Ruoššas, Suomas, Ruotas ja Norggas.
- (sme) *Sámiid ássanguovlu gohčoduvvo Sápmiin.* → Sámiid ássanguovlu gohčoduvvo Sápmiin.
- (sme) *Odne gávdnojit maidái ollu sámiid geat eai šat sápmás ja identitehtavuoddu lea dál varas dehe sogas iige gielas gitta gudii e destii sápmás ja identitehtavuoddu lea dál varas jalli berajvuodan iige gielas gitta.* → Odne gávdnojit maidái ollu sámiid geat eai šat sápmás ja identitehtavuoddu lea dál varas dehe sogas iige gielas gitta gudii e destii sápmás ja identitehtavuoddu lea dál varas jalli berajvuodan iige gielas gitta.
- (sme) *Amerihká ovtastuvvan stáhtain lea 50 oassestáhta, mat leat oalle muddui iešstivrejeaddjit.* → Amerihká ovtastuvvan stáhtain lea 50 oassestáhta, mat leat oalle muddui iešstivrejeaddjit.

Our first goal is to get them to work.

Francis Tyers (Aperitum, Alicante) and Linda Wiechetek (Giellatekno, Tromsø) are currently working on the project.

## 1 First results:

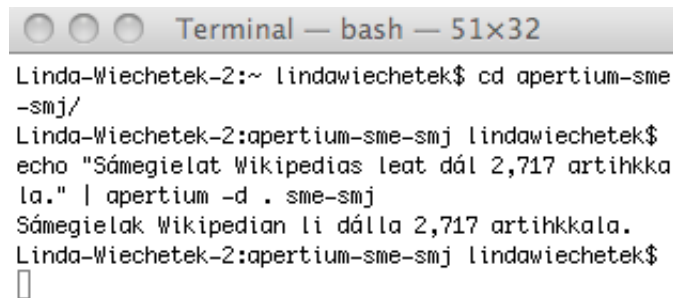
If I want to translate a sentence from North Sámi (*sme*) to Lule Sámi (*smj*), I write for example:

```
echo "Sámegielaht Wikipedias leat dál 2,717 artihkkala".  
| apertium -d . sme-smj
```

and I get the following output

Sámegielaht Wikipedian li dálla 2,717 artihkkala.

In the terminal, this looks as follows:

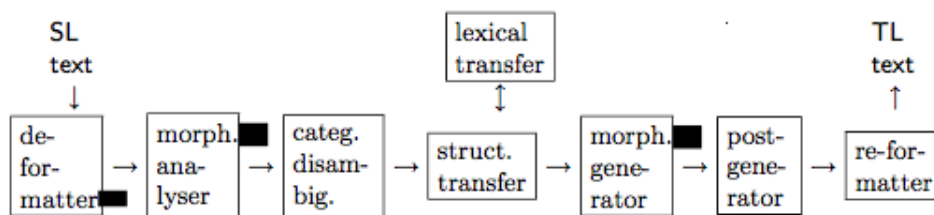


```
Terminal — bash — 51x32  
Linda-Wiechetek-2:~ lindawiechetek$ cd apertium-sme  
-smj/  
Linda-Wiechetek-2:apertium-sme-smj lindawiechetek$  
echo "Sámegielaht Wikipedias leat dál 2,717 artihkka  
la." | apertium -d . sme-smj  
Sámegielaht Wikipedian li dálla 2,717 artihkkala.  
Linda-Wiechetek-2:apertium-sme-smj lindawiechetek$  
□
```

## 2 Prerequisites

In order to get a working MT system, we need different components:

1. a number of lexica
  - the smj monolingual lexicon **apertium-sme-smj.smj.dix**
  - the sme monolingual lexicon **apertium-sme-smj.sme.dix**
  - a bilingual lexicon from sme to smj **apertium-sme-smj.sme-smj.dix**
2. a PoS annotator, a syntactic parser
3. (structural) rules
  - **apertium-sme-smj.sme-smj.t1x**
  - **apertium-sme-smj.sme-smj.t2x**
  - **apertium-sme-smj.sme-smj.t3x**
4. a postprocessor **apertium-sme-smj.post-smj.dix**



### 3 Challenges

- make structural rules that determine when sme locative is translated into smj elative or inessive in a way the computer understands it (e.g. habitive constructions, adverbials of stative verbs)
- get a good bilingual dictionary, right now the dictionary is incomplete and we mostly work with one-to-one relations between words
- postprocessing, orthographic context-dependent changes in smj (such as forms of liehket in vowel contexts)
- change of word order, when does SVO become SOV and again how do we get the computer to distinguish that

The default translation of North Sámi locative is with Lule Sámi elative:

```

echo "Son beroštii arkeologalaš bargguin, ja son čokkii
dávviid ja dávttiid boares hávddiin."
| apertium -d . sme-smj

```

```

Sån berustij arkeologalasj bargojs, ja sån tjåkkij
dávverijt ja dávtijt boares hávdijs.

```

When the sentence contains a stative verb, such as *sset*, locative is translated as inessive.

#### Translation of sme locative with stative verbs such as *ásset*

```

echo "Sámit dahjege sápmelaččat ássset Ruoššas, Suomas ja
Norggas."
| apertium -d . sme-smj
... årru Ruossjan, Suoman ja Vuonan.

```

## 4 Goal

The ultimate goal is to have everything on a nice graphical interface and make it freely available such as other rule-based machine translation programs as Apertium and GramTrans

[what is apertium](#)

[who develops it](#)

[downloading](#)

**[test drive](#)**

[Text translation](#)

[Document translation](#)

[Surf & translate](#)

[DicLookUp](#)

[Apertium unstable](#)

[Apertium alpha-testing](#)

[documentation](#)

[interact!](#)

[latest news](#)

[software](#)

### Text translation

Plain text translation using Apertium 3.0

Translation type: Spanish → Catalan

El catalán (català), también llamado valenciano (valencià) en la Comunidad Valenciana, es una lengua romance occidental que procede del latín vulgar.

Mark unknown words

### Translation

El català (català), també anomenat valencià (valencià) en la Comunitat Valenciana, és una llengua romanç occidental que procedeix del llatí vulgar.