Developing prototypes for machine translation between two Sámi languages

Francis Tyers ftyers@prompsit.com Universitat d'Alacant, Spain
Linda Wiechetek linda.wiechetek@uit.no Tromsø University, Norway
Trond Trosterud trond.trosterud@uit.no

The languages

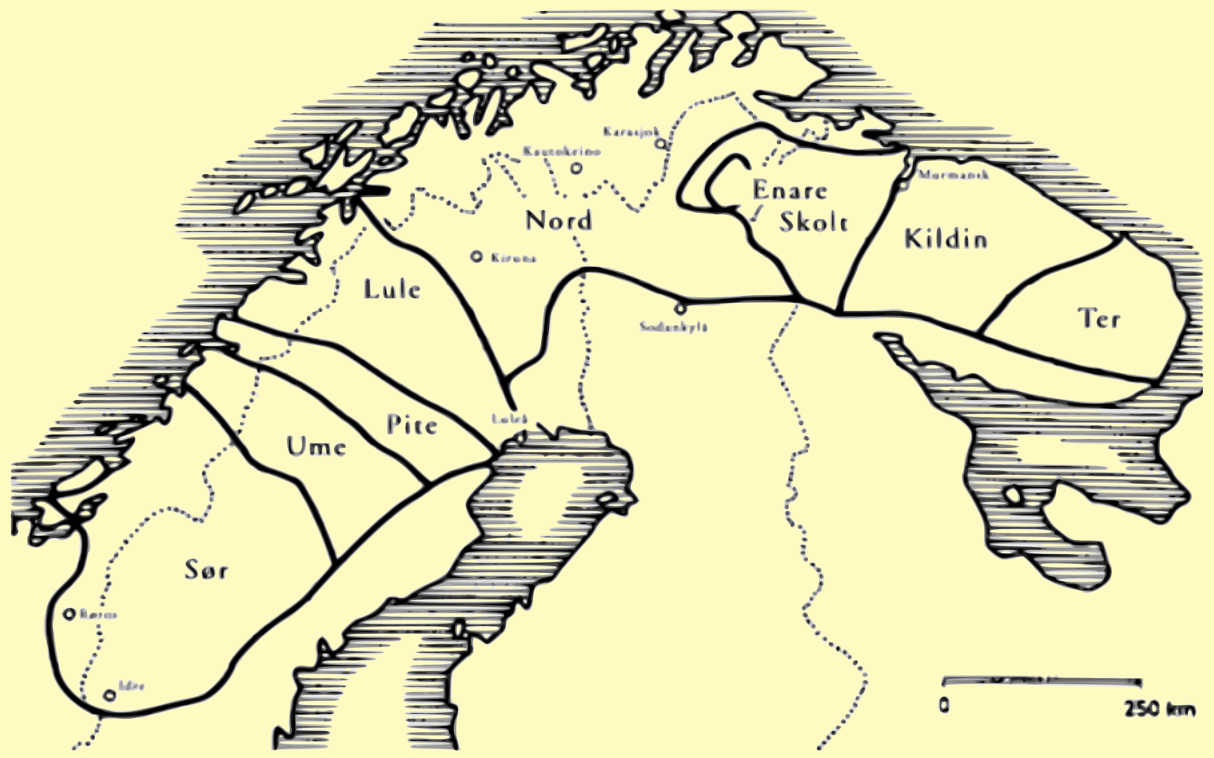


Figure 1 Map of *Sápmi* showing the Sámi language areas

North (N.) Sámi and Lule (L.) Sámi,

- belong to the Finno-Ugric language family
- are spoken in the north of Norway and Sweden, North Sámi also in Finland
- have 15,000 - 25,000 (N.) and < 2,000 (L.) Sami speakers
- are heavily inflective *and* agglutinative
- are largely mutually intelligible

A machine translation system for this pair must be of a quality such that post-editing the output is faster than translating from scratch, since the users will prefer the original to a bad translation.

Existing resources

There are resources for both morphological analysis and disambiguation of N. and L. Sámi

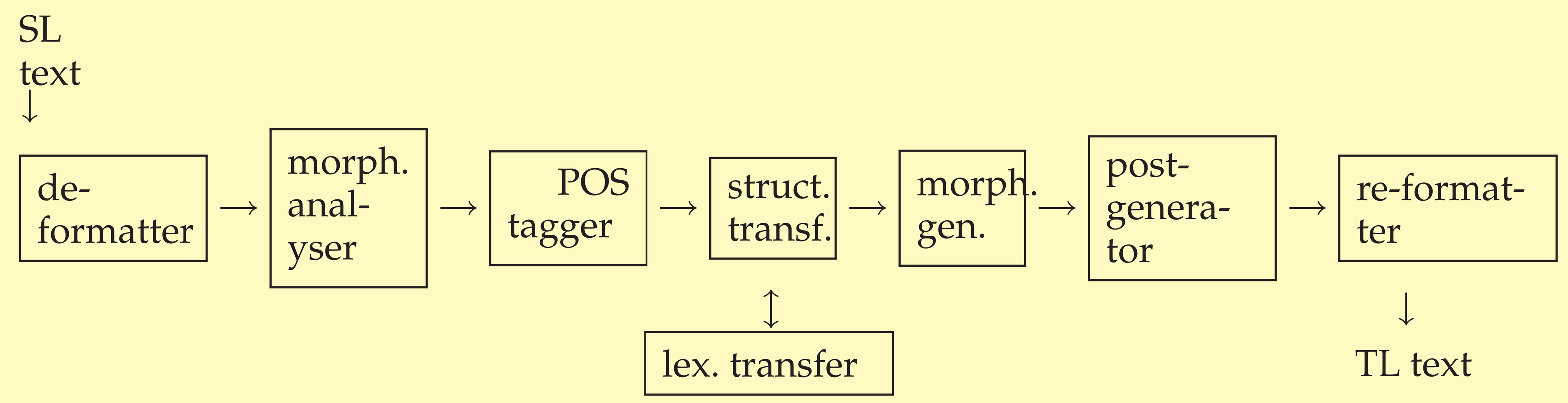
- morpho(phono)logical transducers (lexc and twolc)
- a North Sámi Constraint Grammar parser

`guolli AIGI "fish N" ;` has the accusative plural form *guliid*. Both consonant gradtion (*-ll-* vs. *-l-*) and. diphtong simplification (*uo* -> *u*) take place. The two-level compiler handles both.

```
Vx:0 <=> Vow _ Cns:+ i (...)  X5:  ;
  where Vx in (e o a) ;
```

In the diphtong simplification rule, *X5* marks that the second vowel (*e o a*) in a diphthong has to be simplified if the suffix contains an *i*.
Morphological and syntactic disambiguation is handled by the North Sámi Constraint Grammar parser. By means of context rules both morphological and syntactic analyses are removed except for the last reading.

Rule-based machine translation



Aside from the main modules in the Apertium system (morphological analyser, HMM part-of-speech tagger and structural/lexical transfer) the N.-L. Sámi system also took advantage of:

- The N. Sámi constraint grammar – which made the HMM tagger largely redundant and including syntactic and semantic analysis for use in the structural transfer
- An bilingual transfer lexicon

Construction of the transfer lexicon
A transducer applied regular changes to turn N. Sámi lemmata into L. Sámi candidates (e.g. **š** -> **sj** in *beaivváš* ('sun') -> *bæjvásj*). Words recognised by the L. Sámi morphological analyser with the same POS as the input word were accepted, words not recognised were revised.

Structural differences between North and Lule Sámi

- **Case differences**: N. Sámi locative -> L. Sámi inessive or elative
- **Negation**: the N. Sámi negation verb can inflect for tense, L. Sámi expresses tense by means of the main verb
- **Word order**: L. Sámi allows for a number of SOV (subject object verb) constructions whereas N. Sámi prefers SVO (1)

(1) Anne ráhkada biepmu. (N. Sámi)
Anne biebmov dahká. (L. Sámi)
'Anne makes food.'

The SOV rule captures the pattern (subject, verb, object) and outputs them in the order subject–object–verb by reordering the chunks.

Statistical machine translation

We used the

- Moses decoder,
- the word aligner GIZA++
- the srilm language model

Language models:
For Lule Sámi we made both an unfactored and a factored (wrt POS) trigram language model on our Lule Sámi corpus, 278,000 words (120,000 words New Testament; 106,000 fact; 39,000 fiction).
Translation models:
The models were severely limited by the availability of parallel corpora (we had only The New Testament (9,200 parallel sentences), and curriculum texts (1700 parallel sentences).

Results

RBMT:
The translation of 16 Wikipedia test sentences were compared to a manual reference translation. Structural transfer is unproblematic with a few exceptions, the choice of lexical tags and the lexical choices are the bigger challenge.

Type of deviation	**Example (L. Sámi)**	**Explanation**
One-to-many relations	*dálla* vs. *dál* (both 'now')	two different forms of one word
Tag inconsistencies	*iesjráddijiddje* ('self-governed')	is analysed both as a deverbal form and a lexicalised adjective
POS assymetries	*gullujiddje* ('belonging')	is analysed as a derived verb form
CG disambiguation error	*liehket* (infinitive 'to be')	should be *li* (3rd person plural)
Lexical matters	*tjiehpe* vs. *smidá* (both 'clever')	can be used synonymously in certain contexts
Case	*bargojn* ('work' Ine+Pl/Com+Sg) vs. *bargoj* (Com+Pl/Gen+Sg)	the wrong case form is used
Word order	*manna l ulmmel* vs. *man ulmmen la* ('which is the purpose')	SVO vs. SOV

SMT:

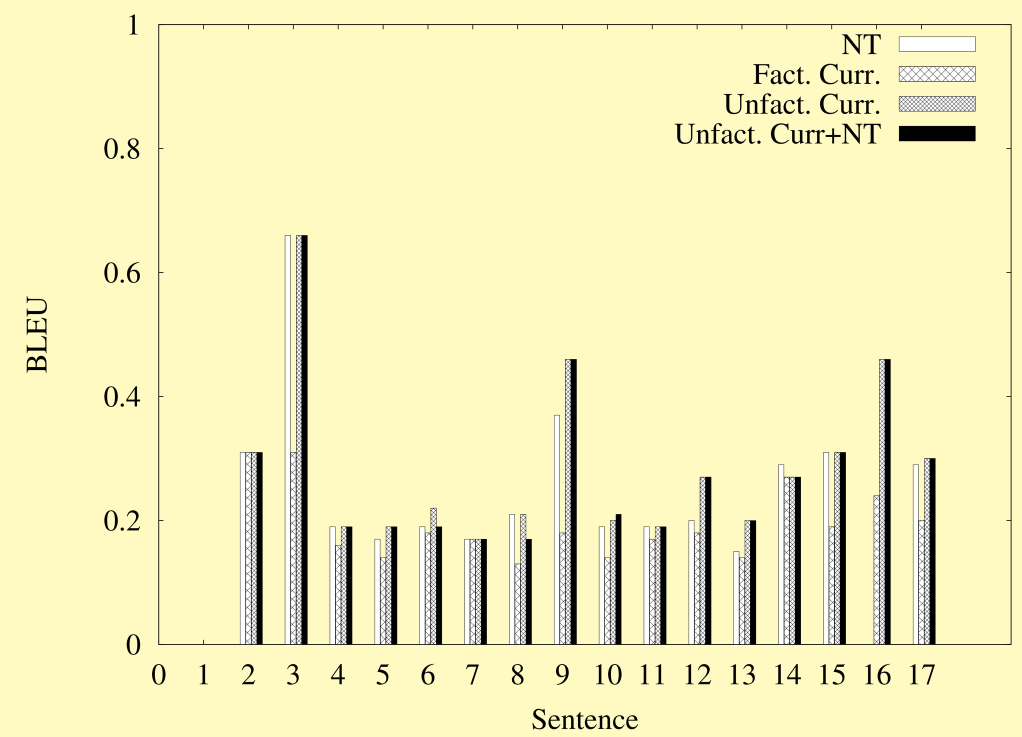


Figure 2 BLEU result for four models

The same 16 test sentences were translated by

- a factored model (curriculum corpus)
- an unfactored model (curriculum corpus)
- an unfactored model (NT corpus)
- an unfactored model (the whole corpus)

Somewhat unexpectingly, the unfactored model was better than the factored one (average BLEU 0.3 vs. 0.2).
Comparing the SMT and RBMT results is problematic, as the lexicon for the rule-based system was small, and the grammar rule set was restricted. The RBMT did well on known constructions (BLEU > 0.9), but badly on new text. The SMT did badly across the board, and much of its success was due to the similarities of the languages causing *free rides*.
The parallel corpus was too small for serious SMT work, but its size is representative for what might be found for low-density minority languages.
The token/type ratio changes from genre to genre, but the relative distance between languages remain the same. This indicates that also an SMT system based upon a larger corpus would fare less well for morphologically complex languages like the Sámi languages.

Conclusions

The rich morphology and especially the paucity of parallel corpora for Sámi make SMT less suited for MT between North and Lule Sámi, despite the close relationship between the two languages. Therefore RBMT is the best approach for this language pair.

Apertium copes well with the structural transfer from North to Lule Sámi. Improving the lexicon and the coverage of the structural transfer rules will be the next steps forward for our RMBT model.