Developing Prototypes for Machine Translation between Two Sámi Languages

Francis M. Tyers
Departament de Llenguatges
i Sistemes Informàtics,
Universitat d'Alacant
E-03071 Alacant (Spain)
ftyers@dlsi.ua.es

Linda Wiechetek
Giellatekno,
Universitetet i Tromsø,
Norway
linda.wiechetek@uit.no

Trond Trosterud
Giellatekno,
Universitetet i Tromsø,
Norway
trond.trosterud@uit.no

Abstract

This paper describes the development of two prototype systems for machine translation between North Sámi and Lule Sámi. Experiments were conducted in rule-based machine translation (RBMT), using the Apertium platform, and statistical machine translation (SMT) using the Moses-decoder. The experiments show that both approaches have their advantages and disadvantages, and that they can both make use of pre-existing linguistic resources.

1 Introduction

In this paper we describe the development of two prototype machine translation systems between two Sámi languages, North Sámi (sme) and Lule Sámi (smj), one rule-based (Apertium), and one statistical (Moses). There are other systems which have been developed with marginalised languages in mind (e.g. (Lavie, 2008)), but, as of writing, these were not available under an open-source licence and thus could not be applied to the task at hand. The content will be split into several sections. The first section will give a general overview of the languages in question, and sketch a typology of MT scenarios for minority languages. The next sections will describe the two machine translation strategies in some detail and will outline how the existing language technology was able to be re-used and integrated. We will follow this by a short evaluation and then some discussion and future work.

1.1 The languages

Both North Sámi and Lule Sámi belong to the Finno-Ugric language family and are spoken in the north of Norway and Sweden, North Sámi also in Finland. North Sámi has between 15,000 and 25,000 speakers, while Lule Sámi has less than 2,000 speakers.

The Sámi proto-language was originally an agglutinative language, but North and Lule Sámi have developed features known from inflective languages (case/number combinations are often expressed by one suffix only, certain morphological distinctions are expressed by means of consonant gradation (i.e. a non-segmental process) only, etc.).

The main objective with the development of the prototype rule-based system was to evaluate how well existing resources could be re-used, and if the shallow-transfer approach was suited to languages with more agglutinative typologies.

1.2 A typology of MT systems for minority languages

Minority language speakers typically differ from the majority in being bilingual, the minority speaks the language of the majority, but not vice-versa. This has some implications for the requirements society will put to machine translation systems.

A majority to minority language system must be of high quality, so high that post-editing the output is faster than translating from scratch. The goal is to produce well-formed text, not to understand the content, since the minority language users will prefer the original to a bad translation. A minority to majority language system, on the other hand, will be useful even as a gist system, answering vital questions such as "what are they writing about me in the minority language newspaper?". The systems presented here are minority to minority language systems. North and Lule Sámi are mutually intelligible, and also in this context a gist sys-

tem will not be that interesting. The importance of the system lies in its ability to produce text. Here, North Sámi is the larger language, possessing close to a full curriculum of school textbooks. A high-quality MT system would help produce the same for Lule Sámi, and moreover from the closely related North Sámi than from Norwegian. The same situation may found for many language communities.

2 Rule-based machine translation

2.1 Apertium

Apertium is an open-source platform for creating rule-based machine translation systems. It was initially designed for closely-related languages, but has also been adapted to work better for less-related languages. The engine largely follows a shallow-transfer approach. Finite-state transducers (Garrido-Alenda and Forcada, 2002) and (Roche and Schabes, 1997) are used for lexical processing, first-order hidden Markov models (HMM) are used for part-of-speech tagging, and multi-stage finite-state based chunking for structural transfer (Forcada, 2006). The original shallow-transfer Apertium system consists of a de-formatter, a morphological analyser, the categorial disambiguator, the structural and lexical transfer module, the morphological generator, the post generator and the reformatter.

2.1.1 Analysis and generation

For the analysis and generation, we used existing finite-state transducers for the two languages.[1]

lttoolbox has been widely used to model romance language morphology, and although it has been used to model the morphology of other languages with complex morphology (e.g. Basque), it is not ideal for these languages. *lttoolbox* is lacking features for dealing with stem-internal variation, diphthong simplification, and compounding.

North Sámi word forms involve both consonant gradation, diphthong simplification and compounding. The North Sámi noun *guolli* ('fish') alternates between *-ll-* (strong stage) and *-l-* (weak stage).

Additionally, one has to deal with diphthong simplification, the diphthong *uo* changes into a monophtong *u* in e.g. accusative plural *guliid*.

In the Apertium lexicon, *guolli* is represented as in figure 1. *gu* represents the stem, the item be-

```
<pardef n="gu/olli__N">
 <e>
   <p>
     <l>liid</l>
     <r>olli<s n="N"/><s n="Pl"/>
                     <s n="Acc"/></r>
   </p>
 </e>
 ...
</pardef>
```

Figure 1: Section of inflectional paradigm for `gu/olli__N`

tween `<l></l>` *liid* the generated ending and the items between `<r></r>` the analysis including the lemma and the morphological tags.

The Divvun and Giellatekno Sámi language technology projects[2] use finite-state transducers for the morphological analyser and closed-source finite-state tools from Xerox (Beesley and Karttunen, 2003). They tools handle two-level morphology model with twolc (two-level compiler) for morphophonological analysis together with lexical tools in a single transducer, consonant gradation, diphthong simplification and compounding are handled by two-level rules. Consonant gradation and diphthong simplification of the noun *guolli* are handled in the following way. *guolli* is listed in the root lexicon with lemma and continuation lexicon *AIGI* and redirected to the sublexicon *AIGI*.

```
guolli AIGI "fish N" ;

LEXICON AIGI !Bisyll. V-Nouns.
+N+Sg+Acc:%>X4 K ;
+N:%>X5 GODII- ; ! weak gr dipth simpl
...
```

From there it is redirected to a further sublexicon *GODII-* which redirects it to the sublexicon *GODII-*, which provides the plural accusative analysis.

```
LEXICON GODII-
 +Pl+Acc:jd9 K ;
```

At the same time, a two-level rule handles diphthong simplification when encountering the diacritical mark *X5* by removing the second vowel (*e o a*) in a diphthong (*ie uo ea*) if the suffix contains an *i*.

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

Consonant gradation is handled in another rule where a consonant (*f l m n r s . . .*) is removed between a vowel, an identical consonant, another vowel, and a weak grade triggering diacritial mark (the rule is slightly simplified, noted by . . .).

```
Cx:0 <=> Vow: _ Cy  Vow (...) WeG: ;
where Cx in (f l m n r s ...)
      Cy in (f l m n r s ...)
```

A general difficulty for generation and analysis are inconsistent tagsets in SL and TL. While verbs are specified with regard to transitivity (*V TV, V IV*) for North Sámi, they were not specified in the Lule Sámi dictionary (only *V*). Another matter of choice and convenience is the degree of lexicalisation as in the case of derived verbs. The North Sámi verbform *gohčoduvvo* ('he/she is called') either goes back to the form *gohččut* ('order') or to *gohčodit* ('call, name'), which is derived from *gohččut* but to some extent lexicalised.

```
gohččut+V+TV+Der1+Der/d+V
      +Der2+Der/PassL+V+Ind+Prs+Sg3
gohčodit+V+TV+Der2+Der/PassL+V+Ind+Prs+Sg3

gåhtjudit+V+TV+Der1+Der/Pass+V+Ind+Prs+Sg3
```

In Lule Sámi, *gåhtjuduvvá* only gets the analysis with the lexicalised verb *gåhtjudit* as a lemma. The parallel derived form to North Sámi is not provided in the analysis. For the construction of the bilingual `sme-smj` dictionary, that means that *gohčoduvvo* is only matched with *gåhtjuduvvá* if *gohčoduvvo* is analysed with *gohčodit* as its lemma. In the bilingual dictionary, both pairs *gohččut - gåhttjot* and *gohčodit - gåhtjudit* exist. But *gåhtjuduvvá* cannot be generated from *gåhttjot*.

```
<e><p><l>gohččut<s n="V"/></l>
     <r>gåhttjot<s n="V"/></r></p></e>
<e><p><l>gohčodit<s n="V"/></l>
     <r>gåhtjudit<s n="V"/></r></p></e>
```

In the previous case, tag assymetry is due to annotation-choices. In other cases tag inconsistencies are linguistically motivated as in the case of the negation verb *ii/ij* ('not (do)'), which is specified with regard to tense in Lule Sámi, but not in North Sámi. This is due to the fact, that Lule Sámi has different present tense and past tense forms of the verb. North Sámi, on the other hand only has one form to express both present and past tense. The tense distinction is made by means of the main verb following the negation verb as in *ii boađe* ('he/she does not come') and *ii boahtán* ('he/she did not come').

```
ii      ii+V+IV+Neg+Ind+Sg3
```

```
ij      ij+V+Neg+Prs+Sg3
ittjij   ij+V+Neg+Prt+Sg3
```

Both for generation and analysis that means that one has to find a possibility to account for the 'missing' tag in North Sámi. 'Missing' means here the lack of tag specification for the *temps* (tempus) variable in the transfer files.

A number of multiword expressions differ from each other in SL and TL. While in North Sámi *gii beare* ('whoever') has inner inflection, the Lule Sámi *vajku guhti* does not. The initial pronoun *gii* corresponds to the second component *guhti* in Lule Sámi.

The last type of generation modification happens in a separate step. Orthographic variants and contractions are handled by the postgenerator. The Lule Sámi copula *liehket ('to be')* has three forms for the tag combination liehket+V+Ind+Prs+Sg3, *le, la, l*. While *le* and *la* are interchangable variants, *l* is a shortened form of *la* after wordforms that end in a vowel. The postgeneration lexicon specifies this change and outputs the correct form.

2.1.2 Disambiguation (Constraint Grammar)

Disambiguation of morphological and shallow syntactic tags is handled by the North Sámi parser. The parser uses Constraint Grammar, a formalism based on Karlsson (1990) and Karlsson (1995) and further developed by Tapanainen (1996) and Bick (2000).

The approach is bottom-up, which means that all input (ideally) receives one or more analyses. Those analyses are then one by one removed except for the last reading, which is never removed. The parser uses the output of the morphological transducer as an input and adds shallow syntactic tags. Syntax tags do not only function as the basis of a dependency tree structure representation, but also disambiguate morphology, e.g. homonymous genitive and accusative forms are distinguished on the level of syntax (genitive premodifier @→*N* vs. accusative object @←*OBJ*). The readings are then disambiguated by means of context rules.

The disambiguation file itself consists of different sections:

- **Sets**: lexical, POS, morphological features, syntactic, semantic lists one wants to abstract over

- **Syntactic annotation rules**: operators MAP and ADD annotate syntactic tags such as @←*OBJ*

- **Disambiguation rules**: operators SELECT and REMOVE either pick or discard a reading

In the Apertium engine, the Constraint Grammar module is added as a pre-disambiguator after the morphological analyser and before the statistical POS tagger. Apertium uses the r21668 version[3] of the parser, which is based on vislcg3.

A syntactic (or even semantic) analysis of the SL is also useful in MT, and the structural transfer in the `sme-smj` Apertium engine profits from syntactic information. By mapping the habitive tag *@HAB* onto locative nouns with habitive syntax/semantics, one can directly translate locative into inessive and a structural transfer rule in one of the MT modules becomes redundant. In the prototype system, the accuracy of the CG disambiguator has made the HMM-based tagger almost redundant.

It would appear that the rule-based Constraint Grammer parser is able to give a better performance than an HMM based tagger.[4]

Trigrams are not suitable for expressing syntactic structure.[5] CG on the other hand successfully expresses syntactic structure as a product of contextual disambiguation. (Bick, 2000, p.137)

2.1.3 Lexical transfer

Lexical transfer is handled in the bilingual dictionary, where entries have the form

```
<e><p><l>beaivi<s n="N"/></l>
    <r>biejvve<s n="N"/></r></p></e>
```

The North Sámi lemma with its POS specification comes first embedded in `<l></l>`, followed by the Lule Sámi lemma and its corresponding POS specification embedded in `<r></r>` In the case of a one-to-many relation between SL and TL, i.e. if several TL items exist for one SL item, the default translation is picked by means of the restriction `<e r="RL">`.

```
<e r="RL"><p><l>dàl<s n="Adv"/></l>
       <r>dàlla<s n="Adv"/></r></p></e>
```

[3]2008-10-29 23:20:45 +0100

[4]Samuelsson and Voutilainen note in their comparison of a a linguistic and stochastic tagger that "at ambiguity levels common to both systems, the error rate of the statistical tagger was 8.6 to 28 times higher than that of EngCG-2." (Samuelsson and Voutilainen, 1997, p.251)

[5]According to Bick (2000), the syntactic structure problem is "unique" to probabilistic HMM grammars and resides in the "Markov assumption" that $p(tn|t1...tn-1) = p(tn|tn-1)$ (for bigrams), or $= p(tn|tn-1 tn-2)$ (for trigrams).

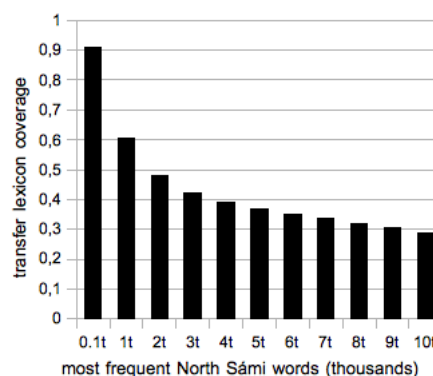**Figure 2:** Coverage of the bilingual `sme-smj` dictionary

```
<e>        <p><l>dàl<s n="Adv"/></l>
           <r>dàl<s n="Adv"/></r></p></e>
```

A lexical selection module as described in the Apertium documentation (Forcada, 2008) is not employed by the system. Lexical transfer is considered to be regular instead of context-dependent. How close that is to the real situation is still to be decided.

The transfer lexicon was constructed in the following way: The orthographical differences between North and Lule Sámi are mostly regular. We thus made a finite state transducer which turned North Sámi lemmata into Lule Sámi candidates. The candidates were run through our Lule Sámi morphological transducer. Words recognised with the same POS as the input word were accepted, whereas words not recognised were manually revised. Semantic pairs which were non-cognates were manually added. Figure 4 shows the coverage of our transfer lexicon, for the n-thousand most common North Sámi lemmata.

2.1.4 Syntactic transfer

There are a number of structural differences between North and Lule Sámi that require structural transfer rules.

- The North Sámi locative case expressing place and source corresponds to either Lule Sámi inessive (place) and elative (source) depending on the context.

- In simple object constructions, the unmarked word order in Lule Sámi tends to be SOV, while it is SVO in North Sámi.

- In negation construction as discussed above, the Lule Sámi negation verb can inflect for

tense, while in North Sámi tense is expressed by means of the mainverb negation form

As the default translation of North Sámi locative (1) the Apertium system chooses Lule Sámi elative as in (2).

(1)   son čokkii dávviriid ja dávttiid boares hávddi-in.
son čokkii dávvriid ja dávttiid boares hávddi-LOC.PL.
'(s)he collected things and bones old graves.from.'

(2)   sån tjåkkij dávverijt ja dávtijt boares hávdi-js.
sån tjåkkij dávverijt ja dávtijt boares hávdi-ELA.PL.
'(s)he collected things and bones old graves.from.'

The default elative becomes inessive

- in habitive constructions,

- in place adverbials of stative verbs,

- before certain adverbs such as *gitta*.

In (3) a structural rule chooses inessive as a translation for locative when encountering the habitive tag *@HAB* distributed by a CG-rule, a verb from the verbs_stative list such as *ássat* ('live'), and an adverb from the ine_adv list such as *gitta* ('dependent on').

(3)   Sámit dahjege sápmelaččat ásset Ruošša-s, Suoma-s ja Norgga-s.
Sámit dahjege sápmelaččat ásset Ruošša-LOC.SG, Suoma-LOC.SG ja Norgga-LOC.SG.
'Sámi or also 'sápmelaččat' live in Russia, Finland and Norway.'

(4)   Sáme jali sábmelattja årru Ruossja-n, Suoma-n ja Vuona-n.
Sáme jali sábmelattja årru Ruossja-INE.SG, Suoma-INE.SG ja Vuona-INE.SG.
'Sámi or also 'sábmelattja' live in Russia, Finland and Norway.'

North and Lule Sámi differ with respect to word order. Especially in written texts, Lule Sámi allows for a number of unmarked SOV (6) construction whereas North Sámi prefers SVO (5).

(5)   Anne ráhkada biepmu.
Anne makes   food.

(6)   Anne biebmov dahká.
Anne food   makes.

Word order is treated in the second transfer module. The SOV rule in figure 3 captures the pattern (subject, verb, object) and outputs them in the order subject–object–verb by reordering

```
<rule>
  <pattern>
    <pattern-item n="SN_Subj"/>
    <pattern-item n="FMainV"/>
    <pattern-item n="SN_Obj"/>
  </pattern>
  <action>
    <out>
      <chunk>
        <clip pos="1" part="whole"/>
      </chunk>
      <b pos="1"/>
      <chunk>
        <clip pos="3" part="whole"/>
      </chunk>
      <b pos="2"/>
      <chunk>
        <clip pos="2" part="whole"/>
      </chunk>
    <out>
  </action>
</rule>
```

Figure 3: Transfer rule to convert SVO → SOV

the chunks indicated by *pos="1", pos="2"* and *pos="3"* into 1–3–2.

The structural rules work successfully in transferring North Sámi to Lule Sámi structures. As the structural differences are minimal, the construction of rules is not very time-consuming. Rather the identification of structural differences is a new task as contrastive North-Lule Sámi grammar has been a rather neglected area within syntactic research.

3 Statistical machine translation

For the statistically based machine translation we used the Moses decoder, the word aligner GIZA++, and the srilm language model.[6]

3.1 Corpora

Minority languages may roughly be divided into three groups: The ones with a (limited) role in public administration or similar domains, the ones with a standardised written language and some text (more often than not the Bible comprises the bulk of the available corpus), and the ones with neither of these. Of our languages, North Sámi falls in the first group and Lule Sámi in the second. This means that the parallel resources available are extremely limited, they consist of the New Testament (approx. 150,000 words each), and a small corpus of school curriculum texts (appr. 15,000 words

---

[6]Available from the urls `http://www.statmt.org/moses/`, `http://www.fjoch.com/GIZA++.html`, `http://www.speech.sri.com/projects/srilm/` respectively

each, describing the content of the curriculum for the Sámi schools in Norway). The two NT versions have been translated in different countries (Norway/Finland and Sweden, respectively), with different Bible versions as source texts, and they differ from each other more than an ordinary parallel corpus would have done. The curriculum texts are probably translations of the same original – in any case the sentences are better matches of each other.

3.2 Training process

For the statistical machine translation, we build both factored and unfactored models. For Lule Sámi (the target language) we made both an unfactored and a factored trigram language model on our Lule Sámi corpus, 278,000 words. Half of the corpus (120,000 words) consists of New Testament (NT) texts, 106,000 belongs to the fact category, and 39,000 words is fiction. The factored model contained POS information, obtained from our Lule Sámi CG parser.

We then built various translation models. The models were severely limited by the availability of parallel corpora. We had one corpus consisting of the New Testament (9,200 parallel sentences), and one containing curriculum texts (1700 parallel sentences).

4 Evaluation

4.1 Qualitative evaluation

For the development of the Apertium system 16 test sentences from Wikipedia were used as regression tests.[7] Their target translations are based on a manual translation. Out of the 16 test sentences, 12 are successfully matched with the target translation at present. Remaining problems are not of a structural kind, but are dependent on one-to-many relations in the bilingual dictionary, tag inconsistencies between the `sme` and `smj` dictionaries, POS assymmetries and disambiguation errors from the Constraint Grammar disambiguator.

For evaluation purposes another independent manual translation is used. The Apertium translation deviates mostly with regard to lexical matters. Other lemmata were chosen. If they are synonyms or more idiomatic than the other ones remains to be studied. With regard to structural deviations, there was one deviating choice of case and one of

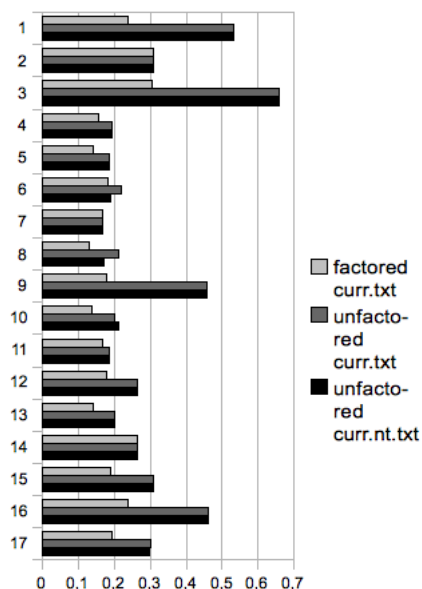[7]http://wiki.apertium.org/wiki/Northern_Sami_and_Lule_Sami/Regression_tests

Figure 4: BLEU result for three models

word order. The word order deviation might hint at a rather optional SOV order.

The evaluation shows that while structural transfer seems to be mostly unproblematic, the choice of lexical tags and the lexical choices are the bigger challenge. Tags should be consistently chosen for both SL and TL whenever the deviation is not linguistically motivated. In the case of lexical choice, one needs to have a closer look at the bilingual lexicon. Are the deviations interchangable translations or is one of them the more idiomatic one?

4.2 Quantitive evaluation

For evaluation, we used the same 16 North Sámi Wikipedia test sentences (manually translated into Lule Sámi). For the SMT system, they were tested by three different translation models, a factored and an unfactored model based upon the curriculum corpus, and an unfactored model based upon the NT corpus (due to technical difficulties we were not able to make a factored model of the NT corpus).

The results are somewhat unexpected. Of the two versions of the curriculum translation model, the unfactored one is better than the factored one, with an average BLEU score of 0.3 as against 0.2. Comparing the two unfactored models, the larger one, containing NT and curriculum texts, performs similarly to the curriculum model for most sentences, but worse in some cases, resulting in a slightly worse overall score.

Type of deviation	Example
one-to-many relations	*dálla* vs. *dál* (both 'now')
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) should be *li* (3rd person plural)

Table 1: Remaining transfer problems

Type of deviation	Example
lexical matters	*tjiehpe* vs. *smidá*, *moattegielak* vs. *ålogielak*, *sáhttá* vs. *máhttá*
case	*bargojn* vs. *bargoj*
word order	*manna l ulmmel* SVO vs. *man ulmmen la* SOV ('which is the purpose')

Table 2: Selection of divergences between North Sámi and Lule Sámi

Comparing the SMT and RBMT results is harder, as the lexicon for the rule-based system was small, and the grammar rule set was restricted. Thus, the RBMT did very well on known constructions (BLEU around 0.9 and better), 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 (unknown words were passed through and now and then were correct).

With such a small training set, the result cannot be but bad. From earlier cross-linguistic research, a morphology-rich language such as Finnish comes out with clearly worse results than the more analytic German and French. Comparing BLEU score from (Banchs, 2005) with the token/-type ratio of Banchs' training set gives the picture in table 3.

	French	German	Finnish
Token/type	189	74	29
BLEU	0,302	0,245	0,203

Table 3: Token/type ratio and BLEU for 4 source languages in a Europarl MT study

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 than good for a morphologically complex language like Sámi.

5 Discussion

The corpora for Sámi are not good enough for SMT systems to be able to replicate the good RBMT results for North Sámi to Lule Sámi but much can be done both with tuning and corpus gathering. The corpora are probably good enough to build a gist system for North Sámi to Norwegian.

Apertium copes well with the structural transfer, but tag inconsistencies and many-to-many relations in the lexicon cause deviations between manual and automatic translations. A good lexicon and a consistent tagset are the basis for successful RBMT.

For morphologically complex languages, the lt-toolbox format for designing transducers might not be ideal, and one might consider other morphological transducers such as lexc and twolc.

Future plans in RBMT aim at making a full-coverage system out of the Apertium prototype. Word alignment can help constructing a more complete and better bilingual dictionary, and statistical methods could be used to choose the most idiomatic wordform in the case of one/many-to-many relations. Alternatively, a statistically-based lexical selection module as proposed in (Forcada, 2008) may be included. For optimisation of structural transfer, the Constraint Grammar could be enhanced by semantic roles that disambiguate between an inessive locative (PLACE) and an elative locative (SOURCE).

The available parallel corpora where Lule Sámi is one of the languages will not be large enough for SMT in the foreseeable future.

Returning to the typology of MT systems for minority languages, we would like to explore the possibility of using SMT to create a gisting system for North Sámi to Norwegian. A corpus of 1,000 sen-

tences has already been tested. For this language pair, the linguistic distance is longer, but the empirical base far better (the present corpus collection contains appr. 120,000 sentences of parallel (but non-aligned) text). Although not much can be expected from a North Sámi–Lule Sámi SMT system, the development of a North Sámi–Norwegian system should be possible.

Acknowledgements

References

Banchs, Rafael E. and Crego, Josep M. and de Gispert, Adrià and Lambert, Patrik and Mariño, José B. 2005. Statistical Machine Translation of Europarl Data by using Bilingual N-grams, *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pp. 133–136

Beesley, K. R. and L. Karttunen. 2003. *Finite State Morphology* Vol. 1 CSLI Publications, Stanford. http://www.fsmbook.com/.

Bick, E. 2000. *The Parsing System 'Palavras': Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, Aarhus.

Forcada, M. L. 2006. Open-source machine translation: an opportunity for minor languages. *Strategies for developing machine translation for minority languages*. 5th SALTMIL workshop on Minority Languages. pp. 1–7

Forcada, M. L. and B. Ivanov Bonev and S. Ortiz Rojas and J. A. Pérez Ortiz and G. Ramírez Sánchez and F. Sánchez Martínez and C. Armentano-Oller and M. A. Montava and F. M. Tyers. 2008. *Documentation of the Open-Source Shallow-Transfer Machine Translation Platform Apertium*. http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf

Garrido-Alenda A. and M. L. Forcada. 2002. Comparing nondeterministic and quasideterministic finite-state transducers built from morphological dictionaries. *Procesamiento del Lenguaje Natural*. No. 29 pp. 73–80

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (eds.). 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. *Natural Language Processing* No. 4 *Mouton de Gruyter*, Berlin and New York.

Karlsson, F. 1990. Constraint Grammar As A Framework For Parsing Running Text. *Proceedings of COLING* Vol. 3 pp. 168–173

Lavie, A. 2008. Stat-XFER: A General Search-based Syntax-driven Framework for Machine Translation *Proceedings of CICLing 2008*, pp. 362–375

Roche, E. and Y. Schabes (eds.). 1997. Finite-State Language Processing. *MIT Press*, Cambridge, Massachusetts.

Samuelsson, C. and A. Voutilainen. 1997. Comparing a Linguistic and a Stochastic Tagger. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* pp. 246–253

Tapanainen, P. 1996. *The Constraint Grammar Parser CG-2*. University of Helsinki Publications Vol. 27 pp. 246–253