Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

A North Saami to South Saami Machine Translation Prototype

Lene Antonsen Trond Trosterud
Francis M. Tyers

Department of Language and Culture
The Arctic University of Norway
{lene.antonsen, trond.trosterud, francis.tyers}@uit.no

Abstract

The paper describes a rule-based machine translation (MT) system from North to South Saami. The system is designed for a workflow where North Saami functions as pivot language in translation from Norwegian or Swedish. We envisage manual translation from Norwegian or Swedish to North Saami, and thereafter MT to South Saami. The system was aimed at a single domain, that of texts for use in school administration. We evaluated the system in terms of the quality of translations for postediting. Two out of three of the Norwegian to South Saami professional translators found the output of the system to be useful. The evaluation shows that it is possible to make a functioning rule-based system with a small transfer lexicon and a small number of rules and achieve results that are useful for a restricted domain, even if there are substantial differences between the languages.

1 Introduction

The paper presents an MT system from North to South Saami. It is intended to work in a setting with manual translation from Norwegian or Swedish into North Saami, acting as a pivot language with postediting to the other, smaller, Saami languages. The Saami languages are closely related, and therefore lend themselves better to MT than a system translating from Norwegian or Swedish directly, but on the other hand, the classical problems of translating via a pivot language apply in this case as well. On the other hand, by using North Saami as a pivot language, one can manage with one MT-system instead of two.

After looking at previous work and at the languages themselves, we present the actual MT system and its evaluation. We presented translated text to translators, alongside with the Norwegian original, without revealing the fact that the texts were actually translated from North Saami. Finally we present some conclusions.

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

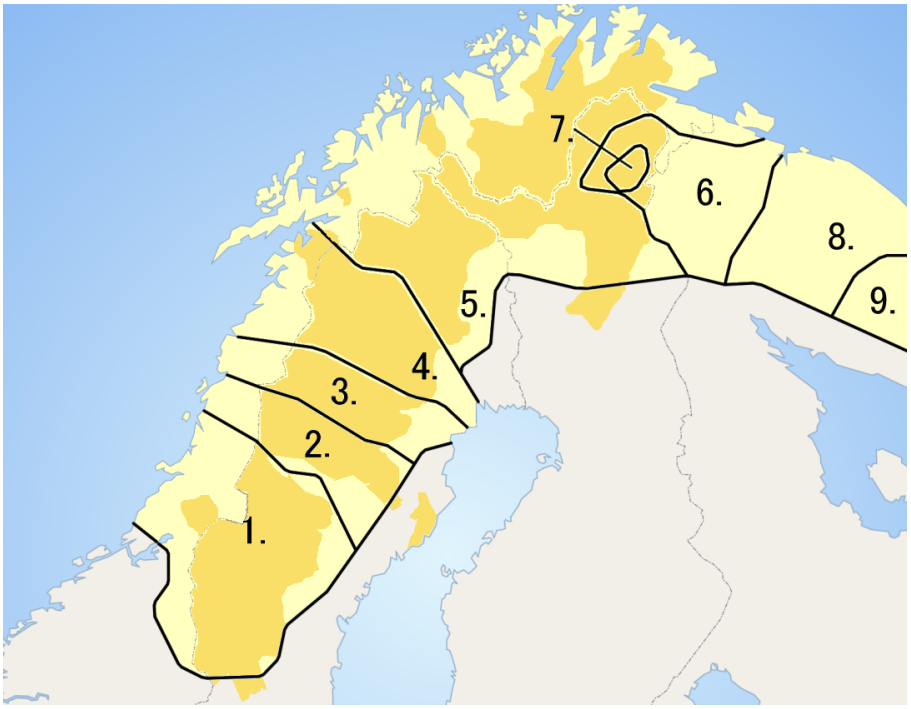

Figure 1: Map of the Saami language area. South Saami (1, Ume Saami (2, Pite Saami (3, Lule Saami (4, North Saami (5, Skolt Saami (6, Inari Saami (7, Kildin Saami (8, and Ter Saami (9. The administrative area for Saami languages is coloured dark yellow.

2 Background

2.1 Motivation

Six of the Saami languages are part of the regular school curriculum, to at least some extent, in the countries where they are spoken, three of them in Norway and Sweden. At Figure 1 these are numbered with 1 and 4–8. North Saami is by far the largest, its approximately 22,000 speakers constitute more than 85 % of the total amount of Saami speakers. South Saami has fewer than 500 speakers in Norway and Sweden.[1] The number of pupils having South Saami as a school subject in Norway in 2014/15 was 74, as compared to 1,943 for North Saami (Rasmussen, 2015. Despite being few, South Saami pupils also need school textbooks in the core subjects throughout all their years at school, school subjects must be planned, defined and evaluated, etc.

South Saami has made its entrance as an administrative language during the last decade. The first municipality within the South Saami area to adopt Saami as an official language was Snåsa in Norway, in 2008, today two municipalities in Norway and ten in Sweden within the South Saami area have Saami as one of their offical la nguages. Saami has status as an administrative language in seven municipalities in the North Saami area

[1] A general overview of the South Saami sociolinguistic situation is given in the introductory chapter of Todal (2007).

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

in Norway and in three municipalities in the North Saami area in Sweden.[2]

It goes without saying that with fewer than 500 South Saami speakers, text production becomes a bottleneck for both education and revitalisation. The motivation for the present work is thus to offer machine translation as a way of making text production more efficient, via manual translation from Norwegian and Swedish to North Saami, which then functions as a pivot language for MT into other Saami languages, in this case to South Saami. South Saami is also the Saami language with the largest distance to North Saami, thus any success in the present project may safely be assumed to be within reach for MT to the other Saami languages as well.

2.2 Previous research

Previous work on Saami machine translation includes Tyers et al. (2009 (on an early MT system North Saami to Lule Saami, comparing RBMT and SMT for the language pair in question, Wiechetek et al. (2010 (on lexical selection rules for the same language pair, and Trosterud and Unhammer (2012 (on an evaluation of a North Saami to Norwegian MT system. These articles do not discuss the role of pivot translations of related languages, and they do not deal with South Saami.

A study of pivot translation for under-resourced languages is Babych et al. (2007, where they show that combining a rule-based system for related languages (Russian and Ukrainian and another rule-based system for distant languages (Russian and English they are able to achieve better results than a rule-based system direct from Ukrainian to English. They put this down to the fact that the Russian to English system was much more developed, and that the task of translating between related languages is much easier, requiring less development time.

A similar approach to ours is Masselot et al. (2010 who use Spanish as a pivot to translate to Brazilian Portuguese, while showing translators only English original and Brazilian Portuguese MT output.

The novelty of our approach is thus applying this model to a more distant language pair, and the inclusion of a new language (South Saami into the machine-translation literature.

3 Languages

As Germanic languages, Norwegian and Swedish differ from the Saami languages in numerous respects. In this article we will especially refer to Norwegian because it is the language we picked for the evaluation.

Norwegian has definiteness as a morphological category, prepositions and verb / particle constructions, no case for nouns, and a relatively strict word order, with V2 in main clauses. However, Norwegian and the Saami languages also show Sprachbund phenomena. Most notably, they have the same tense system. Whereas the Saami languages also

[2] Cf. http://minoritet.prod3.imcms.net/1075 and https://www.regjeringen.no/no/tema/urfolk-og-minoriteter/samiske-sprak/samelovens-sprakregler-og-forvaltningsom/id633281/ Note that the regulations only mention "Saami", which Saami will be supported is dependent upon whatever Saami languages are traditionally used in the municipality in question.

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

possess a number of infinite verb construction, the Norwegian embedded clause pattern is in most cases a possible option for the Saami languages as well.

The Saami languages constitute the westernmost branch of the Uralic language family. They possess most of the classical Uralic characteristics: A rich verbal morphology with a rich repertoire of infinite forms, and a medium-size case system with both grammatical and adverbial functions, and no gender distinction. They also show extensive contact with their Germanic neighbours. Many grammatical structures are head-initial constructions, as compared to the classical Uralic head-final pattern.

Grammatical similarities between North and South Saami include the following: Both languages have three persons and numbers, they have a similar system for postpositions and verb derivation. Noun phrase syntax is similar, and the case systems are almost identical.

Taking a closer look, there are still differences. In both North and South Saami, negation is expressed with a negation verb which inflects for person and number, but in South Saami, this verb is inflected for tense as well. South Saami has OV word order and lacks a copula in predicative constructions, whereas North Saami uses VO in simple sentences, and requires a copula. The possession construction (*to have*) is different from the Norwegian one, normally using copula rather than *possessor*.NOM *Verb*. In North Saami, the construction is *possessor*.LOC *copula possessed*.NOM, whereas the possessor is in the genitive in South Saami.

Although the case systems are similar, North Saami has a locative case that covers the semantic field of both inessive and elative in South Saami. In MT then, the locative must be split into two cases. There is also some differences in case usage, plural objects are accusative in North Saami, but nominative for indefinite and accusative for definite objects in South Saami. Here is an example sentence, translated from Norwegian, first to North Saami and second to South Saami:

(1) Du må møte til den fagprøven lærebedriften melder deg opp til. (Norwegian)
Don galggat boahtit fágaiskosii, masa oahppofitnodat du almmuha. (North Saami)
You should.2SG come apprenticeship-exam.ILL , which.ILL apprenticeship-company you.ACC register.
Datne tjoerh dan faagenpryövenassese båetedh, mïsse learoesïelte datnem bæjjohte. (South Saami)
You should.2SG that.GEN apprenticeship-exam.ILL come, which.ILL apprenticeship-company you.ACC register.
'You have to take the apprenticeship exam that the company has registered you to do.'

For a MT system the orthographic differences imply that apart from person names and acronyms there will be no free rides during the translation process. The core vocabulary is distinct, even recent loans from the same donor languages are different, and the vocabulary coverage for a working system must thus be very good.

An overview over the differences between the two languages can be found in Sammallahti (1998). North and South Saami are not mutually intelligible, due both to linguistic distance and to radically different orthographic principles for their literary languages. An analogy would be the difference between English and Frisian. Both North Saami and South Saami are spoken in Norway and Sweden, but in addition North Saami is spoken

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

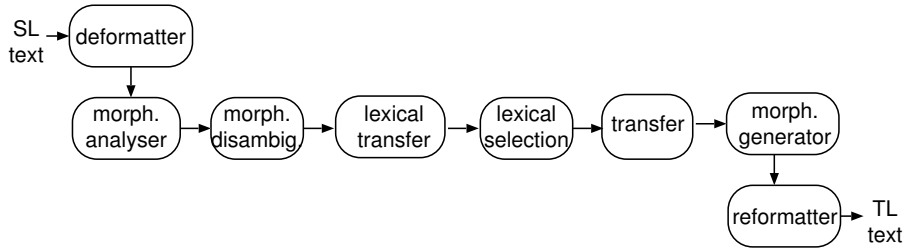

Figure 2: Overview of the translation pipeline.

in Finland, where also the majority language is a Uralic language. In some ways South Saami is more conservative than North Saami (e.g. SOV vs. SVO), but in other ways the contact with Finnish has slowed the rate of adoption of other features from Germanic languages into North Saami.

4 Implementation

The translator was implemented using the Apertium platform (Forcada et al., 2011). Apertium provides a highly modular set of tools for building rule-based MT systems. Apertium language pairs are set up as Unix pipelines, where the typical pipeline consists of:

- deformatting (encapsulating formatting/markup from the engine),
- source-language (SL) morphological analysis with a finite-state transducer (FST),
- disambiguation using a Hidden Markov Model (HMM) and/or Constraint Grammar (CG),
- lexical transfer (word-translation on the disambiguated source),
- lexical selection (choosing the appropriate word out of a set of possible translations),
- one or more levels of finite-state based structural transfer (reordering, and changes to morphological features),
- target-language (TL) generation with an FST
- reformatting (unencapsulating format information)

See Figure 2 for an overview of the modules used in this particular language pair, and Table 1 for an example of the output of the stages.

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

4.1 Analysis

Morphological analysis is done on the input using the Helsinki Finite-State Toolkit (Lindén et al., 2011). For each surface form, a finite-state transducer returns a set of the possible analyses, where an analysis is a combination of lemma, and a sequence of tags which describe the morphological structure of the surface form.

A Constraint Grammar-based disambiguator then selects the most appropriate analysis for each surface form according to the context, and assigns to each analysis a syntactic tag denoting its syntactic function (subject, object, main verb, …). The North Saami analysis pipeline has an F-score of 0.93 for morphological disambiguation and assignment for grammatical functions (Antonsen et al., 2010).

4.2 Transfer

4.2.1 Lexical transfer

There is no North Saami-South Saami dictionary, so we made one by combining the Norwegian words in a general North Saami-Norwegian dictionary and a general Norwegian-South Saami dictionary. The word pairs were manually edited and the work revealed many incorrect pairs because of many Norwegian words with more than one meaning. The result of this work was 3,568 general word pairs and 61 proper nouns. We also added a general name lexicon (names like *Maria, Johannesen, London*, with identical entries for both languages, the dictionary consists of approximately 10,000 such name pairs. Even though the names are identical, their morphology differs, so this addition was important for the general coverage of the system. A small number (80) of central name pairs were added with different forms in the two languages (*Oslo, Oslove; Stockholm, Stuehkie; Norga, Nöörje; Romsa, Tromsö; Ruošša, Russlaante*, all pairs listed as North, South), the number of such pairs should be enlarged in order to obtain better coverage.

The existing Norwegian-South Saami lexical resources are limited, with little terminology from modern society. We chose school administration texts as domain and by comparing Norwegian texts with the translations to North Saami and South Saami from a parallel corpus, we added 873 special domain word pairs to the bilingual dictionary. Note that although we have a parallel North Saami - South Saami corpus it contains only around 3,000 sentences — a fraction of the size necessary to train a broad-coverage statistical MT system.

Large syntactic differences also have consequences for the bilingual dictionary, because many North Saami verbs have to be translated into South Saami as `object` + `verb` and `adverb` + `verb` strings. Examples include Norwegian ‘å gjøre synlig’, (“to make visible”) which is translated into North Saami as one derived verb *almmustahttit*, but into South Saami as adverb+verb *våajnoes darjodh*. Norwegian ‘å presentere’ (“to present”) one can in North Saami translate with a verb with an adverbial suffix *ovdanbuktit* (lit. “forward-deliver”), but in South Saami it is not usual to do this, and the same meaning is expressed as two words, adverb + verb: *åvtese buektedh*. This is solved by giving the South Saami translation as a multiword expression (MWE) to the morphological analyser.

Some MWEs are harder to translate. E.g. the counterpart for the Norwegian word ‘framtid’ (“future”) is in North Saami *boahtteáigi*, but in South Saami an MWE with internal agreement: *båetije aejkie*.NOM, *båetijen aejkien*.GEN and so on. The forms for

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

Input	Don galggat boahtit fágaiskosii.
1.	^Don/don<prn><pers><sg2><nom>$ ^galggat/galgat<v><iv><pri><sg2>$ ^boahtit/boahtit<v><iv><pri><pl1>/boahtit<v><iv><inf>$ ^fágaiskosii/fága<n><sgnomcmp><cmp>+iskkus<n><sg><ill>$^./.<clb>$
2.	^Don/Don<prn><pers><sg2><nom><@SUBJ→>$ ^galggat/galgat<v><iv><pri><sg2><@+FAUXV>$ ^boahtit/boahtit<v><iv><inf><@-FMAINV>$ ^fágaiskosii/fága<n><sgnomcmp><cmp>+iskkus<n><sg><ill><@←ADVL>$^./.<clb>$
3.	^Don<prn><pers><sg2><nom><@SUBJ→>/Datne<prn><pers><sg2><nom><@SUBJ→>$ ^galgat<v><iv><pri><sg2><@+FAUXV>/edtjedh<v><pri><sg2><@+FAUXV>$ ^boahtit<v><iv><inf><@-FMAINV>/båetedh<v><iv><inf><@-FMAINV>$ ^fága<n><sggencmp><cmp>/faage<n><sggencmp><cmp>$ ^iskkus<n><sg><ill><@←ADVL>/pryövenasse<n><sg><ill><@←ADVL>$ ^.<clb>/.<clb>$
4.	^Don<prn><pers><sg2><nom><@SUBJ→>/Datne<prn><pers><sg2><nom><@SUBJ→>$ ^galgat<v><iv><pri><sg2><@+FAUXV>/edtjedh<v><pri><sg2><@+FAUXV>$ ^boahtit<v><iv><inf><@-FMAINV>/båetedh<v><iv><inf><@-FMAINV>$ ^fága<n><sggencmp><cmp>/faage<n><sggencmp><cmp>$ ^iskkus<n><sg><ill><@←ADVL>/pryövenasse<n><sg><ill><@←ADVL>$ ^.<clb>/.<clb>$
5-9.	^Datne<prn><pers><sg2><nom>$ ^edtjedh<v><pri><sg2>$ ^faage<n><sggencmp><cmp>+pryövenasse<n><sg><ill>$ ^båetedh<v><iv><inf>$^.<clb>$
10.	Datne edtjh faagenpryövenassese båetedh.

Table 1: Translation process for the first part of the sentence in Example 1: *Don galggat boahtit fágaiskosii.* 'You should take the apprenticeship exam'. Note that some tags have been modified for readability. The stages are as follows: Morphological analysis (1), Morphological disambiguation (2), Lexical transfer (3), Lexical selection (4), Structural transfer (5–9), Morphological generation (10).

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

these are simply listed in the lexicon.

4.2.2 Lexical selection

Because we were adapting the translator to a single, specialised domain, we only made a couple of rules for lexical selection. A translator for more domains will certainly need more such rules.

The pivot model causes some extra lexical selection. Even if the lexical conceptualisation often is similar in South Saami and North Saami, there are also counter examples, like e.g. the Norwegian verbs 'lese, telle, uttale' ("read, count, utter") which all three can be expressed with the North Saami verb *lohkat*. But South Saami usually uses different verbs for these, and the lexical distribution is like the one for Norwegian: *lohkedh, ryöknedh, jïehtedh*.

4.2.3 Structural transfer

The syntactic differences between North Saami and South Saami are greater than what is usually dealt with between related language pairs in Apertium. In order to be able to transfer VO structures to OV more reliably, the transfer phase is split into five parts instead of the three more typically used in Apertium:

- `chunker`: Chunk input words into groups, e.g. noun groups, verb chains.
- `interchunk1`: Merge chunks which have local coordination, e.g. the sequence [NP x] [CC and] [NP y] is merged into [NP x and y].
- `interchunk2`: Merges relative clauses and adpositional phrases, e.g. [NP x] [REL which] [V y] is merged into [NP x which y]; [NP x] [NP y] [POST on] into [PP x y on].
- `interchunk3`: Reorder constituents, e.g. SVO → SOV.
- `postchunk`: Cleanup.

4.2.4 Word form generation

Word form generation is done using a finite-state transducer (FST), compiled using the HFST compiler. Each lexical unit which is output by the postchunk module is looked up in the generation transducer and the surface form is output.

In order to make the general-purpose FST useful for MT generation, we had to ensure that one and only one form was generated for every grammatical word (lemma-tag combination) In case of dialect differences, it is in principle possible to make two or more parallel MT systems, translating into one dialect or the other. In this case we did not do that, but generated one variant only. The parallel forms were marked in the source code with a tag `+Use/NG` (for *not generate*).

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

Bilingual dictionary (`sme→sma`)	15,204
Transfer rules (`sme→sma`)	57

Table 2: Number of bilingual dictionary entries and transfer rules.

Corpus	**Tokens**	**Coverage (%)**
`general`	17,765,348	87.4 ± 1.82
`schooladmin`	448,424	91.1 ± 0.38

Table 3: Vocabulary coverage of the system. For the `schooladmin` corpus around 2 % of the coverage can be attributed to productive derivations, and 3 % to productive compounds. For the `general` corpus, this was 4 % for derivations and 6 % for compounds. The number for the `schooladmin` corpus was lower because of the addition of domain specific word pairs, many of them compounds.

5 Evaluation

We evaluated the system for both vocabulary coverage and translation quality. For vocabulary coverage, we took two corpora, one from the domain (`schooladmin`) and one from the general domain. To calculate vocabulary coverage, we split both corpora into ten parts and for each of these ten parts we calculated the *naïve coverage*,[3] that is the proportion of words receiving at least one analysis. We then calculated the mean and standard deviation. The results can be found in Table 3.

For translation quality, we selected a text in the domain of school books. The original Norwegian text consisted of 288 words, and the manual translation to North Saami consisted of 222 words (the text had not been used when developing the system). The North Saami text was then translated to South Saami by the system, and the South Saami translation and the Norwegian original were given to three translators. They were asked to postedit the MT output to make an adequate South Saami translation, and also to fill in a feedback questionnaire.

5.1 Quantitative evaluation

For the quantitative evaluation we used the Word-Error Rate (WER) and the Position-independent Word Error Rate (PER).[4] The results are shown in table 4. Compared to other Apertium Machine Translation systems,[5] the error rate is high. Correcting for word

[3]This is *naïve* as forms counted as 'known' by this measure may have other analyses which are not delivered by the transducer.

[4]The Word Error Rate is the sum of substitutions, deletions and insertions per word, made by an evaluator correcting the target text. The WER was calculated using the `apertium-eval-translator` tool.

[5]Unhammer and Trosterud (2009) give 17.7% for Bokmål-Nynorsk, Armentano-Oller et al. (2006) give 8.3% for Spanish-Portuguese.

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

	Eval. A	Eval. B	Eval. C	Mean
WER (%)	62.37	49.46	52.69	54.84
PER (%)	36.20	28.32	28.32	30.94
PER/WER	0.58	0.57	0.53	0.56

Table 4: Results for each of the translators and evaluation metrics. Excluding differences in word order, we can see that the system generates the right translation nearly two thirds of the time on average.

order (PER/WER), we see that more than half the errors are due to differences in word order.

5.2 Qualitative evaluation

5.2.1 Lexicon and morphology

As was the case for North Saami, also the South Saami MT-text consisted of 222 words (one word was not translated). Table 5 (first part, on lexical choice) shows how many times the evaluators preferred another lemma than the one provided by the system (not including cases where choice of lemma was related to a different syntax). The three evaluators had chosen at least 24 other words than the MT-system. But only in 9 of these 24 cases all the evaluators had chosen another word. Here we do not take in account one word which was added to the bilingual lexicon in another dialect, thereby causing a deviating stem vowel. The MT-system offers translation of dynamic compounds, and there were seven such words in the MT-text that at least one of the translators accepted, e.g. the Norwegian 'lærekontrakt, læringsarbeid, lærebedrift, programområde' ("apprentice contract, apprentice work, apprentice company, programme area") was given in North Saami as *oahppasoahpamuš, oahppanbargu, oahppafitnodat, p rográmmasuorgi* and then translated to South Saami as *barkoelotjkese, lïeremebarkoe, learasïelte, programmensuerkie*. One dynamic compound none of the evaluators accepted: 'læretid' ("period of training"), for which the North Saami word *oahppoáigi* was translated to *learoeaejkie*. The evaluators preferred *learoetïjje* or *learoeboelhke*. One additional deviation was the dialectal choice of final wovel – *oe* versus – *a*. The system had *l earalatjkoe, learasïelte, learasijjie*, but the translators preferred *learoelatjkoe, learoesïelte, learoesijjie*, ("apprentice period, apprentice company, apprentice place").

The generated morphology was quite good, but five cases of incorrect or lack of disambiguation of the North Saami text resulted in generation of the wrong morphological form in South Saami. Five words were covered only by derivation, like passive verbs and action nominal, and three of them were accepted by at least one evaluator. In one case the evaluators had preferred a plural noun, like in the original Norwegian text, but the MT-text offered a singular noun grounded in the North Saami translation.

Some North Saami verbs require a certain case which is not the correct one for the corresponding South Saami verb. For example, *oasselastit* 'take part, participate' takes

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

	MT	A	B	C	All evaluators disagree with MT
Different lexical choice		45	31	24	9
Some syntactic constructions					
Verb final (O V and Advl V)	4	10	12	12	6
Dem + N rather than bare N	3	7	3	8	0
Indef + N rather than bare N	0	6	0	1	0
Genitive modifier rather than complement	6	2	3	5	1
Possessive Verb *utnedh*	0	5	2	2	0

Table 5: Some of the translation differences between the MT version and the evaluators' versions. The last column shows in how many cases all the evaluators differed from the MT version.

an illative object in North Saami and an inessive one in South Sami — 'participate into' vs. 'participate in'.

5.2.2 Syntactic issues and variation

The evaluation revealed that the syntax was the poorest part of the MT-translation. In this chapter we look at some central syntactic constructions (core VP and NP structure) with big variation between the evaluators.

In Norwegian nouns are either indefinite or definite. In North Saami there is no definiteness expressed in the morphology of the noun. But in South Saami definiteness can be expressed in plural, as explained in Section 3. There was no rule for this in the MT-system at all, and in one such case all the evaluators disagreed with the MT-text.

Table 5 (second part) reveals the variation between the evaluators. For possessive the MT-system generated only the construction *possessor*.GEN *copula*.V *possessed*.NOM, but the evaluators preferred to some extent also using another construction: *possessor*.NOM *utnedh*.V *possessed*.ACC. One of the evaluators used it five times, the others use it twice, but there was no example of all evaluators disagreing with the MT version.

The MT-system did not generate indefinite articles, because there are none in North Saami, and they were not commonly used in traditional South Saami. One of the evaluators still used indefinite articles six times, another one used it once and the last one did not use it at all. In all cases there was an indefinite article in the Norwegian source text.

There is also no definite article in North or South Saami, but it is possible to express definiteness with a demonstrative pronoun. In the machine translated text there were three demonstrative pronouns in NP-initial position. Two of the evaluators used this possibility much more, respectively seven and eight times, but there was no example of all evaluators using a demonstrative pronoun and the MT system not using it.

NP complements in Norwegian, like in the text 'tilbud fra en lærebedrift' ("offer from a training establishment") were in the North Saami text translated to phrases with gen-

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

itive modifiers, in this case with accusative because it is an object: *oahppofitnodaga*.GEN *fálaldaga*.ACC. Two of the South Saami translations follow the Norwegian construction: *faalenassem*.ACC *aktede*.ART *learoesïelteste*.ELA. One of the translators showed that also the North Saami construction with genitive modifier is possible in South Saami for this particular phrase: *learoesïelten*.GEN *faalenassem*.ACC. Only one phrase with genitive modifier in the MT version was rejected of all evaluators.

In all cases under discussion in this section, there is variation in South Saami, and the MT system consistently chooses the traditional, more Uralic-like, construction.

5.2.3 Word order

As can be seen in Table 5, word order was problematic for the MT system.

The evaluators used 10–12 *X*V constructions in their texts (where *X* is object or adverbial), whereas in the MT-text there were only four such constructions, and for 6 of the 12 constructions all evaluators disagreed with the V*X* suggestions of the MT system.

The reason for this poor performance is that the rules for transforming the word order from VX in North Saami to XV in South Saami were not able to cope with all the different syntactic constructions VX was a part of, like omitting of subject, progressive constructions, complex objects and verb phrases as complement to nouns or adjectives. One could cover these constructions by making more transfer rules, or one could make better use of the syntactic analysis of the North Saami input, which mark the object and the verb also in such complex syntactic constructions as in Figure 3[6]. In the example sentence, the last verb phrase is a complement to the noun (@N<). The object is marked with @-F<OBJ, telling that it is the object for a verb in a infinite verb construction, and the mother verb is to the left. The verb to the immediate left of the object is marked @>N, modifier of the object (modifies noun to the right), and the mother verb is the next verb, marked @N<, complement to the noun to the left. The pronoun is marked <hab> (habitive), telling that it is part of a possessive construction, and the possessed noun is marked with <ext> (existential).

5.3 User feedback

The evaluators were asked to fill in a questionnaire (see Figure 4 for a sample[7]).

According to evaluator A, the MT-translation was of no help at all. Both evaluator B and C were much more positive, and evaluator B reported saving 1/3 of the time by editing the MT-translation instead of translating from scratch. Both B and C found useful terms in the MT-text, and C stressed that the MT-text contained words not found in the dictionaries. Table 5 reveals that these two evaluators' translations were much closer to the MT-translation than evaluator A's translation was.

We also discussed the MT-translation with a South Saami linguist and native speaker of South Saami. This linguist saw the possibility of using MT as a help for discussing terminology and a means of getting the terms recommended by the normative body put

[6]Syntactic tags are in Figure 3 marked with @, and the arrow symbols ">" and "<" point to mother (dependency) node. @N< is thus "right complement to noun", and @<SUBJ is "subject with mother to the left".

[7]The whole questionnaire may be found here: http://giellatekno.uit.no/doc/mt/smesma/questionnaire.html.

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

```
"<Dus>"
    "don" Pron Pers Sg2 Loc <hab> @ADVL>
"<lea>"
    "leat" V IV Ind Prs Sg3 @+FMAINV
"<vuoigatvuohta>"
    "vuoigatvuohta" N Sg Nom <ext> @<SUBJ
"<oažžut>"
    "oažžut" V TV Inf @N<
"<heivehuvvon>"
    "heivehit" V* TV* Der/PassL V IV PrfPrc @>N
"<oahpahusa>"
    "oahpahus" N Sg Acc @-F<OBJ
"<.>"
    "." CLB
```

Figure 3: The input analysis of the sentence ”You have the right to get adapted education”

```
Sammenlikning av tiden du brukte på oversettingene - ett kryss:

1.   Sammenlikne tiden du brukte på oversettingene:
       _ Jeg brukte lenger tid på tekst A enn på tekst B:
            Anslå forskjellen, f.eks. 2 ganger mer, 1,5 mer osv., eller du kan
             oppgi antall timer du brukte på tekst A og på tekst B:
       _ Jeg brukte omtrent like lang tid på tekst A og tekst B:
       _ Jeg brukte lenger tid på tekst B enn på tekst A:
2.   Var tekst A og tekst B like vanskelige? - ett kryss
       _ A var vanskeligere
       _ omtrent like vanskelige
       _ B var vanskeligere Anslå forskjellen:
```

Figure 4: Fragment of the questionnaire sent to translators (for translation, see the appendix).

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

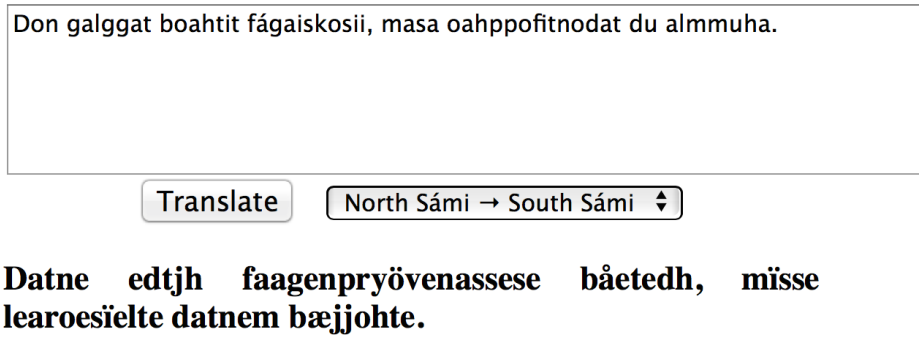

Figure 5: The internal web interface for testing with some example output for the sentence in Example 1.

into use. Seeing texts translated from North Saami, with less impact from the Germanic languages, may also be interesting for the language community. On the other hand, the linguistic impact from North Saami as the 'big brother' may also be controversial.

6 Future work

The system presented here is in an embryonic stage, representing only three person-months of work. An eventual scaling up of the system should concentrate along two lines: improving the lexical coverage, and improving the transfer rules. As the Saami written languages have radically different orthographic principles, there are no free rides within the vocabulary, all terms to be translated must be stated as such in the bilingual dictionary. Lexical coverage is thus crucial to the performance of the systems.

Regarding the interface, we currently have a very limited web-based interface for translating text (see Figure 5), but this could easily be extended for the translation of documents.

We will continue the work on a pivot-translation project with North Saami as the source language and South Saami as one of several candidates for a target language. We have already started working on the next language pair, North Saami to Inari Saami.

7 Conclusions

We have shown that a rule-based translation system is able to deliver a pivot translator good enough to be of use to translators, at least as an aid in the editing process.

A coverage of 89.7 % for the general and 92.4 % for the special domain, with only 4,000 entries in the bilingual dictionary, is a remarkable result. The system's ability to handle dynamic derivation and compounding explains 2 % and 3 % of the coverage, but even without this ability, the lexical coverage is over 85 %.

The evaluators edited the MT text quite much, but there was big variation between their texts, and most constructions in the MT text were accepted by at least one of them.

Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27
DOI 10.3384/nejlt.2000-1533.1642

The evaluators agreed that the word order was wrong in 6 cases for verb-object and verb-adverbial. One reason why two of the evaluators still were that positive might be the paucity of good lexical and terminological resources for translation into South Saami. In this situation, the bilingual lexicon, which was derived from a parallel corpus of text and thereby containing wordpairs not found in existing paper dictionaries, was welcomed as a useful translation resource, despite their syntactic changes.

The linguistic setting surrounding the translation system is actually quite common: a language society with one unrelated majority language, and several closely related minority languages, one of which is larger than the others. The present paper shows that manual translation to the pivot language and MT to related languages (here: to one related language) is indeed a viable option.

This work also shows that it is possible to make a functioning rule-based systems also with a small transfer lexicon and a small number of rules. The challenge in this project was the big syntactic differences between the two languages, but the remaining word order errors are within reach of an improved rule set, based on the informative syntactic analysis of the source language. Obtaining good lexical coverage is far from trivial when the text resources for the target language is so limited as in this case. But with the rule-based approach parallel text is only one of several options, integrating terminological work with the MT transfer lexicon is another.

Acknowledgments

We would like to thank Linda Wiechetek for initial work on the system, and Maja Lisa Kappfjell for work on the bilingual dictionary. We would also like to thank them both for fruitful discussions.

Appendix

Translation of the questionnaire:

```
Compare the time you spent on the translations - tick off once:
   I spent more time on text A than on text b:
   estimate the difference, e.g. 2 times more, 1.5 time more,
   or you may state how many hours you spent on text A and on text B
   I spent approximately as much time  on text A as on text B
   I spent more time on text B than on text A

Were text A and B equally difficult - tick off once
   A was more difficult
   The texts were equally difficult
   B was more difficunt
   Estimate the difference
```

References

Antonsen, Lene, Trond Trosterud, and Linda Wiechetek. 2010. Reusing Grammatical Resources for New Languages. In *Proceedings of LREC-2010*. Valetta, Malta: ELRA.

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez Gema Ramírez-Sánchez, , and Miriam A. Scalco. 2006. *PROPOR*, chap. Open-Source Portuguese–Spanish Machine Translation, pages 50–59. LNAI 3960. Springer-Verlag Berlin Heidelberg.

Babych, Bogdan, Tony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of the MT Summit XI*, pages 29–35.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation* 25(2):127–144.

Lindén, Krister, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen. 2011. HFST-Framework for Compiling and Applying Morphologies. In C. Mahlow and M. Pietrowski, eds., *Systems and Frameworks for Computational Morphology*, vol. 100 of *Communications in Computer and Information Science*, pages 67–85.

Masselot, Francois, Petra Ribiczey, and Gema Ramírez-Sánchez. 2010. Using the Apertium Spanish-Brazilian Portuguese machine translation system for localization. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation, EAMT10*.

Rasmussen, Torkel. 2015. Samisk språk i grunnskolen og videregående opplæring. *Samiske tall forteller Kommentert samisk statistikk 2015* 8:17–41.

Sammallahti, Pekka. 1998. *The Saami Languages. An Introduction*. Davvi Girji.

Todal, Jon. 2007. *Samisk språk i Svahken Sijte. Sørsamisk vitalisering gjennom barnehage og skule*, vol. 1 of *Dieđut*. Sámi Instituhtta – Nordisk Samisk Institutt.

Trosterud, Trond and Kevin Brubeck Unhammer. 2012. Evaluating North Sámi to Norwegian assimilation RBMT. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012)*, pages 13–26.

Tyers, Francis, Linda Wiechetek, and Trond Trosterud. 2009. Developing prototypes for machine translation between two Sámi languages. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation, EAMT09*, pages 120–128.

Unhammer, Kevin Brubeck and Trond Trosterud. 2009. Reuse of free resources in machine translation between Nynorsk and Bokmål. In J. A. Pérez-Ortiz, F. Sánchez-Martínez, and F. M. Tyers, eds., *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 35–42. Alicante.

Wiechetek, Linda, Francis Tyers, and Thomas Omma. 2010. Shooting at flies in the dark: Rule-based lexical selection for a minority language pair. *Lecture Notes in Artificial Intelligence* 6233:418–429.