North Sámi to Finnish rule-based machine translation system

Ryan Johnson[1] and **Tommi A Pirinen**[2] and **Tiina Puolakainen**[3]
Francis Tyers[1] and **Trond Trosterud**[1] and **Kevin Unhammer**[1]
[1] UiT Norgga Árktalaš universitehta, Giela ja kultuvrra instituhtta, Romssa, Norga
[2] Universität Hamburg, Hamburger Zentrum für Sprachkorpora, Deutschland
[3] Institute of the Estonian Language, Estonia
ryan.txanson@gmail.com, tommi.antero.pirinen@uni-hamburg.de,
tiina.puolakainen@eki.ee, {francis.tyers, trond.trosterud}@uit.no
kevin@unhammer.org

Abstract

This paper presents a machine translation system between Finnish and North Sámi, two Uralic languages. In this paper we concentrate on the translation direction to Finnish. As a background, the differences between the two languages is presented, followed by how the system was designed to handle some of these differences. We then provide an evaluation of the system's performance and directions for future work.

1 Introduction[0]

This paper presents a prototype shallow-transfer rule-based machine translation system between Finnish and North Sámi. The paper will be laid out as follows: Section 2 gives a short review of some previous work in the area of Uralic-Uralic language machine translation; Section 3 introduces Finnish and North Sámi and compares their grammar; Section 4 describes the system and the tools used to construct it; Section 5 gives a preliminary evaluation of the system; and finally Section 6 describes our aims for future work and some concluding remarks.

2 Previous work

Within the Apertium platform, work on several MT systems from North Sámi to Norwegian and to other Sámi languages have been developed (Tyers et al., 2009; Wiechetek et al., 2010; Trosterud and Unhammer, 2013; Antonsen et al., 2016)). Besides these systems, several previous works on making machine translation systems between Uralic languages exist, although to our knowledge none are publicly available, except for North Sámi to Norwegian[1], and the translation between Estonian, Finnish and Hungarian being available via English as a pivot language in Google Translate.[2] For non-Uralic pairs there are also numerous similarly laid out systems e.g. in Apertium's Turkic pairs, e.g. (Salimzyanov et al., 2013), that can offer insights on how the pair is implemented, which are detailed later in the article but the main parts are the same.

3 The languages

North Sámi and Finnish belong to the Sámi and Finnic branches of the Uralic languages, respectively. The languages are mutually unintelligible, but grammatically quite similar. The orthographical conventions between Finnish and North Sámi written in Finland were quite similar until 1979, when an unified North Sámi orthography widened the distance to Finnish. Finnish is primarily spoken in Finland, where it is the national language, sharing status with Swedish as an official language. The total number of speakers is at least 6 million people. North Sámi is spoken in the Northern parts of Norway, Sweden and Finland by approximately 24.700 people, and it has, alongside the national language, some official status in the municipalities and counties where it is spoken. North Sámi speakers are bilingual in their mother tongue and in their respective national language, many also speak the neighbouring official language. An MT system between North Sámi and Finnish is potentially of great use to the language communities, although fulfilling different functions. In Finland, it may be used to understand Sámi text, and in Norway and Sweden, it may be used by North Sámi speakers to understand Finnish text. In principle, the system may also be used for North Sámi text production, although further de-

[0] Authors are listed here alphabetically

[1] `https://gtweb.uit.no/jorgal`

[2] `https://translate.google.com`

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 115–122,
Gothenburg, Sweden, 23-24 May 2017. ©2017 Linköping University Electronic Press

velopment will be needed to fulfil such a function.

3.1 Phonological differences

As related languages, Finnish and North Sámi share several phonological processes, the most important one being consonant gradation. However, North Sámi consonant gradation involves the vast majority of stem-internal consonant clusters, whereas the Finnish counterpart involves only the stops *p, t, k*. Vowel length has a more central role in Finnish than in North Sámi, Several instances of final vowel apocopy in North Sámi, as well as a neutralisation of *p, t, k* in word-final position, has also resulted in quite extensive morphological homonymy. A richer inventory of affricates and fricatives in North Sámi, as well as preaspiration, also add to the difference.

3.2 Orthographic differences

In the native vocabulary, neither Finnish nor North Sámi distinguish between voiced and unvoiced plosives, but whereas Finnish writes them as *p, t, k*, North Sámi writes *b, d, g*, as in *kirja : girji* “book”. Finnish marks vowel length with double letter symbols. In North Sámi this distinction is marked for one vowel only, *a*, and with acute accent. Apart from this the orthographic principles of the two languages is quite similar, the almost total lack of free rides is a result of different phonology.

3.3 Morphological differences

There are a number of examples where the morphologies of Finnish and North Sámi are rather different.

North Sámi has a separate dual number, whereas Finnish has not. Otherwise the North Sámi and Finnish finite verb morphology is almost identical. The infinite verb conjugation is more different, though: Finnish has a rich array of infinitives that are inflected in different subsets of the case system.

Finnish has more than twice the number of cases as North Sámi has. Where North Sámi only has one case for the direct object (accusative), Finnish has two (accusative and partitive). The Finnish system of adverbial cases consist of a 2x3 matrix of inner/outer to/in/from cases, North Sámi has only one of these distinctions (to/in˜from), thus the 6 Finnish cases corresponds to 2 North Sámi ones. In principle, Finnish and North Sámi have the same system of possessive suffixes, but in North Sámi its use is far more restricted than in Finnish.

3.4 Syntactic differences

Syntactically speaking, there are two varieties of North Sámi, one used within and one outside of Finland. The Finnish variety is much closer to Finnish than the Scandinavian one. Comparing Finnish with the Scandinavian variety of North Sámi, the most striking difference is participle constructions vs. relative clauses. Where North Sámi uses subordinate clauses, written Finnish often use head-final participle constructions instead. Since both varieties are found in Finnish, at least to some degree, we at the moment let most “Scandinavian” varieties of North Sámi through, thereby giving North Sámi from Norway and Finland a different stylistic flavour in the Finnish output.

The North Sámi passive is a derivational process, whereas it for Finnish is an inflectional one, resulting in quite different syntactic patterns for passive. Finnish has a richer array of indefinite verb forms.

Finnish adjectives agree with their head noun in case and number, whereas North Sámi has an invariant *attribute* form for all but one adjective, the adjective *buorre* ‘good’, and partial agreement for determiners.

Existential and habitive clauses have the same structure in the two languages, *possessor.local-case copula possessed* and *adverbial copula e-subject* (*on me / in street is car* ‘I have a car/There is a car in the street’). except that in Finnish, the possessed/e-subject behaves like objects, whereas it in North Sámi they behave like subjects. Thus, in North Sámi, the copula agrees with the possessed / e-subject, whereas in Finnish, it does not.

4 System

The system is based on the Apertium[3] machine translation platform (Forcada et al., 2011). The platform was originally aimed at the Romance languages of the Iberian peninsula, but has also been adapted for other, more distantly related, language pairs. The whole platform, both programs and data, are licensed under the Free Software Foundation’s General Public Licence[4] (GPL) and all the software and data for the 30 released language pairs (and the other pairs being worked on) is available for download from the project website.

[3] `http://apertium.sf.net`

[4] `https://www.gnu.org/licenses/gpl-3.0.en.html`

4.1 Architecture of the system

The Apertium translation engine consists of a Unix-style pipeline or assembly line with the following modules (see Figure 1):

- A deformatter which encapsulates the format information in the input as *superblanks* that will then be seen as blanks between words by the other modules.

- A morphological analyser which segments the text in surface forms (SF) (words, or, where detected, multiword lexical units or MWLUs) and for each, delivers one or more lexical forms (LF) consisting of lemma, lexical category and morphological information.

- A morphological disambiguator (CG) which chooses, using linguistic rules the most adequate sequence of morphological analyses for an ambiguous sentence.

- A lexical transfer module which reads each SL LF and delivers the corresponding target-language (TL) LF by looking it up in a bilingual dictionary encoded as an FST compiled from the corresponding XML file. The lexical transfer module may return more than one TL LF for a single SL LF.

- A lexical selection module (Tyers et al., 2012b) which chooses, based on context rules, the most adequate translation of ambiguous source language LFs.

- A structural transfer module, which performs local syntactic operations, is compiled from XML files containing rules that associate an action to each defined LF pattern. Patterns are applied left-to-right, and the longest matching pattern is always selected.

- A morphological generator which delivers a TL SF for each TL LF, by suitably inflecting it.

- A reformatter which de-encapsulates any format information.

Table 1 provides an example of a single phrase as it moves through the pipeline.

4.2 Morphological transducers

The morphological transducers are compiled with the Helsinki Finite State Technology (Lindén et al., 2009),[5] a free/open-source reimplementation of the Xerox finite-state tool-chain, popular in the field of morphological analysis. It implements both the lexc morphology description language for defining lexicons, and the twol and xfst scripting languages for modeling morphophonological rules. This toolkit has been chosen as it—or the equivalent XFST—has been widely used for other Uralic languages (Koskenniemi, 1983; Pirinen, 2015; Moshagen et al., 2013), and is available under a free/open-source licence. The morphologies of both languages are implemented in lexc, and the morphophonologies of both languages are implemented in twolc.

The same morphological description is used for both analysis and generation. To avoid overgeneration, any alternative forms are marked with one of two marks, LR (only analyser) or RL (only generator). Instead of the usual compile/invert to compile the transducers, we compile twice, once the generator, without the LR paths, and then again the analyser without the RL paths.

4.3 Bilingual lexicon

The bilingual lexicon currently contains 19,415 stem-to-stem correspondences (of which 8044 proper nouns) and was built partly upon an available North Sámi—Finnish dictionary[6], and partly by hand (i.e., by translating North Sámi stems unrecognised by the morphological analyser into Finnish). The proper nouns were taken from existing lexical resources. Entries consist largely of one-to-one stem-to-stem correspondences with part of speech, but also include some entries with ambiguous translations (see e.g., Figure 2).

4.4 Disambiguation rules

The system has a morphological disambiguation module in the form of a Constraint Grammar (Karlsson et al., 1995). The version of the formalism used is vislcg3.[7] The output of each morphological analyser is highly ambiguous, measured at around 2.4 morphological analyses per form for Finnish and 2.6 for North Sámi[8]. The goal of

[5] `https://hfst.github.io`

[6] `https://gtweb.uit.no/langtech/trunk/words/dicts/smefin/src/`

[7] `http://visl.cg.sdu.dk`

[8] Cf. (Trosterud and Wiechetek, 2007)

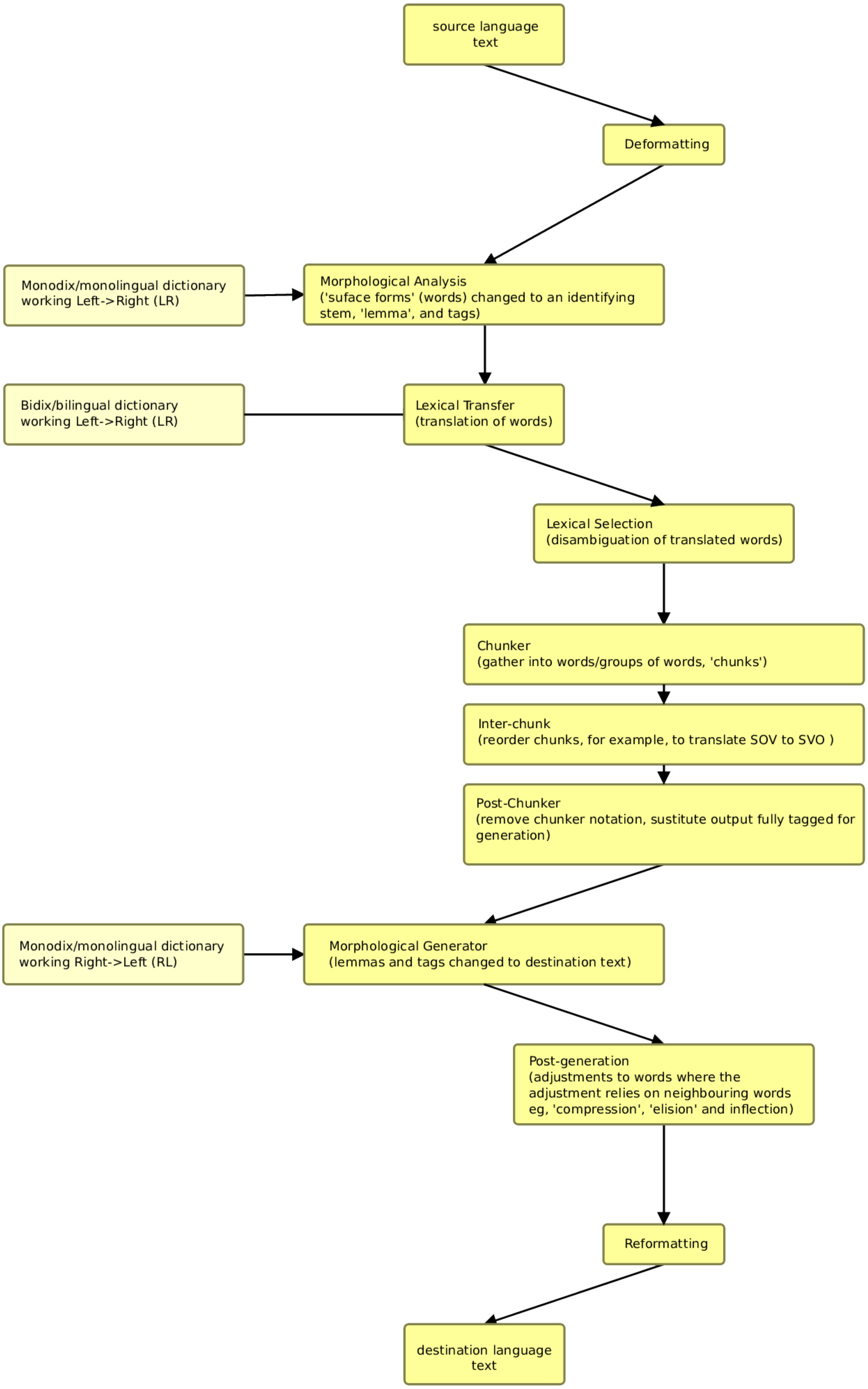
source language text
Deformatting
Monodix/monolingual dictionary working Left->Right (LR)
Morphological Analysis ('suface forms' (words) changed to an identifying stem, 'lemma', and tags)
Bidix/bilingual dictionary working Left->Right (LR)
Lexical Transfer (translation of words)
Lexical Selection (disambiguation of translated words)
Chunker (gather into words/groups of words, 'chunks')
Inter-chunk (reorder chunks, for example, to translate SOV to SVO)
Post-Chunker (remove chunker notation, sustitute output fully tagged for generation)
Monodix/monolingual dictionary working Right->Left (RL)
Morphological Generator (lemmas and tags changed to destination text)
Post-generation (adjustments to words where the adjustment relies on neighbouring words eg, 'compression', 'elision' and inflection)
Reformatting
destination language text

Figure 1: Apertium structure (Image from apertium wiki by user Rcrowther) http://wiki.apertium.org/wiki/Workflow_diagram

North Sámi input: Sámegielat leat gielat maid sámit hállet.
Morphological analysis:
ˆSámegielat/ sámegielat<adj> <attr>/ sámegielat<adj> <sg><nom>/ sámegiella<n> <pl><nom>/ sámegiella<n> <sg><acc><px2sg>/ sámegiella<n> <sg><gen><px2sg>/ sámegiella<n> <sg><acc><px2sg>/ sámegiella<n> <sg><gen><px2sg>$ ˆleat/ leat<vblex><iv> <indic><pres><conneg>/ leat<vblex><iv> <indic><pres><p1><pl>/ leat<vblex><iv> <indic><pres><p2><sg>/ leat<vblex><iv> <indic><pres><p3><pl>/ leat<vblex><iv> <inf>$ ˆgielat/ giella<n> <pl><nom>/ giella<n> <sg><acc><px2sg>/ giella<n> <sg><gen><px2sg>/ giella<n> <sg><acc><px2sg>/ giella<n> <sg><gen><px2sg>$ ˆmaid/ maid<adv>/ mii<prn><itg> <pl><acc>/ mii<prn><itg> <pl><gen>/ mii<prn><itg> <sg><acc>/ mii<prn><rel> <pl><acc>/ mii<prn><rel> <pl><gen>/ mii<prn><rel> <sg><acc>$ ˆsámit/ sápmi<n> <pl><nom>/ sápmi<n> <pl><nom>$ ˆhállet/ hállat<vblex><tv> <imp><p2><pl>/ hállat<vblex><tv> <indic><pres><p3><pl>/ hállat<vblex><tv> <indic><pret><p2><sg>$ ˆ./.<sent>$
Morphological disambiguation:
ˆSámegielat/sámegiella<n> <pl><nom> <@SUBJ→>$ ˆleat/leat<vblex><iv> <indic><pres><p3><pl> <@+FMAINV>$ ˆgielat/giella<n> <pl><nom> <@←SPRED>$ ˆmaid/mii<prn><rel> <pl><acc> <@OBJ→>$ ˆsámit/sápmi<n> <pl><nom> <@SUBJ→>$ ˆhállet/hállat<vblex><tv> <indic><pres><p3><pl> <@+FMAINV>$ ˆ./.<sent>$
Lexical translation:
ˆSámegiella<n> <pl><nom> <@SUBJ→>/ Saamekieli<n> <pl><nom> <@SUBJ→>/ Saame<n> <pl><nom> <@SUBJ→>$ ˆleat<vblex><iv> <indic><pres><p3><pl> <@+FMAINV>/ olla<vblex> <actv><indic><pres><p3><pl> <@+FMAINV>/ sijaita<vblex> <actv><indic><pres><p3><pl> <@+FMAINV>$ ˆgiella<n> <pl><nom> <@←SPRED>/ kieli<n> <pl><nom> <@←SPRED>/ ansa<n> <pl><nom> <@←SPRED>$ ˆmii<prn><rel> <pl><acc> <@OBJ→>/ mikä<prn><rel> <pl><acc> <@OBJ→>$ ˆsápmi<n> <pl><nom> <@SUBJ→>/ saame<n> <pl><nom> <@SUBJ→>$ ˆhállat<vblex><tv> <indic><pres><p3><pl> <@+FMAINV>/ puhua<vblex> <actv><indic><pres><p3><pl> <@+FMAINV>/ mekastaa<vblex> <actv><indic><pres><p3><pl> <@+FMAINV>$ˆ.<sent>/.<sent>$
Structural transfer:
ˆSaamekieli<n> <pl><nom>$ ˆolla<vblex> <actv><indic><pres><p3><pl>$ ˆkieli<n> <pl><nom>$ ˆmikä<prn><rel> <pl><par>$ ˆsaame<n> <pl><nom>$ ˆpuhua<vblex> <actv><indic><pres><p3><pl>$ˆ.<sent>$
Finnish translation:
Saamekielet ovat kielet #mikä saamet puhuvat

Table 1: Translation process for the North Sámi phrase *Sámegielat leat gielat maid sámit hállet* (The Sámi languages are the languages that the Sámis speak)

```
<e><p><l>sálten<s n="n"/></l><r>suolaus<s n="n"/></r></p></e>
<e><p><l>sálti<s n="n"/></l><r>suola<s n="n"/></r></p></e>
<e><p><l>sámeduodji<s n="n"/></l><r>käsityö<s n="n"/></r></p></e>
<e><p><l>sámegiella<s n="n"/></l><r>saame<s n="n"/></r></p></e>
<e><p><l>sámegiella<s n="n"/></l><r>saamekieli<s n="n"/></r></p></e>
<e><p><l>sámi<s n="n"/></l><r>saame<s n="n"/></r></p></e>
<e><p><l>sámil<s n="n"/></l><r>sammal<s n="n"/></r></p></e>
```

Figure 2: Example entries from the bilingual transfer lexicon. Finnish is on the right, and North Sámi on the left.

the CG rules is to select the correct analysis when there are multiple analyses. Given the similarity of Finnish and North Sámi, ambiguity across parts of speech may often be passed from one language to the other and not lead to many translation errors. Disambiguating between forms within the inflectional paradigms in case of homonymy, on the other hand, are crucial for choosing the correct form of the target language, and there has been put much effort into developing CG rules to resolve such ambiguity for North Sámi. Currently, ambiguity is down to 1.08 for North Sámi (analysed with the disambigator used for MT on a 675534 word newspaper corpus[9]. The corresponding number for Finnish is 1.36, for a subcorpus of 770999 words of Wikipedia text. The Finnish CG rules are a conversion of Fred Karlsson's original CG1 rules for Finnish (Karlsson, 1990), and the poorer results for Finnish are due to conversion problems between the different CG version, and between CG1 and our Finnish FST.

5 Evaluation

All evaluation was tested against a specific version of Apertium SVN[10] and Giellatekno SVN[11]. The lexical coverage of the system was calculated over freely available corpora of North Sámi. We used a recent dump of Wikipedia [12] as well as a translation of the New Testament. The corpora were divided into 10 parts each; the coverage numbers given are the averages of the calculated percentages of number of words analysed for each of these parts, and the standard deviation presented is the standard deviation of the coverage on each corpus. As shown in Table 2, the naïve coverage[13] of the North Sámi to Finnish MT system over the corpora approaches that of a broad-coverage MT system, with one word in ten unknown.

Corpus	Tokens	Cov.	std
se.wikipedia.org	190,894	76,81 %	±10
New Testament	162,718	92,45 %	±0.06

Table 2: Naïve coverage of sme-fin system

The coverage over the Wikipedia corpus is substantially worse, due to the fact that this corpus is "dirtier": it contains orthographical errors, wiki code [14], repetitions, lot of English texts, as well as quite a few proper nouns, this is easily seen in the large deviation between divided parts. The New Testament on the other hand is rather well covered and has practically uniformly distributed coverage throughout.

To measure the performance of the translator we used the Word Error Rate metric—an edit-distance metric based on Levenshtein distance (Levenshtein, 1966). We had three small North Sámi corpora along with their manually post-edited translations into Finnish to measure the WER. We have chosen not to measure the translation quality with automatic measures such as BLEU, as they are not the best suited to measure quality of translations for the use case, for further details see also Callison-Burch et al. (2006; Smith et al. (2016; Smith et al. (2014).

For translation post-edition we used three freely

[9] Cf. (Antonsen and Trosterud, forthcoming) for a presentation of the North Sámi CG.

[10] `https//svn.code.sf.net/p/apertium/svn/nursery/apertium-sme-fin`: revision 76019

[11] `https://victorio.uit.no/langtech/trunk/langs/sme`: revision 147464

[12] `http://se.wikipedia.org`

[13] A non-naïve coverage would require manual evaluation of correctness for the cases where word-forms are covered accidentally by e.g., morphological processes.

[14] While we have tried to clean the data from most of the Mediawiki codes and notations, there is always some left after the cleanup, due to new wiki codes after creation of the cleanup script or actual broken data

Corpus	Tokens	OOV	WER
Redigering.se	1,070	95	34.24
Samediggi.fi	570	33	36.32
The Story	361	0	19.94

Table 3: Word error rate over the corpora; OOV is the number of out-of-vocabulary (unknown) words.

available parallel texts from the internet: one from the Finnish Sámi parliament site[15], one from a Swedish regulation of minority people and languages and the story that is used with all Apertium language pairs as initial development set ("Where is James?"). Table 3 presents the WER for the corpora.

Analysing the changes in post-edition, a few classes of actual errors can be identified. One common example arises from the grammatical differences in the case system systems, in particular the remaining adpositions are often turned into a case suffix for the dependent noun phrase, e.g., the North Sámi "birra" has been turned into the Finnish adposition "ympäri" (around), where elative case is required, similarly for the translation "seassa" (among) instead of inessive case. Also visible, especially in the story text is the lack of possessive suffix agreement e.g. "heidän äiti" (their mother n sg nom) instead of "heidän äitinsä" (their mother n sp nom/gen pxsp3), while the former is perfectly acceptable in standard spoken Finnish it is not accepted as formal written language form. Another issue that appeared a number of times, maybe partially due to the genre of the texts selected, i.e. law texts, was the selection of adverb (form), e.g. the word-form "mukana" (with) was corrected to "mukaan" (according to). A large amount of simple lexical problems is due to the vocabulary of the selected texts as well: "hallintoalue" (governmental area), "seurantavastuu" (responsibility of surveillance), "itsehallinto" (autonomy), and their compounds, are all either missing or partially wrong due to lexical selections.

6 Concluding remarks

We have presented the first MT system from Finnish to North Sámi. With a WER of above 30%, it still is far from production-level performance, and it is also at the prototype-level in terms of the number of rules. Although the impact of this relatively low number of rules on the quality of translation is extensive (cf., the difference in WER between the development and testing corpora), the outlook is promising and the current results suggest that a high quality translation between morphologically-rich agglutinative languages is possible. We plan to continue development on the pair; the coverage of the system is already quite high, although we intend to increase it to 95 % on the corpora we have we estimate that this will mean adding around 5,000 new stems and take 1–2 months. The remaining work will be improving the quality of translation by adding more rules, starting with the transfer component. The long-term plan is to integrate the data created with other open-source data for Uralic languages in order to make transfer systems between all the Uralic language pairs. Related work is currently ongoing from North Sámi to South, Lule and Inari Sámi, from North Sámi to Norwegian, and between Finnish and Estonian. The system presented here is available as free/open-source software under the GNU GPL and the whole system may be downloaded from Sourceforge and the open repository of Giellatekno.

Acknowledgements

The work on this North Sámi-Finnish machine translation system was partially funded by the Google Summer of Code and Google Code-In programmes, and partly by a Norsk forsingsråd grant (234299) on machine translation between Sámi languages.

References

Lene Antonsen and Trond Trosterud. forthcoming. Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk. Norsk Lingvistisk Tidsskrift.

Lene Antonsen, Trond Trosterud, and Francis Tyers. 2016. A North Saami to South Saami machine translation prototype. 4:11—27.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *EACL*, volume 6, pages 249–256.

Mikel L Forcada, 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.

[15] http://samediggi.fi

Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Fred Karlsson, Atro Voutilainen, Juha Heikkilae, and Arto Anttila. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th conference on Computational linguistics-Volume 3*, pages 168–173. Association for Computational Linguistics.

Kimmo Koskenniemi. 1983. *Two-level morphology—A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, Department of General Linguistics. University of Helsinki, Finland.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 28–47. Springer.

Sjur N Moshagen, Tommi A Pirinen, and Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16*, number 085, pages 343–352. Linköping University Electronic Press.

Tommi A Pirinen. 2015. Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. *SKY Journal of Linguistics*, 28:381–393.

Ilnar Salimzyanov, J Washington, and F Tyers. 2013. A free/open-source kazakh-tatar machine translation system. *Machine Translation Summit XIV*.

Aaron Smith, Christian Hardmeier, and Jörg Tiedemann. 2014. Bleu is not the colour: How optimising bleu reduces translation quality.

Aaron Smith, Christian Hardmeier, and Jörg Tiedemann. 2016. Climbing mount bleu: The strange world of reachable high-bleu translations. *Baltic Journal of Modern Computing*, 4(2):269.

Trond Trosterud and Kevin Brubeck Unhammer. 2013. Evaluating North Sámi to Norwegian assimilation RBMT. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012)*, volume 3 of *Technical report*, pages 13–26. Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg.

Trond Trosterud and Linda Wiechetek. 2007. Disambiguering av homonymi i Nord- og Lulesamisk. *Suomalais-Ugrilaisen Seuran Toimituksia = Mémoires de la Société Fi nno-Ougrienne. Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007*, 253:375–395.

Francis Tyers, 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.

Linda Wiechetek, 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.