A South Sámi Grammar Checker For Stopping Language Change

Anonymous Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymouser Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

We have developed and evaluated a basic rule-based South Sámi grammar checker for two frequent error types that are caused by and causing language change and a loss of the language's morphological richness. These error types regard adjective forms (confusion of attributive and predicative forms) and the negation paradigm, especially for past tense forms. Our work includes a classification of adjective paradigms, manual error mark-up in an evaluation corpus, and a tool for automatic error detection/correction. While we achieve decent results for adjective error correction, negation error correction still needs more work with regard to error classification and identifying syntactic context. There is a lot of potential for improvement of the tool as grammatical insights and rule-writing go hand in hand. The grammar checker for adjective error correction will be released May 2023 and will be freely available for download.

1 Introduction

South Sámi is in a critical situation that requires concrete measures so that morphological richness is taught to the next generation and does not get lost. In this article we will focus on grammar tool, which will support South Sámi writers in their grammatical choices when other help is not available.

We focus on two very frequent grammatical error types of morphological forms that the language community wishes to preserve. Those include adjective inflection and inflection of verbal periphrastic negation. An investigation in 2018 (Kappfjell and Trosterud, forthcoming) showed tendencies of adjective classes being reduced from four to two classes, and negation in South Sámi has been covered by Blokland and Inaba (2015).

The school system does not provide sufficient language support for South Sámi. Students only have a few hours a week to learn. The teachers of South Sámi then have to select what they teach, which are typically the topics that are satisfactorily described in the grammar books. Both the adjective and negation paradigms are described, but not in all their detail and with regard to the variations in the spoken language. Making a grammar checker that focuses on these neglected grammatical topics has the purpose of improving grammatical knowledge in these areas.

Regression testing shows that error correction for both negation and adjective forms look promising with precions of up to 80% when starting our work.

2 Background

2.1 Language situation

According to Blokland and Hasselblatt (2003, p.110), there are about 2,000 ethnic South Sámi, of which approximately 300-500 are South Sámi speakers. There are two major varieties in South Sámi: northern (or Asele) South Sámi and South (or Jamtland) South Sámi (Sammallahti, 1998, p.24), but the differences between the two are minor, and limited mostly to phonetics and morphology. South Sámi has a written standard, in which literature (especially children's literature) is published, it is also used to some extent on the Internet. South Sámi is an official language in the municipality of Hattfjelldal (South Sámi: Aarborte) in Nordland, Snåsa, (South Sámi: Snåase) and Rørvik (South Sámi: Raarvihke) in Nord Trøndelag and Røros (South sámi: Plaassja) in Trøndelag, four commiunities and a recognized (historical) minority language in Sweden. There are some minor differences between the orthogra-

phies used in Sweden and Norway·

There is a lack of standardization and clarification regarding variants, for example, the language at the system level is undergoing change and simplification, for example, in the field of adjectives and negation (which this article deals with) there are system changes among South Sámi writers. The program will help the writers with correct word forms but also how they behave in sentences.

2.2 Technical background

The technological implementation of the grammar checker is based on rule-based natural language processing: finite-state automata for morphological analysis (Beesley and Karttunen, 2003; Lindén et al., 2013) and constraint grammar (Karlsson, 1990b; Didriksen, 2010) for syntactic and semantic as well as other sentence-level processing. The South Sámi tools are publicly available[1] It is part of a multilingual infrastructure (url removed for anonymity) which includes 130 languages.

The grammar checker is built on a pipeline of modules: we process the input text with morphological analysers and tokenisers to get annotated texts, then disambiguate and then apply grammar rules on the disambiguated sentences.

The grammar checker takes input from the finite-state transducer (*FST*) to a number of other modules, the core of which are several Constraint Grammar modules for tokenisation disambiguation, morpho-syntactic disambiguation and a module for error detection and correction. The full modular structure is described in (removed for anonymity). We are using finite-state morphology (Beesley and Karttunen, 2003) to model word formation processes. The technology behind the *FSTs* is described in Pirinen (2014). Constraint Grammar is a rule-based formalism for writing disambiguation and syntactic annotation grammars (Karlsson, 1990a; Karlsson et al., 1995). In our work, we use the free open source implementation VISLCG-3 (Bick and Didriksen, 2015). All using the *GiellaLT* infrastructure (Moshagen et al., 2013).

2.3 Motivation

A recent survey shows that language technology is used to a far greater extent by minority languages and indigenous languages than by state-bearing majority languages such as Norwegian. (Trosterud, 2019) The size of the language community also plays a role: South Sámi use language technology aids to a far greater extent than Northern Sámi. Language technology tools are therefore central to the revitalization of South Sámi, and our goal is to be able to provide good tools to the South Sámi language community. South Sámi school children of the 80es who were taught by Anna Jacobsen, had a strong grammarian with clear expectations of how correct language should be as guidance. When language experts from the past generation pass away, the bearers of this knowledge disappear. In a reality where South Sámi is not used as frequently in daily life as it used to, we need other tools to ensure that feedback for correct and incorrect language is available. Otherwise, there is a lot of insecurity about it and instead of using the language, people keep quiet and do not dare to write.

3 The South Sámi grammar checker

3.1 Negation errors

Standard negation in South Sámi utilizes a negative auxiliary and a connegative form of the lexical verb. The basic paradigm usually presented in grammars, cf. (Bergsland, 1946, pp.169–170), (Hasselbrink, 1981-1985, p.145), (Henrik and Mattsson, 2012, p.38), is one where the negative auxiliary has two moods (indicative and imperative) and two simple tenses (present and preterite) The connegative form ends in -h and is homonymous with the second person singular of the imperative. Depending on inflection type, it may also be identical to the second person singular or the third person plural of the present indicative. (Blokland and Inaba, 2015)

But according to Blokland and Inaba (2015), there are several tendencies towards different inflectional patterns for negation.

One error type regards the negation verb itself. In past tense it should be in congruence with the subsequent past tense connegative form. In example (1), the form *ean* (1.Du) should actually be *eakan* (3.Du) as in ex. (2) as the connegative form *ligan* is a third person dual.

(1) *__Ean__ ligan dah gåetesne,
NEG.1DU be.PAST.3DU this home.INE.SG
mohte hæhtjosne vaeresne.
this cabin.INE.SG mountain.INE.SG
'They were not at home, but in the cabin in the mountain'

[1] url removed for anonymity

(2) **Eakan** ligan dah gåetesne, mohte hæhtjosne vaeresne.
NEG.3DU be.PAST.3DU this home.INE.SG this cabin.INE.SG mountain.INE.SG
'They were not at home, but in the cabin in the mountain'

```
ADD (&msyn-ConNegPrt-congruence)
TARGET (Prt ConNeg) + $$SG-PERS IF
(-1 ("ij" Prs Neg) - $$SG-PERS) ;
```

A second typical error is the the use of the third person singular form of the negation verb as a default, as in example (3). Here the first person dual form of the connegative form *limen* shows the actual person and number of the verb phrase, and the the negation verb should agreer with it.

(3) Ij limen månnoeh
NEG.3SG be.CONNEG.PAST.1DU there
desnie.

'We were not there.'

```
ADD (&msyn-Neg-VFinitt-ConNeg)
TARGET (Ind Prs) + $$ALL-PERS
OR (Ind Prt) + $$ALL-PERS
(-1 ("ij" Prs Neg) + $$ALL-PERS)
(NEGATE 0 ConNeg) ;
```

A third type changes a finite verb form to a connegative verb form, cf. ex. (4). Here, *edtjigan* should be changed to *edtjh*, and subsequently the tense of the negation verb *eakan* should be changed to past tense as marked by the connegative, i.e. *idtjigan*.

(4) Mohte eakan edtjigan
But NEG.PRS.3DU shallPAST.3DU
juakadidh.
drink
'But they shouldn't drink.'

3.2 Adjective errors

South Sámi grammars that write about the adjective system often state that the adjective paradigm is unclear. In the dictionaries and in the text corpora, there is a big variation.

According to earlier grammarians, two-syllable adjectives usually have two forms in the positive, one of them ending in a vowel and the other of them ending in *-s*.

These two forms can be attributive or predicative forms. Alternatively, there can be only one form for both attributive and predicative. According to earlier grammarians, the comparative forms are built on the predicative form. However, in today's South Sámi there are also comparatives built on attributive forms. Table 1 shows all four attribute-predicative combinations are those according to these grammars

Attributive	**Predicative**
vowel (*buerie*)	vowel (*buerie*)
vowel (*skïemtje*)	-s (*skïemtjes*)
-s (*vihkeles*)	vowel (*vihkele*)
-s (*båeries*)	-s (*båeries*)

Table 1: Possible combinations of adjective forms in positive

In addition to that, some of the adjective forms can also be adverbs. The predicative form *vihkele* 'important' for example is homonymous to the adverbial form. Other adjectives have more part-of-speech homonymies. *buerie* 'good' is for example both attributive and predicative form of an adjective, but can at the same time also be a noun. The form *båetije* 'coming' is both an adjective, deverbal noun and a present participle of a verb.

Kappfjell and Trosterud (forthcoming) show that text collections of modern South Sámi exhibit others tendencies on of adjectives inflection than its mentioned on the earlier grammars. They come to the conclusion that modern South Sámi shows the same system as before, but the attribute is more frequent than a predicative: 60% vs. 30%. The other tendency is that instead of four adjective classes, there are only two of them where attributive and predicative are homonymous, either ending in a vowel or in -s. The investigation shows, that predicative and attribute forms are the same in 98.4% of the cases. Only 8.7% of the adjective types display variation. This system appears to be very stable and consistent. However, there is a desire in the language community to revert the system and go back to and teach morphological richness to new generations.

We have to keep in mind that South Sámi language orthography was approved in 1978, and there has been a careful revitalization at the Sámi schools in Snåsa and Hattfjelldal. There are approximately 500 speakers, but only 1/10 actually write the language as well. South Sámi training has been deficient in that it has been cut short by only a few hours, and the teachers have thus not

been given the space they have needed to be able to provide complete training in the most important grammatical systems.

(5) Saemien kultuvre lea *gånkaladtje* jïh
Sámi culture is royal.PRED and
tjaebpies.
beautiful.PRED
'The Sámi culture is royal and beautiful.'

For a rule based grammar checker this means that we need to distinguish between adjectives that have one form for both attributive and predicative forms and those that differ in their forms. We resolve this by adding an early rule to the syntactic analyzer module preceding the grammar checking rules. The rule below adds a secondary tag <AttrPred> to each adjective with both an attributive (Attr) reading and a predicative reading in the same cohort. Since this rule precedes all disambiguation rules, both readings are still available, and the tag ensures that this information is kept throughout the analysis.

```
SUBSTITUTE (A) (A <AttrPred>)
TARGET A
IF (0 Attr LINK 0 (A Nom));
```

The error detection rules are ADD-rules. They add an error tag, here *&msyn-adj-attr-pred* to the erroneous form in a syntactic context. There are different syntactic contexts that require different types of rules. The one below pays attention to a nominative subject to its left and a possible copula between the adjective and the copula. Since copulas can be dropped in South Sámi, the subject can be an important marker. In addition it excludes a noun to its right.

```
ADD (&msyn-adj-attr-pred)
TARGET (A Attr) IF
(*-1 Nom
BARRIER (*) - REALCOPULAS - Ela)
(NEGATE 0 ATTR-PRED-A
OR A + Sg + Nom OR A-ATTR-ONLY)
(NOT 1 N) ;
```

The second context below is a visible copula that can be either by itself or together with a negation verb. If the subject is dropped, the copula is the decisive marker for predicative forms. Again we do not want a noun to the right of the adjective. This rule explicitly asks for an end of sentence after the adjective form.

```
ADD (&msyn-adj-attr-pred)
TARGET (A Attr) IF
(NEGATE 0 ATTR-PRED-A OR
A + Sg + Nom OR A-ATTR-ONLY)
(1 EOS)
(*-1 (Neg Ind) OR
REALCOPULAS BARRIER NOT-ADV-PCLE);
```

The third case is a coordination context where the predicative adjective is coordinated with another predicative adjective, which shows that the form should be predicative rather than attributive.

```
ADD (&msyn-adj-attr-pred)
TARGET (A Attr) IF

(-1 CC LINK *-1 Nom
BARRIER (*) - REALCOPULAS)
(NEGATE 0 ATTR-PRED-A
OR A + Sg + Nom OR A-ATTR-ONLY)
(NOT 1 N);
```

4 Evaluation

The evaluation is based on a part of *SIKOR*, the South Sámi free corpus containing administrative, law, religious, non-fiction, fiction, and science texts. The evaluation corpus and is marked up for the following error types - spelling errors, morpho-syntactic errors, syntactic errors, formatting errors, real word errors, etc. It consists of a publicly available corpus, *FREECORPUS* (34,512 words) and a part that is restricted by copyright *BOUNDCORPUS* (166,483 words). At this moment we use only *FREECORPUS* as we still need to correct and improve mark-up. For the final version of this paper we will include *BOUNDCORPUS*.

The results of the evaluation are shown in Table 2. The quality is measured using basic precision, recall and f_1 scores, such that recall $R = \frac{t_p}{t_p+f_n}$, precision $P = \frac{t_p}{t_p+f_p}$ and f_1 score as harmonic mean of the two: $F_1 = 2\frac{P \times R}{P+R}$, where t_p is a count of true positives, f_p false positives, t_n true negatives and f_n false negatives.

Adjective rules include attr>pred, pred>attr and attr>adv. Negation rules include finite>connegative, neg-present>neg-past.tense, second-person-past-connegative>person-agreeing-with-neg.

While precision for adjectives is decent (57%), negation rules fail in this test. The corpus only has 11 negation errors altogether, which makes the results very unreliable. 84 adjective form errors, on

the other hand, show that adjective errors are frequent and their correction is relevant for the language community. As a result of this evaluation, only adjective rules are included in the first release of the South Sámi grammar checker. For the final version of the paper, we plan to use the other part of the corpus, *BOUNDCORPUS*, as well for more testing material. It needs to be marked-up for grammatical errors of the type we are investigating. Previous versions did not include certain types of mark-up for the following reasons: 1) The norm was not clear at that point of time. 2) Manual mark-up is cumbersome, and not all error instances are easy to detect.

	Precision	**Recall**	**Positives**
Adjective rules	56.96%	51.72%	84
Negation rules	27.27%	23.07%	11

Table 2: Evaluation of the South Sámi grammar checker on *FREECORPUS*

When further investigating the reasons for the shortcomings of our tool we found the following: In ex. (6) attributive *guelhties* is erroneously corrected to predicative *guelhties*. The reason for that is that rules are missing a condition for possible coordination. This can easily be specified and corrected.

(6) Bovtside leah *guelhties* jih
reindeer.ILL be..3SG cool.ATTR and
gaaloes giesie hijven.
rainy.cool.ATTR summer is good.PRED
'For the reindeer, a cool and rainy summer is good.'

In ex. (7) even though the adjective *aelhkie* 'simple' preceeds a noun, it is not attributive. Instead, it is part of an infinitve construction of the type 'it is easy to do ...'. However, being an SOV language, in South Sámi the infinitive can be preceded by the object *ditnie-laejkiem* 'tin wire'.

(7) Ij leah *aelhkie*
be.3SG be.CONNEG easy.PRED
ditnie-laejkiem giesedh.
tin.wire.ACC pull
'it is not easy to pull a tin wire.'

This is a recurrent false positive type as can be seen in ex. (8), where predicative *vihkele* is erroneously corrected to attributive *vihkeles* since it is followed by a noun. However, this is an infinitive construction with an object before the infinitive just as in the previous example.

(8) lea **vihkele** saemiengïelem
be.3SG important.PRED Sámi language**.acc**
åtnose bertedh bievnese- jïh
use.ILL prepare information and
gaskesadteme teknologijesne
communication technology.ACC
'it is important to prepare the Sámi language for use in information and communication technology'

Another false positive caused by homonymy (adjective-verb) is the mark-up of the present participle verb *båetije* 'coming' as in ex. (9).

(9) *båetije* saemien siebredahken
coming Sámi.GEN society.GEN
dïejveldimmine
discussion.INE
'In future debates of the Sámi society.'

All three syntactic contexts can easily be included in error correction rules as exceptions.

As there are many (more) different types of negation error types, negation rule shortcomings are manyfold are the following. One issue, negation rules have not been paying attention to is homonymy between finite and a connegative forms like *lij* 's/he was' in ex. (10). The grammar checker tries to correct the form based on the assumption that it is a finite form. However, a negative condition excluding possible connegatives, should take care of this problem.

(10) Saemien siebredahken tseegkemisnie
Sámi.GEN society.GEN building.INE,
ij **lij** gaajhkide
NEG.PAST.3SG be.PAST.CONNEG all.ILL
saemientjïertide seamma nuepie
Sámi.groups.ILL same possibility
'In building the Sámi society, there were not the same opportunities for all Sámi groups'

5 Conclusion

In this article we present the first South Sámi grammar checker for adjective error correction, which is a very frequent error type due to language change and simplification of morphological richness. The loss of language arenas in a bilingual society and insufficient grammar teaching in schools have lead to a loss of the distinc-

tion between attributive and predicative forms and a simplification of the negation paradigm. The grammar checker is therefore meant as a tool to help language revitalization and support the wish of the language community to re-establish traditional morphological paradigms and teach them to future generations. While we get decent results for adjective form correction (Precision of 56%), which in addition are easy to improve in the future, negation correction is more complex and has more variants, which requires further testing. The grammar checker will be released in May 2023 and will be freely available for download to be used in text processing like Microsoft Word and Google Docs. Future plans include further the implementation and release of more error correction of other frequent error types, starting with the verbal negation paradigm.

Acknowledgments

References

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

Knut Bergsland. 1946. *Røros-lappisk grammatikk : et forsøk på strukturell språkbeskrivelse / av Knut Bergsland*. H. Aschehoug Co. (W. Nygaard) ; Cambridge, Mass. : Harvard University Press, Oslo.

Eckhard Bick and Tino Didriksen. 2015. CG-3 – beyond classical Constraint Grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa 2015)*, pages 31–39. Linköping University Electronic Press, Linköpings universitet.

Rogier Blokland and Cornelius Hasselblatt. 2003. *The endangered Uralic languages*. John Benjamins Publishing Company.

Rogier Blokland and Nobufumi Inaba. 2015. https://doi.org/10.1075/tsl.108.14blo *Negation in South Saami*, pages 377–398. Uppsala University University of Turku.

Tino Didriksen. 2010. http://visl.sdu.dk/cg3/vislcg3.pdf (Accessed 2017-11-29) *Constraint Grammar Manual: 3rd version of the CG formalism variant*. GrammarSoft ApS, Denmark.

Gustav Hasselbrink. 1981-1985. *Oårj'elsaamien baaguog'ärjaa, Skrifter utgivna genom Dialekt- och folkminnesarkivet i Uppsala Ser. C, Lapskt språk och lapsk kultur nr 4, Lundequistska bokhandeln*. Uppsala.

Magga Ole Henrik and Magga Lajla Mattsson. 2012. *Sørsamisk grammatikk*. Davvi girji, Kárásjohkka.

Maja Kappfjell and Trond Trosterud. forthcoming. Åarjelsaemien göökteli̇htse adjektijvi gramatihke. page .

Fred Karlsson. 1990a. Constraint Grammar as a Framework for Parsing Running Text. In *Proceedings of the 13th Conference on Computational Linguistics (COLING 1990)*, volume 3, pages 168–173, Helsinki, Finland. Association for Computational Linguistics.

Fred Karlsson. 1990b. Constraint grammar as a framework for parsing unrestricted text. In *Proceedings of the 13th International Conference of Computational Linguistics*, volume 3, pages 168–173, Helsinki.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi A Pirinen, and Miikka Silfverberg. 2013. Hfst—a system for creating nlp tools. In *International workshop on systems and frameworks for computational morphology*, pages 53–71. Springer.

Sjur N. Moshagen, Tommi A. Pirinen, and Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. In *NODALIDA*.

Tommi A. Pirinen and Krister Lindén. 2014. State-of-the-art in weighted finite-state spell-checking. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404*, CICLing 2014, pages 519–532, Berlin, Heidelberg. Springer-Verlag.

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

Trond Trosterud. 2019. Kva bruker vi minoritetsspråksordbøker til? ein studie av brukarloggane for tolv tospråklege ordbøker. *LexicoNordica*, pages 177–198.