Supporting Language Users – Releasing the first Lule Sámi Grammar Checker

Inga Lill Sigga Mikkelsen
UiT Norgga árktalaš universitehta
inga.l.mikkelsen@uit.no

Linda Wiechetek
UiT Norgga árktalaš universitehta
linda.wiechetek@uit.no

Abstract

We present the first rule-based L1 grammar checker for Lule Sámi. Releasing a Lule Sámi grammar checker has direct consequences for language revitalization. Our primary intention is therefore to support language users in their writing and their confidence to use the language. We release a version of the tool for MS Word and GoogleDocs that corrects six grammatical error types. For the benefit of the user, the selection of error types is based on frequency of the errors and the quality of our tool. Our most successful error correction, for a phonetically and syntactically motivated copula error, reaches a precision of 96%.

1 Introduction

We release a new L1 grammar correction tool for Lule Sámi that can be integrated in MS Word and GoogleDocs. *GramDivvun* is the first grammar checker for Lule Sámi and has been released May 31st 2023.[1] The underlying purpose is to provide a tool that can give language users the security that their language is right in the absence of a strict norm - a paradox we face in our daily work. Speakers and writers of a language are confident and carefree when they feel secure in their language use.[2] However, minority languages often face loss of language arenas and at the same time have less resources for language teaching than majority languages. The consequence is that (new) language users get insecure in their use of language and are often left to criticism by the language experts when speaking or writing. This can lead to frustration and resistance to use the language among the ones that are not considered language experts. The notion of the language barrier - where older generations take the role of the 'language police' - has also been reported in other indigenous language contexts, for example when learning the Cree language as an adult. (McCreery, 2006, pp.43) (Johansen, 2006)

As one of the authors of this work is herself a member of the Lule Sámi speech community she is familiar with general attitudes, one of which is that the ones that know the language have a clear feeling of how the language should be, even if there is not a written norm. This creates a gap between these experts and the language learners. At the same time there are few contexts/opportunities to improve one's grammar skills and avoid being criticised so that speaking can lead to anxiety and speakers can feel discouraged to use the language. Especially in writing, Lule Sámi text production differs from its coexisting majority language text production. Even official texts and texts written by highly proficient users contain a lot of spelling and grammar errors (Wiechetek et al., 2022). This is due to lower written language proficiency in minority languages, and also a lack of written norms.

A norm and someone enforcing this norm is necessary to teach language competence to the younger generation and pass on expert language knowledge in all its richness. In the absence of sufficient L1 teachers, now many L2 speakers are becoming teachers that need support to teach the language in all its details. There are no books that explain grammatical phenomena in all their details, including contrasting examples and frequent mistakes. Existing grammar books only have text book examples and focus on morphology, rather than syntax. Where sufficient human feedback on our language production is missing, we need a tool that can evaluate the correctness our language on the fly.

[1] https://divvun.no/en/korrektur/gramcheck.html

[2] "A positive attitude is also connected with creating a safe environment for learners." (McCreery, 2006)

Proceedings of the NoDaLiDa 2023 Workshop on Constraint Grammar - Methods, Tools and Applications, pages 37–45

Our language technology tools already have a wide user base including official domains such as the Sámi Parliament, Sámi media and schools that use our proofreading tools. The grammar checker will be included in the automatic updates of future versions of the spellchecker to provide better tools for the users. *Divvun* has been established to provide language technology tools for the Sámi language community, and has an ownership agreement with the Sámi Parliament, which unequivocally states that what Divvun develops belongs to the Sámi people through the Sámi Parliament.

The construction of a Lule Sámi grammar checker started in October 2020 with a general error categorization and smaller experiments with rules. In 2022 we did intensive work to collect regression tests and reported first results (Mikkelsen et al., 2022). The main motivation for making proofing tools are the needs of the language users and the tools' usability. That means that we want to make the tools available at an early stage, even if they do not include all the functionalities yet, and at the same time ensure their quality (i.e. especially good precision). Ensuring the quality means that only those error types that give a certain precision are included. The tools are meant to support teachers, proof readers and individuals by finding errors that are hard to detect because of orthographic similarities. They are also meant to help enforcing the (mostly orthographic) language norm proposed by the normative organ Giellagálldo[3] in a consistent way.

2 Language situation for Lule Sámi

Lule Sámi is an indigenous language spoken in Northern Norway and Sweden. The language is classified as a severely endangered language by UNESCO and has an estimated 800-3,000 speakers (Sammallahti, 1998; Kuoljok, 2002; Svonni, 2008; Rydving, 2013; Moseley, 2010). Lule Sámi is a morphologically complex language, for more details see Ylikoski (2022).

The current orthography of Lule Sámi was approved in 1983, and the first spell checker for the language was launched in 2007. Lule Sámi lacks a long written tradition. According to Kuoljok (1997) most of the speakers can barely read and even fewer write. This situation has changed since 1997. In the education system, Lule Sámi is taught and used as the language of instruction. In Norway, Lule Sámi was for the first time taught as first language in primary school in 1992, and from 2012 it was possible to take a bachelor's degree in Lule Sámi at Nord University. Lule Sámi is also to a greater extent used in public administration, in 2000 the Jåhkåmåhke/Jokkmokk municipality became one of the municipalities in Sweden with a Sámi-language administration and in 2006 the municipality Divtasvuona/Tysfjord was included in Norways Sámi-language administration municipalities. This development means that Lule Sámi is also used in writing to a greater extent than before. However, the written tradition is not very established, and the elderly heritage speakers master the written language only to a smaller extent.

In 2013, a Lule Sámi corpus of writing errors was created to test the spell checker's effectiveness. Today this corpus consist of 39,892 words, written by native Lule Sámi speakers, and it has all together 4,784 writing errors. 2,055 are non-word errors identified by the spell checker, while the remaining 2,729 errors are morpho-syntactic, syntactic and lexical errors that only a grammar checker can detect and correct (Wiechetek et al., 2022). The mark-up of this corpus shows an error rate of 11,9% in written texts, which indicates that Lule Sámi speakers struggle when writing the language.

To fully master a written language one must read a lot (Trosterud, 2021), minority language users therefore have a greater need for help in the writing process, since they do not experience their language in written form as much as majority language speakers. With Lule Sámi classified as a severely endangered language by UNESCO, it is important to increase the use of Lule Sámi to revitalize the language. A grammar checker for Lule Sámi would make it easier for people to write in the language, thus increasing its written use.

To develop a functional Lule Sámi grammar checker, we opted to focus on errors made by proficient speakers instead of second language learners. This approach allows us to create a grammar checker that can handle sentences with fewer errors and gradually introduce more complex errors. A grammar checker for texts written by second language learners would require a different approach as they tend to have more and different types of errors, including more complex errors.

Errors made by high proficiency speakers often arise when the written norm deviates from the spo-

[3] http://www.giella.org

ken dialectal variation or the "errors" might express an ongoing language change.

3 Technical background

All tools described in this article are part of a multilingual infrastructure for 130 languages (Moshagen et al., 2013).

Lule Sámi has a morphological analyser and lexicon, which are both publicly available[4]. The morphological analyser was originally imported with all rules and set specifications from North Sámi and then adapted to Lule Sámi.

GramDivvun 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 Wiechetek (2019). We are using finite-state morphology (Beesley and Karttunen, 2003) to model word formation processes. The technology behind our *FSTs* is described in Pirinen (2014). Constraint Grammar is a rule-based formalism for writing disambiguation and syntactic annotation grammars (Karlsson, 1990; Karlsson et al., 1995). In our work, we use the free open source implementation VISLCG-3 (Bick and Didriksen, 2015).

The challenge consists in writing rules that are as general as possible so one rule can cover many different erroneous forms at once. Most Lule Sámi grammatical errors can be referred to as a combination of morphological features that is confused with another combination, rather than a confusion pair of two lemmata as is typical for languages with less morphological complexity like English (e.g. *theirs–there's*). This allows for a higher degree of abstraction.

The syntactic context is specified in hand-written Constraint Grammar rules. The ADD-rule below adds an error tag (`&real-negSg3-negSg2`) to the negation auxiliary *ij* '(to) not (do)' as in example (1) if it is a 3rd person singular verb and to its left there is a 2nd person singular pronoun in the nominative case. The context condition further specifies a barrier for the rule to apply. Subjunctions, conjunctions, or finite verbs – typically indicating a new clause – stop the scanning of the rule.

[4] https://github.com/giellalt/lang-smj/

Each ADD-rule is accompanied by a COPY-rule that exchanges relevant morphological tags in order to produce the correct sequence for the FST morphological generator to generate the correct form. In this case *Sg3* is exchanged for *Sg2*. At the same time, we add a tag, *&SUGGEST* to mark that this is not the erroneous form anymore, but the correction.

(1) Dån **ittjij** boade guossáj.
you.2SG NEG.PAST.3SG come guest.ILL
'You didn't visit.'

```
ADD (&real-negSg3-negSg2) TARGET ("ij")
IF (0 (Sg3))
(*-1 (Pron Nom Sg2)
BARRIER S-BOUNDARY OR
CS OR CC OR VFIN) ;

COPY (Sg2 &SUGGEST) EXCEPT (Sg3)
TARGET  (&real-NegSg3-NegSg2) ;
```

4 Lule Sámi Grammar checker

4.1 Testset

Having a set of example sentences that show the natural context for a grammatical error is essential for the construction of a grammar checker. We want to correct errors that are actually made by users of the language.

We have collected sentences and made regression tests of representative errors in *Yaml*-formatted[5] files specific to each error type. (Wiechetek et al., 2021) Typically, each regression file contains several hundred sentences. Our standard has been to have yaml tests of at least 50 test sentences. There should be a balance of correct and erroneous sentences covering the same phenomena so that one can test for false positives and false negatives. Test sentences should cover a variety of syntactic contexts and pay attention to long-distance relationships between syntactic functions. The sentence collection is designed to cover a maximally large amount of real-world errors that people make when writing texts, in order to keep the grammar checker usable for people. The file naming is now error-specific,[6] but as they come from an authentic corpus, they can contain multiple errors per sentence including other types of errors and nested errors.

[5] https://yaml.org/spec/1.2/spec.html

[6] https://github.com/giellalt/lang-smj/tree/main/tools/grammarcheckers/tests

At first, we wrote test sentences for yaml tests ourselves and also searched SIKOR (SIKOR) manually for sentences with similar errors. After having written rules, we automatically harvested test sentences corrected by *GramDivvun* in the developer-corpus [7], and used these to improve the rules. At first, we wrote test sentences for yaml tests ourselves and also searched SIKOR manually for sentences with similar errors. After having written rules, we automatically harvested test sentences corrected by *GramDivvun* in the developer-corpus [8], and used these to improve the rules.

Yaml is a mark-up language with a simple syntax that makes writings of the tests convenient and co-operation with programmers and linguists easier [9]. We chose to use the Yaml format for grammar testing because of positive experiences with the use of the same format for spell checker testing.[10]

4.2 Grammar for error correction

It is challenging to write a prescriptive grammar checker for a language without a clear written norm. Even written grammar books of Lule Sámi do not cover all grammatical phenomena. Oral Lule Sámi contains a lot of dialectal variations and is subject to ongoing language change. As all speakers of Lule Sámi are bilingual, oral language can include interference and loans from the majority languages, which is not desired in a written norm. For all these reasons, it is a challenge to build a grammar checker that corrects this language. We face the question of where to put the boundaries between written and oral Lule Sámi. The decision can have serious consequences since Lule Sámi is an endangered language under revitalisation, and the grammar checker can have a standardising effect on the language of the younger generations. It is positive that speakers receive feedback when they write language that is clearly influenced by Norwegian or Swedish, but at the same time the grammar checker can also be thought to give feedback leading to a limitation of dialectal variation.

We do not have the authority to determine the norm, but with the release of the grammar checker, we might have the strongest influence regarding the sentence level norm in the entire Lule Sámi language community. One cannot wait until normative matters are solved before developing tools needed by the language community, the path must be created as we walk. The grammar checker will be further developed and improved after this first version release. Hopefully, the release of the Lule Sámi grammar checker will facilitate discussions around the norm and discussion around the choices made by us. Upon the release of the grammar checker, we had a presentations for the language community where we informed about the choices regarding the grammar checker and also discussed further development.

We have written 18 rule types, and from the evaluation six of these were ready to be released.

The words *oahpásmuvvat* and *oahpástuvvat* both meaning *to get to know* are often confused. The distinction lies in the animacy of what one is getting to know. *getting to know*. The verb *oahpásmuvvat*, in ex. (2) is used in inanimate contexts and requires illative case, whilst *oahpástuvvat*, in ex. (3) is used in animate context and require comitative case. The rules of the grammar checker corrects both verb according to animacy and the case of the referent.

(2) Oahpásmuváv bijllaj.
get.to.know.PRES.1SG car.SG.ILL
'I get to know the car.'

(3) Oahpástuváv sujna.
get.to.know.PRES.1SG PRON.2SG.COM
'I get to know her/him.'

The modal verb *soajttet* meaning 'maybe' should be paired with the infinitive form of the main verb. However, many writers are using the present singular third-person form *soajttá* as an adverb rather than a modal verb, as shown in ex. (4). In this example, the modal auxiliary is not followed by an infinitive as expected, but rather by a finite verb in the first-person singular form. The rules of the grammar checker will replace *soajttá* with the adverb *ihkap*. This correction is in line with the writer's intended adverb construction. An alternative to that would be inflecting *soajttá* according to

[7] https://giellalt.github.io/proof/gramcheck/extracting-precision-sentences.html

[8] https://giellalt.github.io/proof/gramcheck/extracting-precision-sentences.html

[9] The original test framework for morphology testing initiated by Brendan Molloy can be found on GitHub: https://github.com/apertium/apertium-tgl-ceb/blob/master/dev/verbs/HfstTester.py

[10] https://giellalt.uit.no/infra/infraremake/AddingMorphologicalTestData.html\#Yaml+tests

person and number of the subject and changing the following finite verb to an infinitive form. As this bears more risks in correction, especially when the subject is distant from the verb or dropped, we chose to replace the verb with an adverb.

(4) ***Soajttá** *tjálláv nágin
maybe.PRES.3SG write.PRES.1SG some
bágojt
word.SG.ACC
'Maybe I will write some words'

For agreement the grammar checker corrects relative pronouns in inessive case, as the incorrect ex. (5), and the reflexive pronouns *iesj* in nominative, as the incorrect ex. (6), when these do not agree with their anaphora in number. The grammar checker also corrects agreement errors between subject and verb, this is a quite common error done since indicative verbs are inflected for three numbers and three persons.

(5) Álu l má ålmmåjn ***gænna** l
often is PCLE man.PL.INE who.SG.INE have
fábmo
power
'Often it is men who have power'

(6) Mij hæhttup ***iesj**
we.NOM must.PRES.1PL self.REFL.SG.NOM
jáhkket
believe.
'We ourselves must believe.'

Another noun phrase internal error corrected by the grammar checker is the use of an attributive adjective in predicative position, as the incorrect ex. (7).

(7) Ássje l ***gássjelis** munji.
matter is difficult.ADJ.ATTR I.ILL
'The matter is difficult for me.'

For the copula verb *liehket* 'to be' the grammar checker has three different rule types following the system described in Spiik (1989). In sentence-initial position, the copulas have different forms from sentence internal forms, as shown for the present tense in Table 1. Even if this system is explained in (Spiik, 1989), the sentence internal forms are widely used sentence-initially in written texts, and the sentence initial 3. singular forms in both present and past tense are frequently used in sentence internal position. The sentence internal present 3. person singular form also varies between *la* or *l*: *la* is used if the preceding word ends on a consonant, and *l* is used if the preceding word ends on a vowel. Even though there most likely is and has been been dialectal variation in regarding the copula system, we have made rules according to Spiik (1989). We have even fine-tuned the rules for choosing between *la* or *l* since it really is not as straight forward as Spiik (1989) explains it. As developers we are not sure of how well copula correction will be received in the language community. The copula system of the grammar checker is not widely used in texts, for example, the translators of the Lule Sámi New Testament have chosen a different approach to the copula *liehket*. As the grammar checker allows users to turn off and on error types they want to have checked, they can turn certain corrections off, if they find them annoying.

Morphological form	**Sentence internal**	**Sentence initial**
1Sg	lav	lev
2Sg	la	le
3Sg	la/l	le
1Du	lin	len
2Du	lihppe	læhppe
3Du	libá	læbá
1Pl	lip	lep
2Pl	lihpit	lehpit
3Pl	li	le

Table 1: Paradigm for liehket 'to be'

5 Evaluation

For the evaluation of our tool, we use a part of *SIKOR*, the Lule Sámi corpus, containing administrative, law, religious, non-fiction, fiction, and science texts. SIKOR consists of a freely available corpus, *FREECORPUS*, and a corpus that is restricted by copyright, *BOUNDCORPUS*. We distinguish between three different parts: 1. the gold corpus for evaluation, marked-up for spelling and grammar errors, 2. the unmarked testing corpus and 3. the development corpus for developing rules. For simplicity, we will refer to the error marked-up gold corpus as *FREECORPUS* and *BOUNDCORPUS*. This work includes testing for inconsistencies and improvement of the manual grammar error mark-up the first time. Since the goldcorpus consists of text that has not been proof read, there are a lot of grammatical errors. The goldcorpus and its mark-up is described in

Wiechetek et al. (2022).

The testcorpus is not manually marked-up, but put aside for future evaluation and quality assurance as mark-up as the current goldcorpus is still fairly small, and needs enhancement to cover all different grammatical error types sufficiently. The development corpus on the other hand, is being used to test and improve the grammar checker rules on the fly. It is therefore not marked-up.

A preliminary evaluation on *BOUNDCORPUS* in Table 2 served to chose the error types to be included in the first version of *GramDivvun* and improve error mark-up in the gold corpus. Quality is measured using basic precision and recall, such that recall $R = \frac{t_p}{t_p+f_n}$, and precision $P = \frac{t_p}{t_p+f_p}$, where t_p is a count of true positives, f_p false positives, t_n true negatives and f_n false negatives.

	Precison	**Recall**	**# Err**
Copula forms	96.13%	83.71%	117
Rel agreement	72.22%	81.25%	17
soajttá as Adv	100.00%	100.00%	2
Refl agree	60.67%	33.33%	3
Animacy - Rel	33.33%	25.00%	3
oahpásmuvvat	100.00%	100.00%	1
Attr > Pred	0%	-	1
Pred > Attr	80.00%	40.00%	10
Subj-V agree	77.42%	25.53%	31
Num agree	60.00%	100.00%	10
Pass/Act	0%	0%	5

Table 2: Evaluation on *BOUNDCORPUS*

Table 2 shows that some error types have very few instances in *BOUNDCORPUS*. Some of this does not coincide with our manual proofreading experience and knowledge of frequent errors in written texts, and it may not reflect the real distribution of errors in a larger corpus either. Therefore, we use regression test results in Table 3 as a second criterion to select the error types for *GramDivvun*.

Based on the results of Tables 2 and 3, and keeping the quality assurance for the users in mind, we have released functionalities for errors regarding copula form and relative pronoun agreement, the second of which we reduced to errors regarding inessive case relative pronoun agreement. The first two error types have a good precision and perform well in regression testing. All of them have a precision above 70%. In addition, we have released error correction for error types with few instances

	PASS	**FAIL**
Copula form	122	7
Inessive rel number agreement	136	7
Modal verb *soajttá* as adverb	84	0
Refl number agreement	114	5
oahpásmuvvat-oahpástuvvat	63	1
Adjective form (Attr>Pred)	164	5
Subject-verb agreement	129	108
Past tense negation	46	8
Animacy of rel pronouns	140	63
Nominalization > finite verb	11	0
Adjective form (Pred>Attr)	55	17
Genitive before postposition	68	24
Nominative rel number agree	118	92
Numeral agreement	145	111

Table 3: Regression test results (for comparison)

in *BOUNDCORPUS*, which are based on good regression test results and knowledge about high frequency of the errors from experience as a manual proof reader. These error types are: adverbial use of the modal verb in third person singular, *soajttá* 'maybe s/he does'; use of attributive adjective forms instead of predicative forms; lexical confusion of the verbs *oahpásmuvvat*>*oahpástuvvat*; and reflexive pronoun errors. After fine-tuning the existing error mark-up on a bigger corpus that includes more fiction texts, and therefore other error types (*FREECORPUS*), we evaluated the well-performing rules on both *BOUNDCORPUS* and *FREECORPUS*, cf. Table 4.

Copula errors are by far the most frequent ones. In both corpora together, we found as many as 498 copula errors, four times as much as only in *BOUNDCORPUS*. All error types except for two have a precision above 85%. The low precision of reflexive and attributive > predicative adjective form confusion is not as low as it seems. In both cases, false positives are due to other errors in the text which lead to wrong corrections, but not detection. *GramDivvun* finds the error in the sentence, but fails to correct the error in the whole sentence structure based on other errors.

Altogether these are six general error types that have been released with functionalities in the first version of the Lule Sámi grammar checker.

Many of the rule types involve several rules. For example, copula correction includes three different rules: one for correcting from sentence initial to correct sentence internal forms, one for correct-

	Prec	Recall	# Err
Copula forms	92.77%	79.25%	498
Ine Rel agree	100.00%	71.43%	7
soajttá as Adv	91.67%	100.00%	11
Refl agree	50.00%	40.00%	10
oahpásmuvvat	85.71%	85.71%	7
Attr > Pred	50.00%	53.84%	13

Table 4: Evaluation on *FREECORPUS* and *BOUNDCORPUS*

ing the sentence internal form to the correct sentence initial form and one for choosing between the sentence internal forms *la* and *l*.

The benefit of our work has been twofold, we have improved both our tools and our marked-up data. Firstly, we have used rule development for automatic grammatical error detection, and secondly, we have improved grammatical error mark-up after running the grammar checker. This shows that consistency in manual error mark-up can be assisted by automatic grammar checking.

The evaluation shows despite good precision for the six rule types that were released, there are a number of false alarms and cases where *GramDivvun* does not find the error.

In ex. (8) and (9), the sentences are more complex than what we thought of when writing rules. In ex. (8) the grammar checker erroneously changes the attributive adjective *buosjes* 'tough' to predicative *buossje*. In this example there are two attributive adjectives connected with the conjunction *ja* meaning 'and'. Adding coordination conditions to the rules is fairly simple to fix.

(8) Adrian Nystø Mikkelsen gut aj la
Name who also is
buosjes ja vissjalis
tough.ADJ.ATTR and eager.ADJ.ATTR
bállotjiektje.
soccerplayer
'Adrian Nystø Mikkelsen who is a tough and eager soccer player.'

Another false alarm appears in ex. (9) where the subject is dropped and the grammar checker erroneously corrects the verb *vuojnáv* into 3.Pl since the 1.Sg pronoun *mån* 'I' is dropped.

(9) Hådjånav gå **vuojnáv** mijá
get upset.PRES.1SG when see.PRES.1SG our
galba biejsteduvvi.
signs destroy.PASSIVE.PRES.3PL
'I get upset when I see our signs being destroyed.'

Some of false alarms are due to combinations of errors. In ex. (10), *GramDivvun* erroneously changes the plural relative pronoun *ma* 'that' to singular *mij*. Therefore the subject is singular and the verb *guosski* 'regard' is also corrected by the grammar checker. Here *GramDivvun* changes *ma* to singular which is a false positive because of a wrong referent. Consequently it also tries to change the verb *guosski* to singular to correct the agreement with the relative pronoun.

(10) Lav válljim teoritevstajt
have.PRES.1SG choose.pst.ptcp text.PL.ACC
kompendijis **ma** **guosski**
compendium that.PL.NOM regard.PRES.3PL
álggoálmmukmetodologijav.
indigenous.methodology.ILL
'I have chosen texts from the compendium that regard indigenous methodology.'

We also have similar examples where the erroneous correction by the grammar checker is due to a combination of errors, but here it is the writer who has made two different errors. In ex. (11) the grammar checker corrects the attributive adjective *váges* 'reliable' to singular *váhke*, where it should have been corrected to plural *váge*. The writer has made two errors, one of which is a number error in the verb *viertti* 'must' (present 3.Sg) which should be present 3.Pl *vierttiji*. *GramDivvun* misses this subject-verb agreement error and therefore the adjective attribute form is corrected to predicative singular form. Adding an agreement error rule to *GramDivvun* will lead to a correction of the second error.

(11) Moralla subttsasin de máhttá liehket
moral story then might be
rádna ***viertti** liehket
friend.PL.NOM must.PRES.2SG be
***váges** nubbe nuppijn jus
honest.ADJ.ATTR each other if
rádnastallam galggá bissot.
friendship will remain.
'The moral of the story might be that friends need to be honest with each other if the friendship is to remain.'

The same problem with a combination of errors happens in ex. (12), where the writer has misspelled the indefinite pronoun *iehtjádijn* 'with another'. Because of the typo the grammar checker

erroneously corrects *oahpástuvvat* 'get to know' to *oahpásmuvvat*.

(12) Ietja dahki majt hálidi, ja
self do.PRES.3PL what want.PRES.3PL, and
dan båttå máhtá buorebut
that moment can.PRES.2SG better
*ietjadijn **oahpástuvvat**, javllá Inga Lill.
non.word get.to.know.INF, says Inga Lill
'Everyone does what they want, and at the same time you can get to know someone better, says Inga Lill.'

There are also examples where the rules of the grammar checker work fine, but where the grammar checker erroneously corrects because of problems with disambiguating homonymies. In ex. (13) the disambiguator construes *jage* 'year' to be nominative plural, when it actually is genitive singular. Because of the grammar checker construes *jage* to be the subject of the sentence it corrects the sentence-initial present copula form *le* 'is' to the 3Pl form *li* instead of the correct 3Sg form *la*.

(13) Badjel guoktalåk jage
Over twenty years
***le** duodje
be.PRES.3SG.SENT.INIT Sámi.handcraft
munji årrum vájmoássjen ja oasse iehtjam
me be heart.case and part my
identitehtas.
identity.
'For over twenty years Sámi handcraft has been close to my heart and a part of my identity.'

The evaluation shows that even though the grammar checker works well with six rules, there are still complex issues that cause the grammar checker to fail even for these types of errors. More errors in the same sentence makes it harder for the grammar checker. It is therefore important that the users know that this grammar checker is predominantly meant for L1 users and that upon its release, it does not work very well for second language learners' texts, yet. The evaluation shows that building a grammar checker for L1 users before L2 users is a good way to go, as the tool performs better with only one error in the sentence, and proficiency writers are assumed to make less errors.

6 Conclusion and future plans

We have released a tool for grammatical detection and correction of Lule Sámi (*GramDivvun*) to support the Lule Sámi language community in writing. We evaluated our tool and based on the evaluation, we chose six general error types that met our quality requirements and were ready to be released. These are corrections regarding copula forms, lexical confusion of *oahpásmuvvat-oahpástuvvat*, number agreement for reflexive pronouns, the use of the modal verb *soajttá* as an adverb, confusion of attributive and predicative adjective forms, and finally number agreement of inessive relative pronoun forms. While our evaluation corpus is still a bit too small to have a good representation of all errors, it was evident that copula errors are very frequent, and the other error types were also represented. Copula errors also show the best precision with 96% and recall of 84%. In other error types, we rely on our manual proof-reading experience to know about their frequency. This goes hand in hand with our wish to focus on user demands. In the future we will improve precision and recall for the correction of existing error types by testing on more syntactic contexts. This means we will we need to enhance the corpora with error mark-up. In addition we will improve the quality of error rules that have not been included in this version of *GramDivvun* with the goal of releasing them. We can also conclude that even L1 language users typically make several errors in a sentence. This is due to low literacy in Lule Sámi, and interference errors caused by bilingualism. Our focus must therefore be a tool that can handle these types of sentences.

References

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

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.

Inger Johansen. 2006. det er ikkje eit museumsspråk – det har noko med framtida å gjera ei sosiolingvistisk undersøking av revitaliseringa av sørsamisk. Master's thesis, Institutt for nordistikk og litteraturvitskap, NTNU.

Fred Karlsson. 1990. 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, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

Susanna Angéus Kuoljok. 1997. *Nominalavledningar på ahka i lulesamiskan*. Acta Universitatis Upsaliensis.

Susanna Angéus Kuoljok. 2002. Julevsámegiella. *Bårjås: Julevsámegiella uddni - ja idet?*, pages 10–18.

Dale McCreery. 2006. http://www.malsmb.ca/docs/challenges-and-solutions-in-adult-cree-learning.pdf Challenges and solutions in adult acquisition of cree as a second language. Master's thesis, BA, Canadian University College.

Inga Lill Sigga Mikkelsen, Linda Wiechetek, and Flammie A Pirinen. 2022. https://doi.org/10.18653/v1/2022.computel-1.19 Reusing a multi-lingual setup to bootstrap a grammar checker for a very low resource language without data. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics.

Christopher Moseley. 2010. www.unesco.org/culture/en/endangeredlanguages/atlas *Atlas of the World's Languages in Danger*, volume 3. UNESCO.

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.

Håkan Rydving. 2013. *Words and varieties : lexical variation in Saami*. Société Finno-Ougrienne.

Pekka Sammallahti. 1998. *The Saami Languages: an introduction*. Davvi girji.

SIKOR. UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection, Version 06.11.2018. `http://gtweb.uit.no/korp`. Accessed: 2018-11-06.

Nils Eric Spiik. 1989. *Lulesamisk grammatik*. Sameskolstyrelsen.

Mikael Svonni. 2008. Språksituationen för samerna i sverige. *Samiskan i Sverige, rapport från språkkampanjerådet*, pages 22–35.

Trond Trosterud. 2021. Utan tastatur, ingen tekst: om det språkteknologiske grunnlaget for språka våre. In Karin Kvarfordt Niia, editor, *Framgång för små språk.*, pages 68–73. Små språk i Norden.

Linda Wiechetek, Katri Hiovain-Asikainen, Inga Lill Sigga Mikkelsen, Sjur Moshagen, Flammie Pirinen, Trond Trosterud, and Børre Gaup. 2022. https://aclanthology.org/2022.lrec-1.125 Unmasking the myth of effortless big data - making an open source multi-lingual infrastructure and building language resources from scratch. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Linda Wiechetek, Sjur Nørstebø Moshagen, Børre Gaup, and Thomas Omma. 2019. Many shades of grammar checking – launching a constraint grammar tool for North Sámi. In *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar - Methods, Tools and Applications*, NEALT Proceedings Series 33:8, pages 35–44.

Linda Wiechetek, Flammie A Pirinen, Børre Gaup, and Thomas Omma. 2021. https://aclanthology.org/2021.iwclul-1.6 No more fumbling in the dark - quality assurance of high-level NLP tools in a multi-lingual infrastructure. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 47–56, Syktyvkar, Russia (Online). Association for Computational Linguistics.

Jussi Ylikoski. 2022. Lule Saami. *The Oxford Guide to the Uralic Languages*, pages 130–146.