Many shades of grammar checking – Launching a Constraint Grammar tool for North Sámi

Linda Wiechetek linda.wiechetek@uit.no
Sjur Moshagen sjur.n.moshagen@uit.no
Børre Gaup borre.gaup@uit.no
Thomas Omma thomas.omma@uit.no

Divvun UiT Norgga árktalaš universitehta

1 Introduction

This paper discusses the characteristics and evaluation of the very first North Sámi spell- and grammar checker. At its launch it supports *LibreOffice*, *MS Word* and *GoogleDocs*. We describe its component parts, the technology used such as *Constraint Grammar* (Karlsson, 1990; Karlsson et al., 1995; Bick and Didriksen, 2015) and *Hfst-pmatch* (Hardwick et al., 2015) and its evaluation with a new evaluation tool specifically designed for that purpose.

Only the modules of the full-scale grammar checker described in Wiechetek (2017) that have been tested and improved sufficiently for the public to use are released in this launch. More advanced syntactic errors will be included at a later stage. The grammar checker modules are an enhancement to an existing North Sámi spellchecker (Gaup et al., 2006), following a philosophy of *release early, release often*, and using continuous integration (*ci*) and continuous delivery (*cd*) to deliver updates with new error correction types, as new parts of the grammar checker are sufficiently tested. Releasing at an early stage of development gives the user community early access to improved and much needed tools and allows the developers to improve the tools based on the community's feedback.

North Sámi is a Uralic language spoken in Norway, Sweden and Finland by approximately 25 700 speakers (Simons and Fennig, 2018). These countries have other majority languages, making all Sámi speakers bilingual. Bilingual users frequently face bigger challenges regarding literacy in the lesser used language than in the majority language due to reduced access to language arenas (Outakoski, 2013; Lindgren et al., 2016). Therefore, tools that support literacy, like spelling and grammar checkers, are more important in a minority language community than in a majority language community.

The released grammar checker is a light grammar checker in the sense that it detects and corrects errors that do not require rearranging the whole sentence, but typically just one or several adjacent word forms based on a grammatical analysis of the sentence. Additionally, a number of formatting errors are covered. The main new features are implemented by means of several Constraint Grammar-based modules. These include correction of formatting and punctuation errors, filtering of speller suggestions, much improved tokenisation and sentence boundary detection, as well as advanced compound error analysis and correction.

This paper shows how a finite-state based spellchecker can be upgraded to a much more powerful spelling and grammar checking tool by adding several Constraint Grammar modules.

2 Framework

An open-source spelling checker for North Sámi has been freely distributed since 2007, the beginnings of which have been described by Gaup et al. (2006). The tool discussed in this paper includes the open-source spelling checker referenced above, but further developed and using the hfst-based spelling mechanism described in Pirinen and Lindén (2014). The spelling checker is enhanced with five Constraint Grammar modules, cf. Figure 1. It should be noted that the spelling checker is exactly the same as the regular North Sámi spelling checker used by the language community, but with the added functionality that all suggestions are returned to the pipeline with their full morphological analysis. The analyses are then used by subsequent Constraint Grammar rules during disambiguation and suggestion filtering.

All components are compiled and built using the *Giella* infrastructure (Moshagen et al., 2013). Five constraint grammar modules, i.e. a valency grammar (*valency.cg3*), a tokenizer (*mwe-*

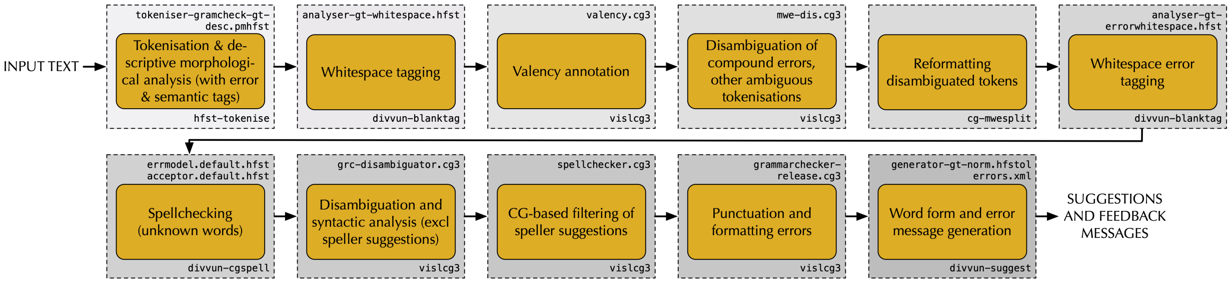
INPUT TEXT
tokeniser-gramcheck-gt-desc.pmhfst
Tokenisation & descriptive morphological analysis (with error & semantic tags)
hfst-tokenise
analyser-gt-whitespace.hfst
Whitespace tagging
divvun-blanktag
valency.cg3
Valency annotation
vislcg3
mwe-dis.cg3
Disambiguation of compound errors, other ambiguous tokenisations
vislcg3
Reformatting disambiguated tokens
cg-mwesplit
analyser-gt-errorwhitespace.hfst
Whitespace error tagging
divvun-blanktag
errmodel.default.hfst
acceptor.default.hfst
Spellchecking (unknown words)
divvun-cgspell
grc-disambiguator.cg3
Disambiguation and syntactic analysis (excl speller suggestions)
vislcg3
spellchecker.cg3
CG-based filtering of speller suggestions
vislcg3
grammarchecker-release.cg3
Punctuation and formatting errors
vislcg3
generator-gt-norm.hfstol
errors.xml
Word form and error message generation
divvun-suggest
SUGGESTIONS AND FEEDBACK MESSAGES

Figure 1: System architecture of the North Sámi grammar checker

dis.cg3) and a morpho-syntactic disambiguator (*grc-disambiguator.cg3*), a disambiguation module for spellchecker suggestions (*spellchecker.cg3*) and a module for more advanced grammar checking (*grammarchecker-release.cg3*) are included in the spelling and grammar checker.

The current order of the modules has shown to be the most optimal one for our use and has been established during the work with the grammar checker. It follows the principle of growing complexity, and information necessary to subsequent modules is made available to them. Valencies for example are used in the disambiguation of compounds, which is why the module preceeds the multi-word disambiguation module. The first whitespace tagging module preceeds other modules because it is used to inserts hints about the text structure, such as `<firstWordOfParagraph>`. Such hints must be available early on. The second whitespace analyser is applied after the multiword disambiguation, but could be applied later. Its purpose is to tag potentially wrong use of whitespace, and must be added before the final grammar checking module, but any position between the multiword disambiguation and grammar checking is going to work. Spellchecking is performed before disambiguation so that more sentence context is available to the disambiguator.

It should be noted that we do not use a guesser for out-of-vocabulary words. A major part of new words are formed using productive morphology of the language, like compounding and derivation, both of which are encoded in the morphological analyser. Also, the lexicon is constantly being updated, so that proper nouns and other potential out-of-vocabulary words will quickly be covered. As updates are made available frequently, this should not be a major issue for users.

The infrastructure is not only valid for North Sámi, but can directly be used by any language in the *Giella* system, e.g. other Sámi languages as well as any other language. Presently there is an early version of a working Faroese grammar checker in addition to the North Sámi one, and initial work has started for a number of the other Sámi languages.

The system does error detection at four different stages of the pipeline. Non-word typos are marked by means of the spellchecker. Secondly, a Constraint Grammar module marks whitespace errors in punctuation contexts based on input from the second whitespace analyser. Compound errors are identified by means of a Constraint Grammar-based tokenisation disambiguation file. And a fourth Constraint Grammar module marks quotation errors.

As a tool intended to be used in production by regular users, it targets all types of errors, from technical typesetting errors such as wrong quotation marks and faulty use of spaces, via spelling errors to advanced grammatical error detection and correction. In the evaluation all of these are counted as grammar checker errors, as we want to evaluate the overall performance of the tool. The only errors not included in the evaluation are those that we do not target at all (which are quite a few in this first beta release).

The sentence in (1) includes a typo (*Norgag* should be *Norgga* ‘Norway’ (Gen.)), a space error (before ‘.’) and a compound error (*iskkadan bargguin* should be *iskkadanbargguin*) ‘survey’ (Loc. Pl.). In addition, there is a congruence error, i.e. *dáid* (Gen. Pl.) ‘these’ should be *dáin* (Loc. Pl.), i.e. it should agree in case and number with *iskkadan bargguin*. The launched grammar checker can detect the first three errors, but not the last one, since syntactic error rules are not included in this initial launch of the grammar checker.

(1) Oktiibuot 13 **Norgag** doaktára leat
altogether 13 Norway.GEN doctor.GEN have
leamaš mielde dáid
been with these.GEN
iskkadan bargguin_.
testing work.LOC.PL
'Altogether 13 Norwegian doctors have participated in these surveys.'

The whitespace analyser detects an erroneous space before '.' The suggested correction is *"<iskkadan bargguin.>"*. The tokenisation disambiguation module detects the compound error and suggests *iskkadanbargu+N+Sg+Com = iskkadanbargguin*. The tokeniser is used to disambiguate between syntactically related n-grams and misspelled compounds, where the misspelling is an erroneous space at the word boundaries.

This module is clearly checking more than spelling conventions, i.e. grammar, as writing, for example, two consecutive nouns as one or two words has syntactic implications. Such noun-noun combinations do not necessarily need to be compounds even if the first element is in nominative case. They can also be syntactically related as in the agent construction in ex. (2) where *addin* 'giving' is a (nominalized) non-finite verb that modifies the second noun, *vuođđu* 'basis'.

This grammar checker finds compound errors by distinguishing between syntactic readings as the previous one and compound readings as *addinvejolašvuođa* 'giving possibility (Gen.)'.

(2) luossabivdu lea lunddolaš
salmon.fishing is natural
Golf-rávnnji **addin**
Golf-stream.GEN give.ACTIO.NOM
vejolašvuođa vuođđu
possibility.GEN basis
'salmon fishing is a natural resource for a possibility given by the Golf-stream'

3 Evaluation (planned)

Previous evaluations of our most advanced error type correction have given good precision (76.6%) and recall (78.6%) (Wiechetek et al., 2019). The development of the grammar checker evaluation tool presented in this paper has been informed by the previous evaluation, and many errors in the infrastructure and in the Constraint Grammar modules have been corrected. Therefore higher precision and recall are expected in this evaluation.

As a reference, when Bick (2015) evaluates his full-fledged grammar checker *DanProof*, the results for correcting both, spelling and compounding errors are the following: precision is 90.8% and recall is 86.8%.

As opposed to the evaluation done in Wiechetek et al. (2019), this evaluation will be automatic and based on an evaluation corpus. This is meaningful as the error corrections intended in this launch typically include a limited amount of adjacent word forms, which is relatively straightforward. Reference-less approaches as proposed in Napoles et al. (2016) are not an option as they require preexisting and independently developed tools that are sensitive to grammatical errors, a luxury not available to most minority languages.

A part of the North Sámi *SIKOR* corpus (SIKOR2016) containing error mark-up for orthographical and grammatical errors is used as the evaluation corpus. It consists of 181 512 words. *SIKOR* is split in one part that contains publicly available texts, and one part that contains texts that cannot be redistributed.[1] The genres represented in the evaluation corpus are news, blogs, and teaching materials.[2]

We evaluate the output of the launched grammar checker tool against a reference text that is manually marked for spelling errors (only non-words), compound errors, space errors and punctuation errors (only quotation marks for now), i.e. only the error types we actually try to correct. The performance of the grammar checker is measured by means of precision and recall. Good precision has typically priority over good recall as users tend to react more critically to flagging correct input as errors as opposed to not flagging error input.

4 Conclusion

This paper has presented the first released spelling and grammar checker tool for North Sámi that, in addition to spelling errors, checks and corrects punctuation errors, compound errors and white space errors. All error detection and correction, including context-aware filtering of spelling suggestions, is performed by Constraint Grammar modules at all stages of the process. The automatic evaluation of the launched grammar checker tool is expected to give good results based on recent

[1] https://giellalt.uit.no/ling/corpus_repositories.html (Retrieved 2019-06-19)

[2] https://giellalt.uit.no/proof/nordplus/StavekontrolltestingOgNorplusprosjektet.html (Retrieved 2019-06-19)

previous evaluations.

Additionally this work also describes the full infrastructure for a full-scale grammar checker and facilitates the implementation of any kind of grammatical error correction as soon as these are considered to be working well enough to be released. The infrastructure is available for any language within the *Giella* framework.

As the infrastructure is mainly ready, the plan is to include a wide range of real-word errors and syntactic error detection in the next release.

References

Eckhard Bick. 2015. DanProof: Pedagogical spell and grammar checking for Danish. In *Proceedings of the 10th International Conference Recent Advances in Natural Language Processing (RANLP 2015)*, pages 55–62, Hissar, Bulgaria. INCOMA Ltd.

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.

Børre Gaup, Sjur Moshagen, Thomas Omma, Maaren Palismaa, Tomi Pieski, and Trond Trosterud. 2006. From Xerox to Aspell: A first prototype of a north sámi speller based on twol technology. In *Finite-State Methods and Natural Language Processing*, pages 306–307, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sam Hardwick, Miikka Silfverberg, and Krister Lindén. 2015. http://aclweb.org/anthology/W/W15/W15-1842.pdf Extracting semantic frames using hfst-pmatch. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, (NoDaLiDa 2015)*, pages 305–308.

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.

Eva Lindgren, Kirk P H Sullivan, Hanna Outakoski, and Asbjørg Westum. 2016. Researching literacy development in the globalised North: studying trilingual children's english writing in Finnish, Norwegian and Swedish Sápmi. In David R. Cole and Christine Woodrow, editors, *Super Dimensions in Globalisation and Education*, Cultural Studies and Transdiciplinarity in Education, pages 55–68. Springer, Singapore.

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

Courtney Napoles, Keisuke Sakaguchi, and Joel R. Tetreault. 2016. http://aclweb.org/anthology/D/D16/D16-1228.pdf There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2109–2115.

Hanna Outakoski. 2013. Davvisámegielat čálamáhtu konteaksta [The context of North Sámi literacy]. *Sámi dieđalaš áigečála*, 1/2015:29–59.

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.

SIKOR2016. 2016-12-08. SIKOR UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection. **URL:** *http://gtweb.uit.no/korp (Accessed 2016-12-08).*

Gary F. Simons and Charles D. Fennig, editors. 2018. http://www.ethnologue.com (Accessed 2018-10-09) *Ethnologue: Languages of the World*, twenty-first edition. SIL International, Dallas, Texas.

Linda Wiechetek. 2017. *When grammar can't be trusted – Valency and semantic categories in North Sámi syntactic analysis and error detection.* PhD thesis, UiT The Arctic University of Norway.

Linda Wiechetek, Kevin Brubeck Unhammer, and Sjur Nørstebø Moshagen. 2019. https://www.aclweb.org/anthology/W19-6007 Seeing more than whitespace – Tokenisation and disambiguation in a North Sámi grammar checker. In *Proceedings of the third Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 46–55.