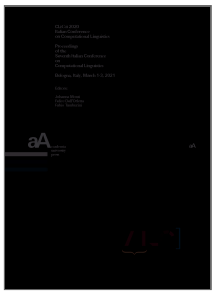

Felice Dell'Orletta, Johanna Monti and Fabio Tamburini (dir.)

Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020
Bologna, Italy, March 1-3, 2021

Accademia University Press

Suoidne-varra-bleahkka-mála-bihkka-senet-dielku 'hay-blood-ink-paint-tar-mustard-stain' – Should compounds be lexicalized in NLP?

Linda Wiechetek, Chiara Argese, Tommi A. Pirinen and Trond Trosterud

DOI: 10.4000/books.aaccademia.8979
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2020
Published on OpenEdition Books: September 3, 2021
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic EAN: 9791280136336

Suoidne-varra-bleahkka-mála-bihkka-senet-dielku ‘hay-blood-ink-paint-tar-mustard-stain’ – Should compounds be lexicalized in NLP?

Linda Wiechetek linda.wiechetek@uit.no
Chiara Argese chiara.argese@uit.no
Tommi A Pirinen tommi.pirinen@uit.no
Trond Trosterud trond.trosterud@uit.no

Divvun & Giellatekno, UiT Norgga árktalaš universitehta

Abstract

English.

Lexicalizing compounds, in addition to treating them dynamically, is a key element in giving us idiomatic translations and detecting compound errors. We present and evaluate an e-dictionary (*NDS*) and a grammar checker (*GramDivvun*) for North Sámi. We achieve a coverage of 98% for *NDS*-queries and of 96% for compound error detection in *GramDivvun*.

Italiano.

La lessicalizzazione delle parole composte, in aggiunta a trattarle in maniera dinamica, è un elemento chiave per ottenere traduzioni idiomatiche e rilevare errori nelle stesse. Presentiamo e valutiamo un e-dizionario (NDS) e un correttore grammaticale (GramDivvun) per il Sami del Nord. Otteniamo una copertura del 98% per le ricerche in NDS e del 96% per il rilevamento di errori nelle parole composte in GramDivvun.

1 Introduction

In this paper[1], we discuss the use and necessity of the lexicalization of compounds – in addition to the dynamic approach to compounding – in two rule-based Natural Language Processing (NLP) applications, a grammar checker *GramDivvun* and an electronic dictionary *NDS* (short for *Neahttadigisánit*). We argue for a dual approach and support this view with an evaluation of these tools. For comparison, we also look at a third application, a corpus tool (*Korp*) for the North Sámi corpus *SIKOR*. SIKOR, the Sámi International KORpus, is the collection of texts in different Sámi languages compiled by UiT The Arctic University of Norway and the Norwegian Sámi Parliament.

In the past, we have mostly focussed on the dynamic approach to morphological analysis. This means that we have a lexicon with lemmata and stems, which in a finite-state manner are combined with inflectional and derivational affixes and other stems and modified when morpho-phonological processes apply. In this way the linguistic processes inflection, derivation and compounding are modelled in a dynamic way, i.e. by means of concatenation and composition as opposed to listing of all forms. Lexicalization, i.e. listing compounds or inflected word forms as such, is the alternative approach to the dynamic one. In addition to these two approaches we also use guessers for certain tasks, i.e. proper name guessing in morpho-syntactic parsing. Our approach is entirely rule-based and open source. Within our 20 year experience with language tools for the Sámi languages and other languages with complex morphology, we have achieved good results and produced reliable tools.

There are a number of approaches to error detection of a few errortypes for morphologically complex - although less complex than North Sámi - languages like Latvian (Deksne, 2019) and Russian (Rozovskaya and Roth, 2019). The Latvian neural network grammar checker focusses on preposition-postposition confusion, adjective-noun agreement, mood errors in verb forms, number and case in noun forms, definiteness of adjectives and missing commata. All of these error types have a good performance with precisions between 78% and 98.5%. Judging from their regular expressions to insert artificial errors, most of their error types seem to be fairly local errors that can be resolved based on bigrams.

The Russian system focusses on more advanced error types - case, number agreement, gender agreement, preposition and aspect. However, the results show that the system is still in its initial phase with low precision and recall for most error types (precision is between 22% and 56%, only gender agreement reaches 68%, and recall is significantly lower, between 9% and 36%). None of these approaches deals with compound error de-

tection.

For neural network approaches, large corpora with error mark-up are necessary, which are not available for North Sámi. The error marked-up corpus contains 120 459 words, and when looking at specific error types – as in this case compound errors – the corpus is even smaller. The Russian system is based on an error-marked corpus of 200k words (deemed too small by its authors), the Latvian system works with artificial errors, an approach that can be problematic as it does not reflect real text errors.

In compounding, two or several words are combined to form a new word. In Sámi, Finnic and Germanic languages, compounding is a productive process and new compounds like in (1) can be made on the fly.[2] In Romance languages, these compounds typically correspond to prepositional constructions (ital. 'la federa del cuscino del divano').[3]

(1) soffá|guoddá|olggoža (North Sámi)
sofa|pute|trekk (Norwegian)
'sofa pillow cover (English)'

The initial motivation for extensive lexicalization of compounds of North Sámi goes back to adapting the spellchecker to users' needs, i.e. avoiding false alarms in *Ávvir* newspaper's texts.

North Sámi is a Uralic language spoken in Norway, Sweden and Finland by approximately 25 700 speakers (Simons and Fennig, 2018). It is a synthetic language, where the open parts of speech (PoS) – nouns, adjectives, etc. – inflect for case, person and number. The grammatical categories are expressed by a combination of suffixes and stem-internal processes affecting root vowels and consonants alike, making it perhaps the most fusional of all Uralic languages. In addition to compounding, inflection and derivation are common morphological processes in North Sámi.

North Sámi has seven morpho-syntactic cases, i.e. nominative (Nom.), genitive (Gen.), accusative (Acc.), illative (Ill.), locative (Loc.), comitative (Com.), and essive (Ess.). Case plays a more central role in Sámi than in preposition-based case languages, since here syntactic functions are identified based on case only. In addition, nouns can bear possessive suffixes. Verbs are inflected for person, number (singular, dual, plural), tense (present and past tense) and mood (indicative, conditional, and potential). Derivational processes (passive, causative, inchoative, diminutive, reflexive, to name only some of them) enhance the combinatory possibilities of each verb.

Table 1 illustrates that compounding in North Sámi is by no means restricted to noun noun combinations, but includes a number of other *parts-of-speech* (PoS) as well, also as heads.[4]

Type	**Example**	**Gloss and translation**
N N	láhka|rievdadusat	law|change.PL 'law changes'
A.Attr N	boahtte|áigi	coming|time 'future'
Adv N	dáppe|olmmoš	here|person 'person from here'
Pron A	iešguđet|lágan	each|alike 'different kinds of'
Pron N	eanet|lohku	more|number 'majority'
Adv Pcle	dušše|fal	only|really 'just'
Adv V	vuostái|váldojuvvo	against|take.PASS.3SG 'received'
PrfPrc N	mearridan|fápmu	decide.PRFPRC|power 'authority'
Num Num	okta|nuppe|lohkái	one|second|ten.ILL 'eleven'
Num N	1978|-láhka	1978|-law '1978 law'
Num A	3|-ivnnat	3|-colored '3-colored'
Num A	golmma|ivnnat	three|colored 'three colored'

Table 1: Compound types according to PoS; '|' is used to mark word boundaries

In North Sámi, compounds are formed without a hyphen, except for those involving a proper noun, a digit, or an acronym like *Davvi-Norgii* 'Northern Norway (Ill.)', *3-juvllatsykkel* 'tricycle', and *ILO-álgoálbmotsoahpamuš* 'ILO-indigenous people agreement' (Riektačállinrávvagat, 2015, p.46). There are a number of multiwords where a space is obligatory (*albma ládje* 'properly' and *duollet dálle* 'sometimes'). Also genitive first compounds have an alternative interpretation when written apart, which makes error detection more difficult.

2 Background

The North Sámi tools described in this article – *NDS*, *Korp* for *SIKOR* and *GramDivvun* (Wiechetek, 2012) – all rely on the *Giel-*

[2]To avoid confusion with hyphenated compounds, "|" is used to mark word boundaries in compounds

[3]Although there are a number of real compounds in Italian, such as *fruttivendolo*, as well.

[4]The following abbreviations are used: N=noun, V=verb, A=adjective, Attr=attributive, Adv=adverb, Pron=pronoun, Pcle=particle, PrfPrc=past participle, Num=numeral, Prop=propernoun.

laLT infrastructure (Moshagen et al., 2013), a technological framework for managing lexical data and building it into language technology applications including e-dictionaries and grammar checkers. All of them make use of a morphological analyzer, an *FST* (Finite-State Transducer) described in Pirinen (2014), where word formation processes are moduled. Additionally, *SIKOR* and *GramDivvun* include a Constraint Grammar-based syntactic analysis. The full modular structure of the latter is described in Wiechetek (2019b).

The computational modeling of the language is done using finite-state morphology (Beesley and Karttunen, 2003). The method of recognizing grammatical words as well as querying their grammatical information is based on looking up the words in an FST that contains the morphological dictionary of the language. There are two types of compounds in the language model: the ones that are stored in the lexicon as lexicalized units and the ones generated dynamically using a compounding model. Table 2 gives the statistics over the length of lexicalized compounds.[5]

Lexicalized four-element compounds are quite common in the noun lexicon, e.g. *davvisámegielterminologiija* 'North Sámi language terminology'. Even six-element compounds (*sáivačáhceguollevuostáiváldindilli* 'fresh water fish receive situation') can be found.

The different types of North Sámi compounds in Table 1 are not treated equally in the morphological analyzer. Only the compounds in the first two lines can be derived dynamically. All others need to be lexicalized, i.e. listed in the lexicon, to receive a compound analysis. Numeral compounding is not treated dynamically in the FST. The dynamic compounds are generated from the dictionary by concatenating word forms (such as a genitive or nominative noun followed by other noun) and adding a compound tag `+Cmp`. The main dynamic compounds are (derived and non-derived) noun + noun pairs. One feature of the underlying technology is that the compounding mechanism is capable of modeling infinitely long compounds: for example nouns of any magnitude are compounds and modeled by the finite-state automaton. Since the compounding mechanism of an FST is very powerful, it also leads to ambiguity. When we allow arbitrary lexemes to combine to form compounds, some will overlap other existing lexemes, cf. ex. (2).

(2) Davvi **regiuvdna**
North region;direction.oven
'The northern region'

Here, *regiuvdna* 'region' has a typical spelling error, o>u. The FST analyzes it as a misspelling of *regiovdna* 'region', but also as a compound with the elements *regi*, a common wrong form of *regiija* 'direction', and *uvdna* 'oven'. While this example has only two possible analyses, twenty or more different analyses are not uncommon.

Roots / PoS	2	3	4	5	6+
N	16 603	1 048	1 665	86	15
Num	408	1 048	42	0	4
Prop	11 680	3 005	115	9	1
A	3 854	333	13	0	0
V	478	4	0	0	9
Adv	896	109	1	0	0
Adp	152	49	0	0	0
Conj	3	0	0	0	0

Table 2: Lexical compounds in the lexicon by the PoS of their head and the number of their roots

3 Compounds in three NLP applications

We present three applications, an e-dictionary, a corpus tool, and a grammar checker tool.

3.1 An e-dictionary (NDS)

The North Sámi – Norwegian dictionary contains 25 000 lemmata and uses an FST. The e-dictionary was first implemented in 2013 with no use of relational databases (all linguistic resources are contained within static files and external command-line tools) (Ryan Johnson, 2013). It is an intelligent dictionary in the sense that is able to look up North Sámi word forms and find lemmas via the FST. It also allows a tolerant mode, which accepts the letters *acdnstz* for *áčđŋšŧž* in addition to their usual values. The e-dictionary can split compounds to provide the user with its elements as well as the whole compound if a translation is available. The lexicalization of compounds is important since the translation of the compound cannot necessarily be derived from the translation of its parts (Antonsen, 2018, p.54).

[5]The table is based on the dictionary size at the time of the writing (September 2020); it is actively developed daily. Further abbreviations are Adp=adposition, Conj=conjunction.

In the FST 90% of the 100 000 nouns, and in the dictionary 75% of the 25 000 nouns are compounds.

3.2 A corpus tool

The web application and corpus search tool *Korp* (Borin et al., 2012) does not show the internal structure of compounds in *SIKOR*. Neither lexicalized, nor dynamic compounds are searchable as either the lexicalized analysis is picked instead of the dynamic one or – in the case of compounds that are not listed in the lexicon – a lexicalized compound is made by the preprocessor. This is a problem inherent in the implementation of the tool. However, when searching for the compound tag used in the FST (`+Cmp`), there are 94 658 results. The reason for that is that the first element in split compounds in coordination receives a specific compound tag (`+Cmp/SplitR`) as well.

Table 3 shows the statistics for compounds in *SIKOR*.[6] The results are obtained using the scripts that can be found in *GiellaLT*.[7] According to our analyses 8.6% of the tokens in corpus are compounds, and 86% are lexicalized. The rest is mainly composed of 2-elements compounds (13.4%) and a very small part of 4-7 elements (0.5%).

Many of the longer compounds in *SIKOR* are quite creative and are hyphenated as the one in ex. (3).

(3)
suoidne-varra-bleahkka-mála-bihkka-senet-dielku
hay-blood-ink-paint-tar-mustard-stain
mu báiddis lei dušše lihkohisvuohta.
my shirt.LOC was only mishap
'The hay-blood-ink-paint-tar-mustard-stain on my shirt was only a mishap.'

Parts / PoS	2	3	4	5	6/7
N	96.2	98.9	89.2	80	66.7
Prop	3.8	1.1	10.8	20	33.3

Table 3: Compound types in SIKOR by the PoS of their head and the number of their root (amounts given in percentage)

The current public version of the Sámi corpus *SIKOR* (SIKOR, 2018) (in Korp) consists of 32.2 million words. It was analyzed with a preprocessor that does not distinguish between lexicalized and dynamic compounds. The (non-public) version of SIKOR used in this article makes this distinction, though, as will future versions in Korp.

A search for compound tags only returns split compounds, i.e. the first coordinated hyphenated nominal element, cf. in ex. (4), i.e. *riddo-* 'coast-'.

(4) **riddo-** ja vuotnaguovlluin
coast- and fjordregion.LOC.PL
'in coastal and fjord regions'

GiellaLT has already produced a solution, i.e. a tag for cohorts with a dynamic compound (`<with-dynamic-compound>`) added by a Constraint Grammar module. However, this tag does not provide any information about the number of elements and the beginning and ending of each element.

3.3 A grammar checker (GramDivvun)

GramDivvun, the North Sámi grammar checker (Wiechetek et al., 2019b) takes input from the FST to a number of other modules, the core of which are several Constraint Grammar modules. 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). All components are compiled and built using the *GiellaLT* infrastructure (Moshagen et al., 2013).

Lexicalization of compounds is relevant for grammar checking within compound error detection. One common error that cannot be resolved by a spellchecker is the spelling of compounds as two or more words. *GramDivvun* performs this type of error detection as part of the tokenization. The tokenization is done in two steps. In the first step potential compounds are tokenized ambiguously (either as one or as two words, the first of which is accompanied by an errortag). In the second step, a Constraint Grammar module[8] selects or removes the error reading. Two conditions need to be met to find the compound error: 1. the compound needs to be lexicalized, and 2. the syntactic context needs to support the compound reading.

The syntactic context is specified in hand-written Constraint Grammar rules. The

[6] The search was done on 2020-09-07.

[7] `https://github.com/giellalt/conf-clicit2021`

[8] `https://github.com/giellalt/lang-sme/blob/3a43911929458fd39da309ed23178bf5dbd04bcd/tools/tokenisers/mwe-dis.cg3`

REMOVE-rule below removes the compound error reading (identified by the tag Err/SpaceCmp) if the head is a 3rd person singular verb (cf. l.2) and the first element of the potential compound is a noun in nominative case (cf. l.3). The context condition further specifies that there should be a finite verb (VFIN) somewhere in the sentence (cf. l.4) for the rule to apply.

```
1 REMOVE (Err/SpaceCmp)
2 (0/0 (V Sg3))
3 (0/1 (N Sg Nom))
4 (*0 VFIN);
```

All possible compounds written apart are considered to be errors by default, unless the lexicon specifies a two or several word compound or a syntactic rule removes the error reading. There are numerous syntactic contexts where the potential parts of compounds make perfectly sense. In the case of noun-noun compounds, the second element can for example be a simple adverbial, as in ex. (5). The second element can be homonymous with another PoS, it can be a finite verb or an infinitive.

(5) son lea boarráseamus **mánná joavkkus**.
s/he is oldest child group.LOC
's/he is the oldest child in the group.'

4 Evaluation

We evaluate the e-dictionary (coverage) and the grammar checker (precision, recall) for compounding (errors). The corpus search tool does not exhibit compounding information and is therefore not evaluated.

4.1 An e-dictionary (NDS)

We analyzed the logs for NDS (*Neahttadigisánit*) for 2019, and found that 12.6% of the types in the user queries are compounds. The results are obtained using the scripts that can be found in *GiellaLT*[7]. The amount of lexicalized compounds in the logs (72.1%) is approximately the same as in the dictionary, where it is 75% (cf. Section 3.1 above). As much as 98% of the compound queries get a translation, either a lexicalized one or of its parts. Thus dynamic compounding contributes with a substantial improvement to dictionary coverage. If the alternatives are "getting no help from the dictionary" and "getting help to translate the parts" then the latter is to be preferred, even though the correct translation would be different from just joining the parts. For example, the compound word *ruhtahearrá* 'rich man' is not lexicalized in NDS but it does get a translation of its parts *ruhta* 'money' and *hearrá* 'man', which can help the user to understand the meaning of the compound word itself.

Most of the non lexicalized compounds are composed of 2 elements (96% in the logs and 93% in the entries). When analyzing the entries in the dictionary, we found that 24.8% are compounds and of those 97.6% are lexicalized. Table 4 shows PoS for compounds in NDS logs and entries.

	Logs				**Entries**			
Parts / PoS	L	2	3	4	L	2	3	4
N	90	87	85	100	86	87	82	0
A	3	0	0	0	2	0	0	0
Prop	3	0	0	0	12	4	0	0
V	2	13	14	0	0	8	18	0
Adv	1	0	0	0	0	0	0	0

Table 4: Compounds according to the number of their parts and PoS in NDS logs and entries (L=lexicalized)

4.2 A grammar checker (GramDivvun)

We evaluate error detection for syntactic compound errors (i.e. words that are written apart and should be a compound) in *GramDivvun* in two ways. Firstly, we compare last year's results in Wiechetek (2019a) with a newer version of *GramDivvun*, from now on referred to as the *Nodalida*-corpus. Last year's results are based on version *r183544* (Wiechetek et al., 2019a)[9]. The new results are based on version r28510[10] of *GramDivvun*.

However, as the focus in the last analysis was a different one, i.e. we evaluated other error types as well, we ran a second evaluation on a 2 363 word-corpus[11] specifically made to test compound error detection, i.e. every sentence contains a potential compound. These sentences are hand-selected from *SIKOR*.

The results of the evaluation are presented in Table 5. We can see that precision has gone significantly up, i.e. the average precision is 95.5%.

[9] https://github.com/giellalt/lang-sme/releases/tag/nodalida-2018 on 2019-09-26

[10] https://github.com/giellalt/lang-sme/releases/tag/clicit on 2020-09-07

[11] http://gtsvn.uit.no/freecorpus/orig/sme/odda_mahppa/compounds.correct.txt

However, the recall has gone down to average 46%. We are investigating the reasons for that. But in general, a high precision is desirable in grammar checking, even at the cost of a lower recall.

The results of the evaluation of *GramDivvun* compound grammar checking are shown in Table 5.

Measure	(2019)	(2020)	
	Nodalida corpus		**Compound corpus**
Precision	75.0%	93.1%	98.0%
Recall	72.9%	43.2%	48.5%
F1-Score	73.9	59.0	64.9
TP	51	54	50
FP	17	4	1
FN	19	67	53

Table 5: Measures for GramDivvun (TP/FP= true/-false positives, FN=false negatives)

False negatives are typically due to the lack of lexicalization. Many of those are proper noun combinations which are very productive, e.g. *Murmánska-aviisa* 'Murmansk newspaper', *Várggát-festiválas* 'at the Várggát festival', *km-galba* 'km sign' and *Divttasvuotna-regiovnna* 'Divttasvuotna region'.

Other reasons are certain (unlikely) analyses of especially the first element, e.g. that generally suggest a syntactic construction rather than a compound as in ex. (6). Here the first element *duorastat* 'Thursday' has a finite verb reading as well.

(6) dán **duorastat veaiggi**.
this.GEN Thursday twilight.GEN
'this Thursday evening'

The false positive is due to an error in the recognition of the span of the target. In ex. (7), *lulli sámi guvlui* is concatenated, but it should only be *lulli sámi*.

(7) dohko **lulli sámi** guvlui.
thither South Sámi area.ILL
'thither towards the South Sámi area.'

5 Conclusion

We have shown that the lexicalization of compounds – in addition to their dynamic treatment – is useful and necessary for two language applications for North Sámi, an e-dictionary (*NDS*) and a grammar checker (*GramDivvun*). The evaluation of *NDS* shows that we get a good coverage: 98% of the compounds logged do get a translation and 72% are lexicalized in the FST. The evaluation of *GramDivvun* has shown that we manage to identify compound errors with a precision of 98% and a recall of 49% utilising a combination of information from the lexicon and syntax.

We conclude that there are perfectly good reasons for lexicalizing compounds, i.e. providing idiomatic translations for when it cannot be derived from the parts, and to support compound grammar checking. At the same time, lexicalization can dissimulate word formation information in corpus tools. This can be resolved and we have already implemented a solution in Constraint Grammar to make the information available in a future version of the corpus tool. As dynamic compounding is limited to few PoS at the moment, in the future we want to investigate and model compounding of other PoS (in the FST). Also experiments with neural network approaches and a comparison of the results to our rule-based grammar checker could be an interesting future project.

Acknowledgments

Thank you to Thomas Omma for doing the error corpus mark-up and for fun linguistic discussions, and to Lene Antonsen for digging in our corpus and helping to find just the right example.

References

Lene Antonsen. 2018. *Sámegielaid modelleren – huksen ja heiveheapmi duohta giellamáilbmái. [Modeling Saami languages. Construction and adaptation to real-world linguistic issues]*. Ph.D. thesis, UiT The Arctic University of Norway, Tromsø.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford.

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

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA).

Daiga Deksne. 2019. Bidirectional lstm tagger for latvian grammatical error detection. In *Ekštein K. (eds) Text, Speech, and Dialogue. TSD 2019. Lecture Notes in Computer Science, vol 11697. Springer*.

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

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

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.

Riektačállinrávvagat. 2015. Riektačállinrávvagat. Sámedikki giellaossodat/Sámedikki oahpahusossodat, Guovdageaidnu.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of russian. In *Transactions of the Association for Computational Linguistics, vol. 7, pp. 1–17, 2019*.

Trond Trosterud Ryan Johnson, Lene Antonsen. 2013. Using finite state transducers for making efficient reading comprehension dictionaries. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa 2013)*, Proceedings Series 16: 59–71.

SIKOR. 2018. SIKOR uit norgga árktalaš universitehta ja norgga sámedikki sámi teakstačoakkáldat, veršuvdna 06.11.2018. `http://gtweb.uit.no/korp`. Accessed: 2018-11-06.

Gary F. Simons and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, twenty-first edition.

Linda Wiechetek, Kevin Brubeck Unhammer, and Sjur Nørstebø Moshagen. 2019a. 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.

Linda Wiechetek, Sjur Nørstebø Moshagen, Børre Gaup, and Thomas Omma. 2019b. 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. 2012. Constraint Grammar based correction of grammatical errors for North Sámi. In G. De Pauw, G-M de Schryver, M.L. Forcada, K. Sarasola, F.M. Tyers, and P.W. Wagacha, editors, *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8/AFLAT 2012)*, pages 35–40, Istanbul, Turkey, may. European Language Resources Association (ELRA).