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, Trond Trosterud

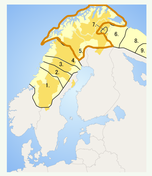
Abstract

Adding lexicalized compounds to a rule-based system of dynamic compounding provides improved handling of idiomatic translations and compound error detection. We present and evaluate an e-dictionary (*NDS*) and a grammar checker (*GramDivvun*) for North Sámi. We achieve a coverage of 98% for *dictionary*-queries and 96% for compound error detection in *grammar checking*.

Keywords: grammar checking, e-dictionary, compounding, tokenization, rule-based methods, complex morphology, Sámi languages.

Introduction

North Sámi is a morphologically complex language with 25,700 speakers in Norway, Sweden and Finland.



Wikimedia Commons

Motivation for lexicalizing compounds: adapt the spell-checker to users' needs and avoid false alarms.

Our rule-based open source approach builds on 20 year experience with tools for 117 languages with complex morphology at *GiellaLT* infrastructure.

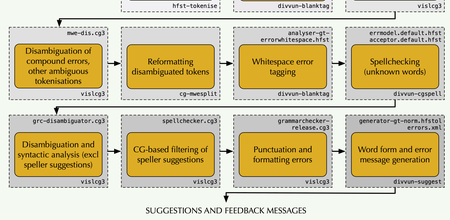
An e-dictionary

Our e-dictionary (*NDS*) is a combined electronic dictionary (25,000 lemmata) and grammatical model. This means that the dictionary:

- processes compound, derived and inflected forms and looks up the lemma(s) in the dictionary (far from trivial for a morphologically complex language like North Sámi).
- allows a tolerant mode which accepts the letters *acd* for *áčđ* etc.
- can split compounds to provide the user with its elements as well as the whole compound if a translation is available.

A grammar checker

GramDivvun, the North Sámi grammar checker takes input from the FST to a number of (Constraint Grammar) modules.



GramDivvun detects compound errors as a part of two-step tokenization:

1. potential compounds are tokenized both as one word (=compound error) and two words
2. Constraint Grammar rules select/remove the error reading under two conditions

(a) the compound needs to be lexicalized
(b) the syntactic context needs to support the compound reading

Compounding in North Sámi

Most compounds are noun-noun combinations and written as one word.
We found compounds of up to 6 elements in our dictionary.

(1) davvisámegielterminologiija
'North Sámi language terminology'.

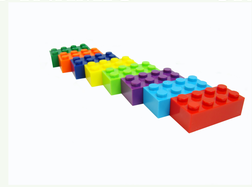
(2) Davvi-Norga
'Northern Norway'

(3) 3-juvllatsykkel
'tricycle'

(4) ILO-álgoálbmotsoahpamuš
'ILO-indigenous people agreement'

Compounds are modelled by *finite state transducers* (FSTs) dynamically or they are lexicalized.

DYNAMIC OR **LEXICALIZED**?



Fernando Barozza/Mostphotos.com

Boggy/Mostphotos.com

A - DYNAMIC:
adding any noun to any other noun (in particular morphological forms)

B - LEXICALIZED:
listing valid combinations in the dictionary

Evaluation

	(2019)	(2020)	
	Large corpus		**Small corpus**
Precision	75.0%	93.1%	98.0%
Recall	72.9%	43.2%	48.5%
F1-Score	73.9	59.0	64.9

Grammar checker:

- We compare previous evaluations (2019) with current results and evaluate a larger and a smaller (more specific) corpus
- Precision has improved significantly, but recall has gone down
- Recent results have shown an improvement of recall
- We manage to identify compound errors utilizing a combination of grammar rules and information from the lexicon

E-dictionary:

- *NDS* coverage: 98% of the compound queries logged in usage logs (N=939112) do get a translation and 72% are lexicalized in the FST
- we analyzed the logs for *NDS* (*Neahttadigisánit*) for 2019, and found that 12.6% of the types in the user queries are compounds

	User logs 2019				Dict. entries			
# elem	Lex	2	3	4	Lex	2	3	4
Noun	90	87	85	100	86	87	82	0
Adj	3	0	0	0	2	0	0	0
Prop	3	0	0	0	12	4	0	0
Verb	2	13	14	0	0	8	18	0
Adv	1	0	0	0	0	0	0	0

Corpus tool

The current North Sámi searchable corpus SIKOR (http://gtweb.uit.no/korp) contains 32 million words and many compounds:

(5) **suoidne-varra-bleahkka-mála-bihkka-senet-dielku**
hay-blood-ink-paint-tar-mustard-stain

Conclusion & Future Research

Reasons for lexicalizing compounds in the FST:

- it provides idiomatic translations when derivation from the parts is not possible
- it supports compound grammar checking

Downsides:

- it dissimulates compounding in corpus tools

Future plans:

- implement compound searchability in the corpus tool
- experiment with neural network approaches & compare with our rule-based grammar checker

Contact Information

Web divvun.no and http://giellatekno.uit.no/index.eng.html
Email feedback@divvun.no
Original article http://ceur-ws.org/Vol-2769/paper_49.pdf
GitHub https://github.com/giellalt