One grammar – different syntaxes

Lene Antonsen
UiT The arctic university of Norway
lene.antonsen@uit.no

Abstract

This paper presents a solution to the problem of providing a good syntactic analysis, within one grammar, for texts from different genres. The environment is a machine translation system which is used, among other things, to translate newspaper texts and transcriptions of oral language.

1 Introduction

This paper presents a way of adapting a syntactic disambiguator implemented within Constraint Grammar framework (CG) (Karlsson, 1990), to different kind of text genres. The morphological and syntactic analysers are described in Antonsen et al. (2010) and Antonsen and Trosterud (2017). The grammar is among other things used within a machine translation system.

2 Different rules for different genres

In running text in North Saami almost all sentences contain a finite verb. The rules in the CG grammar looks for the finite verb, and then the main non-finite verbform, if the finite verb is an auxiliary verb. The next step is disambiguating the case of the nouns, based on the valency of the verb. For this it is crucial to find the verb.

The CG grammar is integrated in a machine translation (MT) system from North Saami to Norwegian. The MT system is implemented with Apertium (Forcada et al., 2011), which is a modular set of tools for building rule-based MT systems. The pipeline consists of the following modules[1]:

1. Deformatting (encapsulating formatting/markup from the engine)
2. Morphological analysis of the source language by means of a Finite-State Transducer (FST)
3. Disambiguation and syntactic analysis with Constraint Grammar
4. Lexical transfer (word translation of the disambiguated source)
5. Lexical selection with Constraint Grammar (choice of contextually appropriate lemma)
6. One or more levels of FST-based structural transfer (reordering and changes to morphological features)
7. Generation of target language by means of FST

2.1 Newspaper headlines

The MT program is integrated in the webpages of the Saami University of Applies Sciences, but the program is located on the internet and can be used on all types of websites. It is possible to translate the webpage for the only newspaper published in North Saami. These webpages contain headlines, very often without a finite verb. The headlines can contain only a non-finite verb, or there is no verb at all.

I have examined a small corpus of 1539 headlines from the newspaper Ávvir, and 38% of the headlines don't contain a finite verb. These headlines are presented in table 1.

Almost half of the sentences are fragments with a head noun in nominative and no verbal, like example (1).

(1) – Viimmat midjiide
– viimmat.Adv mii.Pers.Ill
deaivvadanbáiki.
deaivvadanbáiki.N.Nom
– Finally for.us a.meeting.place

[1] It has the same setup as in Antonsen et al. (2017).

Type of sentences without finite verb	**Amount**	**%**
Head noun in nominative	283	49%
Subject + predicative	44	8%
Infinite verbform:		
Perfect participle	152	26%
Actio essive	16	3%
Infinitive	22	4%
Other	64	11%
Sum	581	100%

Table 1: Headlines without a finite verb.

Before disambiguation the adverb *viimmat* also gets analysis as a Sg2 form of the verb *viibmat* 'to eager'.

8% of the sentences have subject and predicative (in nominative or essive), but the copulas is omitted. In example (2) the noun *lokten* could also be analysed as a Sg1 form of the verb *loktet* 'to lift'.

(2) Vearu lokten viehka
vearru.N.Gen lokten.N.Nom viehka.Adv
vealtameahttun.
vealtameahttun.A.Nom
Tax rise quite unnecessary

33% of the headlines contain a non-finite verb form. Sentences with only perfect participle are perhaps the biggest challenge because this form very often gets the analysis as indicative Sg1 in the present tense as well, as in example (3), where *nuoskkidan* also is a Sg1 form of *nuoskkidit* 'to pollute'.

(3) Čábbámus báikki
čáppat.A.Sup báiki.N.Acc
nuoskkidan.
nuoskidit.V.PrfPrc
The.most.beautiful place polluted.

An examination of the sentences reveals that Sg1 form follows the personal pronoun *mun* or a dash to mark direct speech. The CG rules should therefore choose perfect participle analysis to Sg1 analysis in all other cases. This is a genre dependent feature, because North Saami is a pro-drop language, and the personal pronoun is usually not necessary in such cases.

Sentences with actio essive (equivalent to progressive in English) is usually not so problematic, because this form is never homonymous to any finite verb analysis, but the word can get morphological analysis as a noun in essive, a derived form of the same verb, like the word *gazzamin* in example (4).

(4) Sámedikkis oahpu
Sámediggi.N.Loc oahpu.N.Acc
gazzamin.
gazzat.V.ActioEss
At.the.Saami.parliament education
eating.with.spoon

In example (5) is an adjective followed by a verb in infinitive. All infinitives also get morphological analysis as finite verbs, in this sentence *dieđihit* is analysed also as Pl1, Pl3 and Sg2 before disambiguation.

(5) Čuorbbit dieđihit
čuorbbit.A.Pl.Nom dieđihit.V.Inf
luosaid.
luossa.N.Pl.Acc
Clumsy to.report salmon

There are several frequent adverbs, like *dušše* 'only', which also gets analysis as a finite verb (Prt Pl3 of *duššat* 'to perish'), and also frequent nouns, like *skuvlii* (illative) 'to the school', which also is analysed as Prt Sg3 of *skuvlet* 'to teach'.

Like the examples show, the headlines need special treatment in the CG analyser. A solution is to mark the sentences. Before translation, a *html deformatter* is run on the webpage. The task of this program is to hide html tags [inside brackets] so the linguistic parts of the pipeline can ignore them. Additionally, when it sees an end-of-heading tag like *</h1>* (up to *</h6>*) it will output the symbol ¶ immediately before the hidden html tag, into the text itself. So the input *<h1>Headliner</h1>* turns into *[<h1>]Headliner[]¶[][</h1>]* (with some empty brackets as separators, so we e.g. avoid treating ¶ as part of the preceding word).

Then this is sent to the morphological analyser, where ¶ gets an analysis like other words, letting CG rules match it: LINK *1 ("¶") ;

After CG, the mark is translated into an empty string.

2.2 Oral language

A corpus of transcriptions of oral language is available on internet, LIA Sápmi[2]. The North Saami part contains 174,000 words. In the LIA Sápmi user interface each sentence can be sent to the machine translation system.

The transcriptions consist of 22,153 segments, which the analyser treat as a sentence, and 21% of them have only one or two words, see table 2. That means that there is little context for disambiguation of homonymous forms. 78% of the sentences with one word consist of an interjection, adverb, particle or conjunction. The other one-word-sentences mostly consist of a noun.

Words in sentences	**Amount**	**%**
One word	3557	16%
Two words	1208	5%
Three words	1025	5%
Four words	1398	6%
Five words or more	14,965	68%
Sum	22,153	100%

Table 2: The sentences in LIA Sápmi corpus.

Unlike the headlines, omitting auxiliary verb and copula is unlikely. An exception is the expression *leat leamaš* 'have been', in which the auxiliary verb is often omitted in oral language. There are no abbreviations in the transcriptions.

In the newpaper headlines Sg1 verbforms follow the personal pronoun *mun* or a dash to mark direct speech. In the LIA-transcriptions there are no dashes, and 24% of the sentences with finite verb inflected for Sg1, are pro-drop sentences.

To give the LIA-transcriptions a special treatment in the CG grammar, they are marked with a ¥ in the end of the sentence in the LIA interface before they are sent to the MT system. The sentence is added to the morphological analyser, where ¥ gets an analysis like other words, letting CG rules match it. After CG, it is translated into the empty string.

An example of safe rule for the LIA material is removing analysis with abbreviation:
REMOVE ABBR IF *1 ("¥") ;

[2]LIA (Language Infrastructure made Accessible) is a collaboration project between four universities (UiO, UiB, UiT and NTNU), the Norwegian Dictionary 2014 and the National Library.

3 Conclusion

An additional mark to sentences from genres with syntax that differs from regular running text, enables grammar to use customized rules to provide a proper analysis of the sentences. In a machine translation system, this solution provides a more accurate translation of texts from different genres. An evaluation of the rules for newspaper headlines and transcriptions of oral material remains to be done.

Acknowledgments

Thanks to Kevin Unhammer for implementing the solution in the html deformatter module in the MT-system, and to Trond Trosterud who is especially working on the modul generating the Norwegian translation. Behind the North Saami analyser is work done by several people in the Giellatekno- and Divvun-groups at UiT The arctic university of Norway.

References

Lene Antonsen, Ciprian Gerstenberger, Maja Kappfjell, Sandra Nystø Rahka, Marja-Liisa Olthuis, Trond Trosterud, and Francis M. Tyers. 2017. Machine translation with North Saami as a pivot language. In *Proceedings of the 21st Nordic Conference of Computational Linguistics (NoDaLiDa 2017)*, volume 29 of *NEALT Proceedings Series*, pages 123–131, Linköping, Sweden. Linköping University Electronic Press. http://www.ep.liu.se/ecp/131/015/ecp17131015.pdf (14.02.2018).

Lene Antonsen and Trond Trosterud. 2017. Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk. *Norsk lingvistisk tidsskrift*, 35(1):153–185. http://ojs.novus.no/index.php/NLT/article/view/1416 (05.08.2019).

Lene Antonsen, Trond Trosterud, and Linda Wiechetek. 2010. Reusing grammatical resources for new languages. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2782–2789, Stroudsburg. The Association for Computational Linguistics, ELRA. http://www.lrec-conf.org/proceedings/lrec2010/pdf/254/_Paper.pdf (14.02.2018).

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas,

Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *Papers Presented to the 13th International Conference on Computational Linguistics (COLING-90) on the Occasion of the 25th Anniversary of COLING and the 350th Anniversary of Helsinki University*, volume 3, Helsinki. Yliopistopaino. https://dl.acm.org/citation.cfm?id=991176 (14.02.2018).