Some problems related to the development of a grammar checker

Kristin Hagen, Janne Bondi Johannessen and Pia Lane
University of Oslo

{kristiha, jannebj, pial}@mail.hf.uio.no

1. Introduction

Developing a grammar checker presents problems to the linguist and to linguistic software that are far from trivial. In this paper we will discuss various kinds of problems that we have encountered during our work with developing a grammar checker for Norwegian (Bokmål). The grammar checker in question, developed in cooperation with Lingsoft OY and to be used in Microsoft Word from the summer 2001, will be applied on text that has been grammatically tagged and disambiguated by the Constraint Grammar method (Karlsson et al 1995, Hagen and Johannessen 1998, and Hagen, Johannessen and Nøklestad 2000a, 2000b). The fact that the input text has been multi-tagged and disambiguated before grammar checking has certain advantages, but also presents certain problems, as we shall see.

Below, we shall give a brief review of statistical measures and an outline of the structure of the grammar checker, and then give an overview of the error types that our grammar checker is meant to cover. Many people find it hard to believe that grammar errors occur to any large extent. We shall therefore show some authentic examples from manually proofread tests. The errors can be categorised in two main groups: 1) Errors caused by a sloppy use of the "cut and paste" options of modern text editors, and 2) errors due to people's mistaken beliefs about the written language norms.

We shall then focus on problems that present a challenge in more ways than the obvious ones that deal with recognising mistakes and suggesting improvements. First, we shall focus on mistakes that have caused us having to rewrite the tagger that is used for the input text (undo disambiguation and redo it differently). Second, we shall look at mistakes that are caused by the fact that certain non-standard inflectional forms are homographous with other words found in the lexicon, and suggest a way of dealing with those. Third, we shall see what kinds of problems occur when words are missing, and hence not tagged at all. Finally, we shall show briefly why it is important for a grammar checker to include a component of grammatical analysis.

2. Statistical measures and other grammar checkers

The Norwegian grammar checker has been developed using Constraint Grammar rules. This method has previously been used to develop a Swedish grammar checker (see Birn 2000, Arppe 2000). The resulting precision for this grammar checker is reported to be 70% (according to Birn 2000:38), counted as good alarms divided by the sum of good alarms and false alarms. The resulting precision for the Norwegian grammar checker is 75 % from a test corpus of 890 000 words from the newspapers *Nordlys* and Sarbsborg Blad. However, depending on what kind of text that is used as a test corpus and what kind of rules that are included, these numbers can vary a great deal. For example, if we include a rule that tests whether there are any finite verbs in a sentence, and apply the grammar checker on newspaper text with a lot of verb-less headlines, the precision rises to approximately 91 % (and the same goes for the Swedish grammar checker).

3. The structure of the grammar checker

The grammar checker is composed of two main components: a tagger (consisting of three subparts) and an error detector:

(1) *The main structure of the grammar checker:*

I Tagger

•A preprocessor that among other things splits the text into sentences, identifies abbreviations and fixed expressions and marks capitalised unknown words in non-sentence initial position as proper nouns.
•A multi-tagger that uses NORTWOL (Karlsson et al 1995) to give a morphological analysis. Each word form is given all the possible tags that are appropriate for it.
•A morphological disambiguator, which is a modified version of the disambiguator made by *Taggerprosjektet* (Hagen and Johannessen 1998, and Hagen, Johannessen and Nøklestad 2000a, 2000b).

II Error detector that identifies different kinds of grammatical errors (see below).

•For each error that is identified, the user gets a short error message and if possible a suggestion of how to correct the error. The user is also given the possibility of reading a longer help text.

4 Error types

The Norwegian grammar checker is designed to detect the following main error types:

- Noun phrase internal agreement:
  - Definiteness
    (*et huset --> et hus*)
  - Gender agreement
    (*en nytt hus --> et nytt hus*)
  - Number agreement
    (*et grønt epler --> et grønt eple*)

- Two adjacent adjectives without a comma or conjunction
  en rød rask bil --> en rød, rask bil/en rød og rask bil
- Subject complement agreement
  Bilen er rødt --> Bilen er rød
- *Ingen* and *noen* and negation
  Jeg kjøpte ikke ingen bok i bokhandelen --> Jeg kjøpte ikke noen bok i bokhandelen
  Jeg kjøper noen bok i bokhandelen --> Jeg kjøper ei bok i bokhandelen
- *og/å* errors
  De gikk å sang --> De gikk og sang
  Hun skal å vise meg den nye kjolen --> Hun skal vise meg den nye kjolen
  Den lille gutten kan både snakke å synge --> Den lille gutten kan både snakke og synge
  Jeg trenger og sove --> Jeg trenger å sove
  Han skal prøve og skrive korrekt --> Han skal prøve å skrive korrekt
  Han orket sykle til jobben. --> Han orket å sykle til jobben.
- Too many finite verbs in a sentence or no verb in the sentence at all
  De kan forsøker å hjelpe deg --> De kan forsøke å hjelpe deg
  I Norge er var det slik --> I Norge er det slik.

Den gamle mannen syk. --> Den gamle mannen er syk

- Word order errors

  Jeg går ikke ut hvis det slutter ikke å regne --> Jeg går ikke ut hvis det ikke slutter å regne.

  Nå gutten kommer --> Nå kommer gutten.

5 Some errors from authentic texts

Below we will show a few of the errors we have found when testing the Norwegian grammar checker. They are all found in published newspaper texts from *Nordlys*, *Stavanger Aftenblad* and *Sarpsborg Blad*.

(2)

(a) De første kampen spilles torsdag.

(b) Som leder av Kommunalstyret for kultur, idrett og kirke (kik) har han tatt initiativ til at alle kultursøknader som sendes til staten, underskrives av de fire kommunen sammen.

(c) Det har vært en del tyverier på Sølvberget, men aldri tidligere har noen brukt skrujern for å hente en gjenstand i et monter på en utstilling. --> en monter

(d)Vi merker oss at for eksempel Illiosfestivalen i Harstad fikk økt sitt budsjett fra 95.000 kroner i fjor til 250.000 kroner i år, at Nordland Musikkfestuke fikk ei økning på 90.000 (...) --> en økning

6. Challenging problems

6.1 Errors that have caused us to rewrite the tagger

Usually, a disambiguating tagger can assume that the text to be disambiguated is correct with respect to spelling and grammar. A tagger that is to be used with a grammar checker, however, has to assume that the text may contain errors. Therefore, a lot of rules in the disambiguating part of our original tagger had to be rewritten or discarded. Let us take one example:

In Norwegian, *den* can both be a determiner, as in (3), or a pronoun, as in (4):

(3) *Den* bilen likte han godt
(4) *Den* likte han godt

In the original disambiguator there are CG rules based on the knowledge that a Norwegian determiner agrees in gender and number with the rest of the noun phrase of which it is a part. These rules state that *den* is a pronoun if

- there is no masculine determiner immediately to the right (*den neste bilen*)
- there is no masculine adjective immediately to the right (*den røde bilen)*
- there is no masculine noun immediately to the right (*den bilen)*

In all other cases the word *den* would be a pronoun since the above rules have made it clear that it is not part of a noun phrase. I.e., if there is no agreement, the pronoun reading is chosen.

But when we want the grammar checker to find errors like *den eplet* or *den bilene*, i.e., possible noun phrases with erroneous gender and number agreement, we cannot discard the determiner

reading. We therefore had to remove the original pronoun rules from the disambiguator, leading to the resulting text being more ambiguous than before.

6.2 Errors due to non-standard inflections and spellings being homographous with other words found in the lexicon

The spell checker handles ordinary misspellings and is run before the grammar checker. But what happens if a word is wrongly spelled or inflected and at the same time homographous with other words or inflections found in the lexicon?

The word *seire* is a good illustration of this. According to the official norm, this word is the infinitive form of the infinitive *seire* 'win', as in (5). However, it turns out that many people also use this word as a plural indefinite form of the noun *seier* 'victory' instead of the correct form *seirer*, as in (6*).*

(5) Folket vil *seire*
(6) a. Laget vant to viktige *seirer* (Correct plural form of *seier*)
b. Laget vant to viktige *seire* (Incorrect plural form of *seier*)

The spell checker cannot reveal the error since *seire* is a correctly spelt word in the lexicon. The result of the multi-tagging and disambiguation is therefore that *seire* is analysed as an infinitive, and the grammar checker has no chance of discovering the fact that *seire* is a wrongly inflected noun. A solution would be to expand the lexicon, so that it would contain commonly misspelled words like these:

(7) Lexicon entries for misspelled words:
seire N MASC PL INDEF NOM <Incorr>

6.3 Errors because of missing words

Finding that a word is missing can be a real problem for the grammar checker. The reason is simple: The grammar checker has no semantic knowledge - it does not understand the meaning of words. Determining whether there is a word missing, and what it might be, is almost impossible. Still, we have included CG rules for one seemingly simple case: that of a missing infinitive marker *å*.

Simplifying somewhat, infinitives generally cannot occur on their own without a modal verb or an infinitive marker:

(8) Han kan sykle (with a modal verb)
(9) Han har lært å sykle (with an infinitive marker)
(10)* Han har lært sykle (missing infinitive marker)

It is easy to make the grammar checker find errors like the one in (10): If there is no infinitive marker or modal verb in front of the infinitive, then notify the user that there is something missing. Below is an authentic example:

(11) Hun bruker _ låne deres fres fordi hennes egen står der med brukket splint, sår bihule og tett snabel (...)

However, there is still an overall problem: How does the grammar checker actually know that a given word is an infinitive? Infinitives are frequently ambiguous with other words.

7. Advantages of a system based on a morpho-syntactic analyser

We have seen above that the grammar checker is applied on text that has been grammatically tagged and disambiguated by a Constraint Grammar component. The fact that the input text has been morphologically analysed before grammar checking is an important feature of the system: It is often the case that a sequence of words may be grammatical in one larger context, and ungrammatical in another. Let us take as an example the sequence below:

(12) de situasjonen

In (13a), this sequence is grammatically incorrect, and in (24b), correct:

(13) a. * Vi liker ikke de situasjonen
b. I dag forstår de situasjonene

In (13a), *de* is a determiner with the wrong number agreement features with respect to the following noun; in (13b), *de* is a pronoun. With no morphosyntactic analysis, the difference between these two sentences would be impossible to detect.

References

Arppe, A. 2000. Developing a grammar checker for Swedish. In Nordgård, T. (ed.) *Nodalida '99 Proceedings from the 12th Nordiske datalingvistikkdager*, Department of Linguistics, University of Trondheim, p. 13-27.

Birn, J. 2000. Detecting grammar errors with Lingsoft's Swedish grammar checker. In Nordgård, T. (ed.) *Nodalida '99 Proceedings from the 12th Nordiske datalingvistikkdager*, Department of Linguistics, University of Trondheim, p. 28-40.

De Smedt, K. and V. Rosén. 2000. Automatic proof reading for Norwegian: The challenges of lexical and grammatical variation. In Nordgård, T (red.) *Nodalida '99 Proceedings form the 12th "Nordiske datalingvistikkdagar"*, Department of Linguistics, University of Trondheim, p. 206-215.

Hagen, K., and J.B. Johannessen. 1998. *Disambiguering uten syntaks.* In Faarlund, J.T., Mæhlum, B. og T. Nordgård (eds.) MONS 7, p. 68-79,Novus forlag, Oslo.

Hagen, K., J.B. Johannessen and A. Nøklestad. 2000a. The shortcomings of a tagger. In Nordgård, T (red.) *Nodalida '99 Proceedings form the 12th "Nordiske datalingvistikkdagar"*, Department of Linguistics, University of Trondheim, p. 66-75.

Hagen, K., J.B. Johannessen and A. Nøklestad. 2000b. A Constraint-based Tagger for Norwegian. I Lindberg, Carl-Erik og Steffen Nordahl Lund (red.): *17th Scandinavian Conference of Linguistics. Odense Working Papers in Language and Communication 19*, 31-48, University of Southern Denmark, Odense.

Karlsson, F., A. Voutilainen, J. Heikkilä og A. Anttila. 1995. *Constraint Grammar.* A Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin.