A Constraint-Based Tagger for Norwegian

Kristin Hagen, Janne Bondi Johannessen and Anders Nøklestad
The Text Laboratory, University of Oslo

1. Introduction

At the University of Oslo, work is currently being carried out to develop an automatic morphosyntactic tagger for Norwegian *bokmål* and *nynorsk*. The project comprises seven man-labour years and extends from April 1996 to December 1998[1].

2. Morphosyntactic taggers

2.1 General

Now, what is a morphosyntactic tagger? It is a computer program whose task is to provide correct grammatical information for every word in any stretch of running text. Ideally, every word should have a grammatical description, and the description should be unambiguous. This means that the tagger must be able to disambiguate words that have more than one possible reading, like the word *så* in (1):

(1) *'On the tram to town today, I saw her again':*

På trikken til byen i dag	**så**	jeg henne igjen
	"se" verb past tense	
	"så" conjunction	
	"så" adverb	
	"så" noun masc. appellative sg. indef.	
	"så" verb imperative	
	"så" verb infinitive	

The tagger should have the ability to decide that *så* is a past tense verb in this context.

[1] The project is mainly financed by the Norwegian Research Council, the Documentation Project, the Text Laboratory and the Finnish company Lingsoft.

Our disambiguating tagger consists of the following parts:

(2) I. Preprocessor
- Splits the text into sentences
  - uses special lists of abbreviations to decide when a full stop is a period and when it belongs to an abbreviation
  - decides when a word sequence is a headline rather than a sentence, based, among other things, on the number of words in the sequence

II. Multi-tagger
- Uses a full form list, i.e., a list of inflected words, to mark each word form with all tags that are appropriate for it
- Puts tags on remaining delimiters like commas, colons and dashes
- Analyses unknown words in a compounding module, using linguistic rules and lists of full forms and stems
- Marks capitalised unknown words as proper nouns

III. Morphological disambiguator

IV. Morphosyntactic mapping
- Performs mappings from morphological description to a set of possible syntactic functions for each word in the text

V. Syntactic disambiguator

In this talk we will concentrate on parts 3 to 5, i.e., morphological disambiguation, morphosyntactic mapping and syntactic disambiguation.

For these parts we are using a formalism called Constraint Grammar, which is developed at the University of Helsinki by Fred Karlsson and co-workers. The computer implementation of Constraint Grammar that we are using is developed by Lingsoft, a Finnish company. Constraint Grammar is similar to so-called Dependency grammar. A special feature of this kind of grammar is that it does not split larger grammatical structures into hierarchical sub-structures. Phrases like NP or VP do not occur in this grammar; only relationships between individual words are described.

2.2 Statistical goals

If the tagger provides the correct tags for each word, it reaches what is called a hundred percent *recall*. Furthermore, if no words receive any incorrect tags, we obtain a hundred percent *precision*. Although a hundred percent recall and precision are the ideal goals for the tagger, we are not aware of any other taggers that have reached such high numbers. Based on results from other taggers, a more realistic goal might be about 96% precision and a recall close to 100%.

However, the larger the number of morphosyntactic distinctions that is reflected in the set of tags, the more difficult it will be to reach a high degree of disambiguation. We have nevertheless chosen a considerably larger set of tags than what seems to have been used with comparable taggers for other languages, because we want to retain as many distinctions as possible in the tagger's output. At the moment we have reached a recall of 99.1% and a precision of 94%, but these numbers should increase somewhat before the end of the project.

3. Morphological disambiguation - how to find the finite verb

In Norwegian the finite verb functions as the second constituent in declarative sentences. As first constituent we can find all kinds of phrases containing an indefinite number of words and phrases:

(3) Different NPs as first constituent:

a. Du **går** ...
You walk ...

b. En mann **går** ...
A man walks ...

c. En sliten mann **går** ...
A tired man walks ...

d. Mannen **går** ...
The man walks ...

e. Mannen i parken **går** ...
The man in the park walks ...

f. Mannen med brun frakk og svart hatt **går** ...
The man with a brown coat and a black hat walks ...

g. Mannen med en frakk som er så skitten at den stinker, **går** ...
The man with a coat that is so dirty that it stinks walks

h. Mannen med den evnen at han kan snakke med månen, **går** ...
The man with the skill that he can talk to the moon walks ...

(4) Different AdvPs as first constituent:

a. Da **går** du ...
Then you walk ...

b. I dag og i morgen **går** du ...
Today and tomorrow you walk ...

c. Ofte **går** du ...
Often you walk ...

d. Heller **går** du ...
Rather you walk (= you rather walk) ...

e. Enten **går** du ...
Either you walk ...

f. Ganske ofte **går** du ...
Quite often you walk ...

Since we are using the Constraint Grammar model without a phrasal level it's difficult to determine where the first constituent ends. Many words are also multiply ambiguous with both verbal and non-verbal readings, as shown in (5).

(5) Ambiguous example: adverb as first constituent

Likevel **koster** huset mange penger.
Still the house costs a lot of money.

```
"<Likevel>"
        "likevel" adv
"<koster>"
        "kost" subst mask appell fl ub         (”brushes”- plural noun)
        "koste" verb pres                      (”costs”- present tense verb)
"<huset>"
        "hus" subst nøyt appell ent be
        "huse" adj <perf-part> mask fem ub ent
        "huse" adj <perf-part> nøyt ub ent
        "huse" verb perf-part
        "huse" verb pret
"<mange>"
        "mange" adj pos ub be fl
"<penger>"
        "penge" subst mask appell fl ub
"<.>"
         "." CLB  <PUNKT>
```

Here, both *koster* and *huset* could be the finite verb of the sentence. Note that this is an example of a multi-tagged text that has not yet been disambiguated.

Since the constraints must be formulated locally, without reference to phrases or other hierarchical structures, we have developed a large number of detailed constraints for each possible correct reading. Today we have about 90 constraints for verbs in the present or past tense. One of them, shown in (6), deals with sentences like the one above. (For each constraint I will show you, I will concentrate on an explanation of what the constraint does. The actual form of the constraint, which is given below, is only included for the benefit of those who are familiar with the formalism and are interested in what the constraint looks like).

(6) Constraint: adverb as first constituent

Choose the "verb pres"-reading if you can find nothing but adverbs between a clause boundary and the potential verb.

```
(@w =! (verb pres)
        (*-2C SETN-GR *R)
        (NOT *R IKKE-ADV)
        (-1C ADVERB)
        (NOT -1 "IKKE/OGSÅ"))
```

This constraint is used by the tagger to choose *koster* as the finite verb of the sentence, as you can see in (7):

(7) Disambiguated example: adverb as first constituent

Likevel **koster** huset mange penger.
Still the house costs a lot of money.

```
"<Likevel>"
        "likevel" adv
"<koster>" D:543
        "koste" verb pres
"<huset>" D:819, 817, 559, 447
        "hus" subst nøyt appell ent be
"<mange>"
        "mange" adj pos ub be fl
"<penger>"
        "penge" subst mask appell fl ub
"<.>"
        "." CLB <PUNKT
```

We have tried to make the constraints as specific as possible in order to avoid mistaggings.

(8) Ambiguous example: One noun functioning as first constituent

... at produktene **var** ufarlige
... that the products were not dangerous

```
"<at>"
        "at" sbu
"<produktene>"
        "produkt" subst nøyt appell fl be
"<var>"
        "var" adj pos mask fem ub ent       (”sensitive” – positive adjective)
        "var" subst nøyt appell ent ub      (”pillow case” – singular noun)
        "var" subst nøyt appell fl ub       (”pillow cases” – plural noun)
        "vare" verb imp                     (”last” – imperative verb)
        "være" verb pret                    (”was/were” – past tense verb)
"<ufarlige>"
        "ufarlig" adj pos mask fem nøyt be ent
          "ufarlig" adj pos ub be fl
```

For instance, we have a constraint for sentences like (8), where there is one noun functioning as first constituent.

(9) Constraint: One noun functioning as first constituent

Choose the "verb pret" reading if you find a noun in the first position to the left and a clause boundary in the second one.

```
(@w =! (verb pret)
       (NOT 0 PERF-PART)
       (-2C SETN-GR)
       (-1C SUBST)
       (NOT -1 GEN)
       (NOT *1 OVERSKRIFT))
```

(10) Disambiguated example: Noun as first constituent:

```
"<at>"
        "at" sbu
"<produktene>"
        "produkt" subst nøyt appell fl be
"<var>" D:834, 833, 588
        "være" verb pret
"<ufarlige>" D:661
        "ufarlig" adj pos ub be fl
```

Another constraint works on sentences where the first constituent consists of a whole noun phrase containing a PP. A constraint of this kind must of course be more complicated:

(11) Ambiguous example: NP with PP as first constituent

Det gamle huset med den store verandaen **kostet** mange penger
The old house with the big veranda cost a lot of money.

```
"<Det>"
	"det" det dem nøyt ent
	"det" pron pers 3 ent nøyt
"<gamle>"
	"gammal" adj pos mask fem nøyt be ent
	"gammal" adj pos ub be fl
	"gammel" adj pos mask fem nøyt be ent
	"gammel" adj pos ub be fl
"<huset>"
	"hus" subst nøyt appell ent be
	"huse" adj <perf-part> mask fem ub ent
	"huse" adj <perf-part> nøyt ub ent
	"huse" verb perf-part
	"huse" verb pret
"<med>"
	"med" prep
"<den>"
	"den" det dem fem ent
	"den" det dem mask ent
	"den" pron pers 3 ent mask fem
"<store>"
	"stor" adj pos mask fem nøyt be ent
	"stor" adj pos ub be fl
"<verandaen>"
	"veranda" subst mask appell ent be
"<kostet>"
	"kost" subst nøyt appell ent be
	"koste" adj <perf-part> mask fem ub en
	"koste" adj <perf-part> nøyt ub ent
	"koste" verb perf-part
	"koste" verb pret
"<mange>"
	"mange" adj pos ub be fl
"<penger>"
	"penge" subst mask appell fl ub
"<.>"
	 "." CLB  <PUNKT>
```

(12) Constraint: NP with PP as first constituent

Choose the "verb pret" reading if the word is preceded by a clause boundary followed by a sequence of adverbs, adjectives, nouns, pronouns, conjunctions, determiners and prepositions, and the first word in the sequence is not a preposition.

```
(@w =! (verb pret)
        (NOT 0 PERF-PART)
        (*-2C SETN-GR *R)
        (NOT *R IKKE-ADV-ADJ-SUBST-PRON-KONJ-DET2-PREP)
        (*-2C SETN-GR R+1)
        (NOT R+1 PREP)
        (-1C SUBST)
        (NOT -1 GEN)
        (NOT *1 OVERSKRIFT)
        (NOT 1 PRES/PRET))
```

The first word after the clause boundary cannot be a preposition because that would mean that the clause is introduced by a PP rather than an NP containing a PP, and this particular constraint is not meant to deal with that kind of construction. (13) shows the result of using this constraint:

(13) Disambiguated example: NP with PP as first constituent

Det gamle huset med den store verandaen **kostet** mange penger
The old house with the big veranda cost a lot of money.

```
"<Det>" D:261
        "det" det dem nøyt ent
"<gamle>" D:786
        "gammal" adj pos mask fem nøyt be ent
        "gammel" adj pos mask fem nøyt be ent
"<huset>" D:673
        "hus" subst nøyt appell ent be
"<med>"
        "med" prep
"<den>" D:286
        "den" det dem mask ent
"<store>" D:806
        "stor" adj pos mask fem nøyt be ent
"<verandaen>"
        "veranda" subst mask appell ent be
"<kostet>" D:817, 583
        "koste" verb pret
"<mange>"
        "mange" adj pos ub be fl
"<penger>"
        "penge" subst mask appell fl ub
"<.>"
        "." CLB <PUNKT>
```

4. Syntactic disambiguation

As we saw in (2), the last two components of the system perform syntactic analysis of the text. Thus, while morphological constraints are used for deciding on a morphological description, the last components try to decide what syntactic function these words have in the sentence. Syntactic analysis requires that morphological disambiguation is carried out first, since it would be impossible to determine syntactic functions like Subject and Object without knowing, for instance, which word is the finite verb of the sentence.

Like the morphological constraints, the syntax used in Constraint Grammar is word based. This means that no hierarchical phrase structure is constructed. Instead, phrasal *heads* are marked with syntactic main functions like Subject or Object, while other words in the phrase are simply marked as modifiers of some word to the left or right in the phrase. This is illustrated in (14). The Subject of the sentence is the noun phrase *det gamle huset* "the old house", but only the head of the phrase, i.e., *huset*, is marked as Subject. The other two words in the phrase, *det* and *gamle*, are marked as modifying determiner and adjective, respectively, with arrows pointing in the direction of the head they modify – that is, *huset*.

(14)

The old house cost a lot of money.

Det	gamle	huset	**kostet**	mange	penger
@DET>	@ADJ>	@SUBJ	@FV	@ADJ>	@OBJ

We don't have time to go into details about the entire inventory of syntactic labels we are using, but the labels, with explanations, are listed in (15) for reference.

(15)

@<ADV	ADVERBIAL that modifies something to the left
@<DET	DETERMINER that modifies something to the left
@<P-UTFYLL	COMPLEMENT OF PREPOSITION to the left
@<SBU	SUBJUNCTION that modifies something to the left
@<SBU-REL	RELATIVE SUBJUNCTION ("SOM") that modifies something to the left
@ADJ>	ADJECTIVE that modifies something to the right
@ADV	ADVERBIAL
@ADV>	ADVERBIAL that modifies something to the right
@DET>	DETERMINER that modifies something to the right
@FV	FINITE VERB
@I-OBJ	INDIRECT OBJECT
@INTERJ	INTERJECTION
@IV	NON-FINITE VERB
@KON	CONJUNCTION
@LØS-NP	NP without any syntactic function
@O-PRED	OBJECT COMPLEMENT
@OBJ	DIRECT OBJECT
@S-PRED	SUBJECT COMPLEMENT
@SUBJ	SUBJECT
@SUBST>	NOUN that modifies something to the right
@TITTEL	TITLE
@APP	APPOSITION

As was shown in (2), syntactic analysis is done in two stages. In the first stage, the morphological descriptions for each word are mapped to a set of possible syntactic tags. Most mapping rules put some constraints on the word's environment, like the one shown in (16), which was used for marking *gamle* in (14) as a modifying adjective:

(16) Morphosyntactic mapping for definite neuter singular adjectives

If a definite neuter singular adjective is directly followed by a word that can be a neuter singular noun, and that cannot be anything but a noun, mark the adjective as a modifier for the noun.

```
((adj nøyt be ent)
        ((1 ENT-NØYT-SUBST)
        (NOT 1 IKKE-SUBST))
        (@ADJ>))
```

The mapping rules can only specify context in terms of morphological descriptions. This is sufficient for giving unambiguous syntactic labels to verbs and to most modifiers. Thus, after morphosyntactic mapping has been carried out, the words in our example sentence will have the syntactic labels shown in (17), where we have included a PP as part of the Subject NP:

(17) *The old house with the big veranda cost a lot of money.*
Det gamle huset med den store verandaen **kostet** mange penger

Det	@DET>
gamle	@ADJ>
huset	@SUBJ @OBJ @I-OBJ @S-PRED @<P-UTFYLL @TITTEL @SUBST> @APP @LØS-NP
med	@ADV
den	@DET>
store	@ADJ>
verandaen	@SUBJ @OBJ @I-OBJ @S-PRED @<P-UTFYLL @TITTEL @SUBST> @APP @LØS-NP
kostet	@FV
mange	@ADJ>
penger	@SUBJ @OBJ @I-OBJ @S-PRED @<P-UTFYLL @TITTEL @SUBST> @APP @LØS-NP

In order to get any further with the analysis, however, we must be able to refer to other syntactic labels in the sentence. The last syntactic component, the *syntactic disambiguator*, enables us to do this. This component contains constraints that choose syntactic labels by referring to other syntactic labels and/or the morphological tags of other words in the sentence. We will see later on that the position of words relative to the finite verb of the sentence can be used to determine functions like Subject and Object. First, however, we need to make sure that only one word in each phrase is available as the carrier of a syntactic main function. Otherwise we could, for instance, be running the risk of choosing *verandaen* ("the veranda") as Subject in our example sentence, since this is the word that occurs immediately in front of the finite verb. Thus, we first run a number of syntactic constraints that will check whether any of the words are modifiers. One of these is shown in (18); this constraint will identify *verandaen* as the complement of the preposition *med* in our example sentence.

(18) Syntactic constraint for complement of preposition

If there is an unambiguous preposition to the left and there are no syntactic labels other than premodifiers (@DET>, @ADJ>, @ADV>, @SUBST>, @TITTEL and @KON) or Adverbials between the preposition and the word we are examining, mark the word as a complement of the preposition.

```
(@w =s! (@<P-UTFYLL)
        (*-1C PREP *R)
        (NOT *R SETN-GR)
        (*-1C PREP R+1)
        (NOT R+1 TEGNSETTING/KONJ)
        (*-1C PREP *R)
        (NOT *R IKKE-PREMOD-@ADV))
```

Note that this constraints refers both to purely morphological information, such as the part of speech *preposition*, and to syntactic labels like premodifying determiner and Adverbial. When this constraint has been applied to the sentence, we get the set of labels shown in (19):

(19)

Det	@DET>
gamle	@ADJ>
huset	@SUBJ @OBJ @I-OBJ @S-PRED @<P-UTFYLL @TITTEL @SUBST> @APP @LØS-NP
med	@ADV
den	@DET>
store	@ADJ>
verandaen	@<P-UTFYLL
kostet	@FV
mange	@ADJ>
penger	@SUBJ @OBJ @I-OBJ @S-PRED @<P-UTFYLL @TITTEL @SUBST> @APP @LØS-NP

What remains now is to determine the syntactic functions of *huset* and *penger*. The constraint that chooses Subject function for *huset* is based on the fact that indefinite nouns occurring after the verb rarely function as Subject, so that in such cases the Subject is more likely to be found in front of the verb. The constraint is fairly complicated, but what it does is essentially this:

(20) Syntactic constraint for Subject

For a noun, nominative pronoun or determiner that occurs before the finite verb, choose the Subject label if there is an indefinite common noun after the verb as long as that noun can have some syntactic function other than Subject.

```
(@w =! (@SUBJ)
        (0 PRON/SUBST/DET)
        (NOT 0 "DET/DETTE")
        (NOT 0 @ADV)
```

```
(NOT 0 AKK)
(NOT 0 SP)
(NOT 0 "SAMME")
(*-1 SETN-GR *R)
(NOT LR0 PREP R+1)
(NOT R+1 PREP)
(*-1 SETN-GR*R)
(NOT *R IKKE-MOD-@<P-UTFYLL-@ADV-@APP)
(*1 IKKE-MOD-@<P-UTFYLL-@ADV-@APP *R)
(*R IKKE-MOD-@<P-UTFYLL-@ADV-@APP *L)
(NOT *L CLB)
(*1 IKKE-MOD-@<P-UTFYLL-@ADV-@APP *R)
(LR0 @FV *R)
(NOT LR0 SITATVERB *R)
(*R IKKE-MOD-@<P-UTFYLL-@ADV-@APP *R)
(LR0 IKKE-@SUBJ *R)
(LR0 SUBST *R)
(NOT LR0 BE *R)
(NOT LR0 UBSLEKTSTERM *R)
(NOT LR0 PROP *R)
(*R SETN/INFKONSTR-GR *L)
(NOT *L @IV)
(*1 @FV *R)
(*R IKKE-MOD-@<P-UTFYLL-@ADV-@APP *L)
(*L DEM/POSS/GEN *L)
(NOT *L @FV))
```

We see that in this constraint the position of the word we are examining, as well as the position of the other, indefinite, noun, are defined with respect to the finite verb of the sentence. This shows how important it is for the syntactic analysis that the verb was identified during morphological disambiguation.

Another point to note about this constraint is that it is a *heuristic* constraint. This means that we cannot be a hundred percent certain that the word will in fact be a Subject in this context, and so this constraint will only be applied if no other, more reliable constraint has already chosen a Subject for the sentence. (In fact, a number of other constraints have removed some of the other possible labels listed for *huset* in (18 and 20) before (20) is applied).

Now that we have decided what is the Subject of the sentence, it has become possible to identify the syntactic function of the word *penger*, through the use of the following constraint:

(21) Syntactic constraint for Direct Object

For a word that occurs after the finite verb, choose the Direct Object label if there is an unambiguous Subject in front of the finite verb, there are no copula verbs in the sentence, and the word we are examining cannot be an Adverbial.

```
(@w =! (@OBJ )
    (NOT 0 @ADV)
    (NOT 0 @S-PRED)
    (NOT 0 "MYE/LITE")
    (*-1 IKKE-MOD-U-@KON-@<P-UTFYLL-@ADV-@APP *L)
    (LR0 @FV *L)
```

```
(NOT LR0 KOP *L)
(*L IKKE-MOD-@<P-UTFYLL-@ADV-@APP *L)
(LR0 @SUBJ *L)
(NOT LR0 IKKE-@SUBJ *R)
(NOT *R SETN-GR)
(*-1 @FV *R)
(NOT *R KOMMA)
(*-1 @FV *R)
(NOT *R KOP)
(*-1 IKKE-MOD-U-PERFPART-@<P-UTFYLL-@ADV R+1)
(NOT R+1 PREP)
(*1 PERIODE-SLUTT/CLB-KONJ *L)
(NOT *L @OBJ))
```

Again, we see that the constraint uses the position of the finite verb as the point of reference for the position of the word we are examining as well as for the position of the Subject.

So, when the Subject and Direct Object constraints have been applied, we arrive at the set of labels shown in (22), which constitutes the final analysis of the sentence:

(22)

Det	@DET>
gamle	@ADJ>
huset	@SUBJ
med	@ADV
den	@DET>
store	@ADJ>
verandaen	@<P-UTFYLL
kostet	@FV
mange	@ADJ>
penger	@OBJ

5. Conclusion

In this talk we have shown how morphological and syntactic disambiguation of Norwegian texts is feasible even without recourse to a phrasal level. Such disambiguation is important because it makes it possible to tag large amounts of text which can be used, for instance, in linguistic research. Since no tagged Norwegian corpus is currently available (with the exception of a hand-tagged corpus of Ibsen texts) we think that the development of the tagger we have presented here will make a valuable contribution to research into the Norwegian language.