The Construction of a Tagged Danish Corpus

Thomas Bilgram
Dept. of Linguistics, University of Aarhus, Denmark. bilgram@ling.aau.dk

Britt Keson
DSL, Christians Brygge 1,1. 1219 København K, Denmark. parole@coco.ihi.ku.dk

Abstract

The object of this paper is to present ongoing work on the construction of a morphosyntactically tagged Danish corpus, which is an integral step in the making of a Constraint Grammar (CG) parser for Danish and also constitutes a part of the Danish contribution to the European PAROLE project. This paper discusses various aspects of the morphological description of Danish used here as well as some of the guidelines developed for the manual disambiguation process. Finally, it also briefly gives an overview of the objectives of the two projects involved.

1. Introduction

The work on the construction of a morphosyntactically tagged Danish corpus as described here was undertaken as a joint effort by stud. PhD Thomas Bilgram, University of Aarhus, and cand. mag. Britt Keson and stud. mag. Dorte Haltrup Hansen from Det Danske Sprog- og Litteraturselskab (DSL), one of the two Danish partners in the PAROLE project. One of the aims of the PAROLE project is to produce a morphosyntactically tagged corpus of approx. 250,000 running words for the languages of each of the European partners. For Thomas Bilgram, the tagged corpus will serve as testing material for the development of a Constraint Grammar automatic tagger/parser for Danish.

The Danish text material to be tagged was composed of a random selection of approx. 1600 excerpts containing one or more consecutive paragraphs (each excerpt totalling approx. 150 - 180 words) extracted in part from the 40 mill. word corpus of the Danish Dictionary at DSL. This subcorpus was then analysed using a Two-level description of Danish morphology, DAN-TWOL, to assign to each word all of the possible morphosyntactic analyses, after which the correct analysis was chosen during a manual disambiguation process. In order to make the result as consistent as possible, this disambiguation process was for the most part carried out in parallel on the same texts, and later consensus on possible differences was achieved among the human taggers. The first 50,000 running words were treated as a pilot project to ensure that the information yielded by DAN-TWOL was sufficiently detailed for the purposes of both projects, and to develop a tagging manual to guide the human taggers in the disambiguation process.

2. DAN-TWOL

2.1 TWOL: A short introduction

The Two-level algorithm was originally designed by Kimmo Koskenniemi [Koskenniemi 83]. A TWOL is based on the principle that a word can be regarded as having two separate levels, a surface and a lexical level, as well as a description of the relation between these two levels. The surface level is the word as it appears in a text. The lexical level consists of (i) the morphemes (i.e. lemmas, derivational morphemes and inflectional morphemes), (ii) a description of the morphotax – i.e. ·inflections, derivations, and compounding – and (iii) a description of allomorphy. The DAN-TWOL was made during the graduate studies of Thomas Bilgram [Bilgram 94].

The TWOL tagset vs. other tagsets

The fundamental idea in a TWOL is that a word is analysed by a process of accepting one letter of the word at a time. When a sequence of letters is accepted as a morpheme, the part of the analysis that this morpheme constitutes is appended to the analysis string. Hence, the resulting analyses have a very modular appearance[1] (See Fig. 1).

Fig. 1.

bil+Ø+Ø+Ø	=	**N**	**FLS**	SG	UBEST	NOM
bil+Ø+Ø+s	=	**N**	**FLS**	SG	UBEST	GEN
bil+er+Ø+Ø	=	**N**	**FLS**	PL	UBEST	NOM
bil+er+Ø+s	=	**N**	**FLS**	PL	UBEST	GEN
bil+Ø+en+Ø	=	**N**	**FLS**	SG	BEST	NOM
bil+Ø+en+s	=	**N**	**FLS**	SG	BEST	GEN
bil+er+ne+Ø	=	**N**	**FLS**	PL	BEST	NOM
bil+er+ne+s	=	**N**	**FLS**	PL	BEST	GEN

This modularity makes the analyses look somewhat different from other PoS analysis systems (such as the CLAWS word tagging system [Garside 97]), where a single, and often quite compact, tag bears all the morphological (and often syntactic and semantic) information. However, in general the full set of features in the TWOL can be compared to a tag in other systems, and this leads to a larger and much more detailed tag inventory. The definitions of the tags (i.e. strings of features) in a TWOL is focused on the contents of the morphological features, and hence the number of tags is a direct product of the number of features. At present, a Danish corpus of 250,000 running words was given just over 500 different tags when analysed by DAN-TWOL. This figure can be compared to other tag systems, such as the Penn Treebank corpus (approx. 50 tags) and the Brown Corpus (just over 200 tags).[2]

2.1.2 DAN-TWOL

The result of a DAN-TWOL analysis is a list of all the possible morphosyntactic analyses of a given word. More than half of the words in the corpus were ambiguous, and the average level of ambiguity was approx. 2 analyses for every word. Making this ambiguity explicit is the primary goal for the TWOL, and a word and the possible analyses given to it by DAN-TWOL is referred to as a cohort.

[1] The PoS and gender information (bold type) is part of the lexicon. The number, definiteness and case information (underlined) is a result of the acceptance of – possibly nil – morphemes.

[2] The differences in the tagsets in different taggers render it virtually impossible to make a comparison of performance and output from different automatic taggers. An attempt has been made by Jochen Leidner in a CLUE report [Leidner 97]. The figures mentioned here are from this report.

DAN-TWOL is composed of a description of a lexicon of Danish (i) lemmas, (ii) inflectional and (iii) derivational morphemes, as well as a description of some allomorphy. The lemmas in the lexicon are based on a machine-readable version of the 1986 edition of *Retskrivningsordbogen* [Sprognævn 86] the official Orthographical Dictionary of Danish. From this, all the regular lemmas were extracted and converted into the format needed in DAN-TWOL. For example, nouns were coded for gender, number and definiteness morphemes, and verbs were coded for tense and participle morphemes. All irregular lemmas were handcoded in the lexicon. The lexicon has since been incrementally updated from its original size of approx. 42,000 entries to presently just over 49,000 entries. The inflectional and derivational parts were implemented on the basis of [Arndt 92] and are very similar to the morphology outlined in the Orthographical Dictionary [Sprognævn 96].

The assumption in TWOL that the lexicon offers full coverage of all the words found in a given text is very rarely true. Large real-life corpus texts will always contain words not included in the lexicon, and these words are returned as unanalysed by DAN-TWOL, and hence have to be handled by other means[3]. A consequence of the relatively free compounding in DAN-TWOL is that words which should not have received an analysis by DAN-TWOL erroneously are analysed as various compounds. For example, the name 'Bilgram' is analysed as a compound of 'bil' (*car*) and 'gram' (*gram*). This situation can be regarded as a continuum; at one end we have a system based on a closed list of fully inflected words recognised by the analyser, and at the other end we have a very liberal morphotactical system. The former system yields very good analyses of all the recognised words, but leaves a great number of words unrecognised, while the latter system yields an analysis for almost every word in the text, but the analyses proposed by the system are not always acceptable. The TWOL system is clearly located at the liberal end of this continuum.

The Orthographical Dictionary [Sprognævn 96] contains a few notes (§ 35 and §36) on the change of PoS that takes place in connection with the inflected forms of certain lemmas. For example, past participles can appear inflected according to number and definiteness in the same way that adjectives are, and (typically the *-t* inflected form of) adjectives can also occur with an adverbial function. In the CG approach there is a clear-cut distinction between the PoS and the syntactic function and also an acceptance of the fact that words of a certain PoS in the lexicon can appear in syntactic positions typically occupied by words of another PoS. Hence, the option of adding, for example, all past participles to the lexicon as possible adjectives was deliberately avoided. Instead, the possibility of adding this syntactic information during the manual disambiguation process was introduced.

2.2 PoS Definition in DAN-TWOL

The PoS definition in DAN-TWOL is based on the difference in the set of applicable features. Nouns have one set of features, verbs another, pronouns a third, etc. Fig. 2 is a listing of the different PoS categories and possible features for those used in DAN-TWOL.

[3] A natural part of a CG tagger is a heuristic module that supplies the unrecognised word with a cohort.

Fig. 2.

PoS	1	2	3	4	5	6	7
N	Gender	Number	Definiteness	Case			
V	Mood	Tense	Voice	*(Gender*	*Number*	*Definite-ness*	*Case)*
A	Degree	Gender	Number	Definiteness	Case		
PRON	Number	Definiteness	Case				
NUM	Ord/Cad	Case					
EGEN	Case						

If this is compared to the PAROLE tagset in section 5, one will notice that generally the same features are present. The major differences are due to the TWOL distinction between the truly morphologically derivable information and information that is more syntactically orientated. The latter information is given in '< >': E.g. U <cc>, U <cs>, U <inf>, U <prep>, U <adv>, U <midlsubj>. The '< >' are used to represent other kinds of distributional or textual information as well: For example, <*> (initial capital letter), <upl> (noun without a plural form) and <x-sg> (referent of pronoun is singular).

3. The Disambiguation Process

During the disambiguation process, the human tagger evaluated all the analyses in the cohort and marked the analysis regarded as correct with the tag <Correct!>. As mentioned in 2.1.2, the tagset defined for the PAROLE necessitated the addition of 'transcategorization' tags during this disambiguation process (examples in section 4.2).

The actual disambiguation process was performed using telnet connections to a linux server. The texts were handled in a restrictive mode for the text editor emacs, allowing only movement in the text and the addition and deletion of the various types of 'correct' tags. Hence, it was not possible to edit the actual texts by mistake.

4. Some examples from the tagging guidelines

4.1.1 Common Nouns

In general, a capitalised noun which has received a morphosyntactic analysis as a common noun from the DAN-TWOL was tagged as a common noun, unless the word clearly is the name of person, a geographical location or the like. The motivation for this distinction between common and proper nouns was: (i) to take into consideration the widespread variation in the use capitalisation in various 'naming' expressions, and (ii) to avoid overpopulating the lexicon with potential (i.e. capitalised) proper nouns, creating an undesirable ambiguity for words in sentence-initial position. Since complex naming expressions like 'det konservative folkeparti' (*the conservative party*) are not recognised as a single unit by the morphological analyser, they had to be treated on a strict word-by-word basis. As seen below, a number of Danish personal names indeed overlap with common noun readings, but in these cases the proper noun tag was preferred (marked by ☞), and if missing, was added later to complete the cohort (marked by ✋).

Fig. 3.

common noun
>
proper noun
>
foreign word

nu har statsministeren bedt departementschefen undersøge sagen
(now te prime minister has asked the permanent undersecretary to investigate the matter)
☞ "stat\s#minister" N FLS SG BEST NOM

Det kunne være interessant at møde Statsministeren
(It could be interesting to meet the Prime Minister)
☞ "stat\s#minister" <*> N FLS SG BEST NOM

Det konservative Folkeparti er det største oppositionsparti
(The conservative Party is the largest opposition party)
"den" <*> PRON SG BEST NOM <x-int> <w-pers>
"det" <*> U <midlsubj>
☞"den" <*> PRON SG BEST NOM <x-int> <w-demo/art>
"konservativ" A POS UK PL UB NOM
☞"konservativ" A POS UK SG BEST NOM
☞"folk\e#parti" <*> N INT SG UBEST NOM

et folketingsmedlem for Det Konservative Folkeparti
(a member of parliament for The Conservative Party)
"den" <*> PRON SG BEST NOM <x-int> <w-pers>
"det" <*> U <midlsubj>
☞"den" <*> PRON SG BEST NOM <x-int> <w-demo/art>
"konservativ" <*> A POS UK PL UB NOM
☞"konservativ" <*> A POS UK SG BEST NOM
☞"folk\e#parti" <*> N INT SG UBEST NOM

"Det giver en masse rutine," siger Thomas Bjørn
("It gives a lot of experience," Thomas Bjørn/'bear' says)
"bjørn" <*> N FLS SG UBEST NOM
☞ "*bjørn" EGEN NOM

Peter Søndergård har leveret en levende montage
(Peter Søndergård/'southern farm' has given a vivid montage)
"sønder\#gård" <*> N FLS SG UBEST NOM
✋ "*søndergård" EGEN NOM

4.1.2 Proper Nouns

By far the most words to be marked as proper nouns during the disambiguation process were names of people and geographical locations (countries, cities, rivers etc.). In addition, the proper noun tag was given to the 'proper noun element' of complex naming expressions for companies, institutions, sports teams, music groups, book titles, film titles etc., leaving all words that overlap with lexical common nouns to be tagged as common nouns. Early on, it became clear that most of the capitalised word unrecognised by the DAN-TWOL analyser in fact turned out to be foreign words with what could be regarded as proper noun denotations, and hence the 'proper noun element' definition was extended to include either (i) the (capitalised) name of a person or geographical location or the like, or (ii) any capitalised word unknown to the DAN-TWOL analyser. In the few instances where the DAN-TWOL analyser had assigned a Danish morphosyntactic analysis to a capitalised foreign word that was part of a non-Danish phrase, it was decided to disregard this analysis and give consistent proper noun tags to each element of the whole expression.

det betalingsstandsede firma Accumulator Invest
(the suspended company 'Accumulator Invest')
✋ "*accumulator" EGEN NOM
✋ "*invest" EGEN NOM

i foråret 1990 blev Rungsted Gymnasium inviteret
(in spring 1990 Rungsted Grammar School was invited)
☞ "*rungsted" EGEN NOM
☞ "gymnasium" <*> N INT SG UBEST NOM

kvindlige og mandlige betjente fra Københavns Politis station
(female and male police officers from Copenhagen Police station)
☞ "*københavn" EGEN GEN
☞ "politi" <*> <upl> N INT SG UBEST GEN
☞ "station" N FLS SG UBEST NOM

bogen, som kommer fra Royal Botanic Gardens
(the book, which comes from the 'Royal Botanic Gardens')
"royal" <*> A POS FLS SG UBEST NOM
✋ "*royal" EGEN NOM
✋ "*botanic" EGEN NOM
"garde" <*> N FLS SG BEST GEN
✋ "*gardens" EGEN NOM

Compound nouns are relatively frequent in Danish, and here the generative element of the DAN-TWOL morphological analyser proved particularly useful for identifying the compound words not already listed in the lexicon. Incorporated in the analyser is the decision always to assign the PoS of the last element of a compound word, regardless of the PoS of the preceding element(s).

dansetruppen har banandansen med i programmet
(the dance group has the banana dance in its program)
"banan\#dans~en" <kentaur> N FLS SG UB NOM
☛ "banan\#dans" N FLS SG BEST NOM

at give den gamle Dickens-historie en flyvende start
(to get the old Dickens story off to flying start)
☛ "*dickens\#-historie" N FLS SG UBEST NOM

4.1.3 Foreign Words

Another common potential dilemma exists between analysing a word as a (common or proper) noun or as a foreign word. In our work, the **<foreign word>** tag was retained only as an absolute last resort for unrecognised, non-capitalised words that are not of Danish origin, and which, due to their relative obscurity or infrequency, we felt would be unreasonable to add to the lexicon, and which therefore remained completely unanalysed after the disambiguation process. Hence, the **<foreign word>** tag was *not* used to mark contemporary foreign loan-words that for one reason or another do not appear in the Orthographic Dictionary. Instead, these loan-words were given 'proper' Danish morphosyntactic analyses, whenever it was felt that enough information was present to do so. The decision to tag capitalised unrecognised words as proper nouns and non-capitalised unrecognised words (most frequently) as foreign words proved to be useful in most instances, an obvious exception being foreign song/film/book etc. titles and names of people, where capitalised and non-capitalised words may occur in succession.

Tjeneren undredede sig over aquadenten
(the waiter wondered about the aquadente)
✋ "aquadente" N FLS SG BEST NOM

fra den dag tilhører han Erlanders pojkar
(from that day on he belongs to Erlander's pojkar)
✋ "pojkar" **<foreign word>**

at sidde med en pint og "The Sun"
(to sit with a pint and "The Sun")
"pine" V INF PCP2 UK SG UBEST NOM
✋ "pint" N FLS SG UBEST NOM

den saudiske generalløjtnant Khalid bin Sultan
(the Saudi lieutenant-general Khalid bin Sultan)
☛ "*khalid" EGEN NOM
✋ "bin" EGEN NOM **<foreign word>**
"sultan" <*> N FLS SG UBEST NOM
✋ "*sultan" EGEN NOM

4.2 Adjectives, Adverbs and Participles

Adjectives, adverbs and participles also represent an interesting ambiguity class in Danish. In this case, with regard to the different needs of the potential users of the tagged texts, it was decided to adhere closely to the PoS assignments in the Orthographic Dictionary for these words, and not to assign the members of this ambiguity class the various possible PoS by expanding the DAN-TWOL lexicon. Instead, 'transcategorization' tags, such as **<use-adv>**, were used to mark the syntactic usage of these words in a particular context and thereby create a more informative, and hence more flexible, resulting tagset. As will become evident from the examples below, the general tendency with this ambiguity class was to prefer a participle analysis to an adjectival analysis, and to prefer an adjectival analysis to an adverbial analysis (See Fig. 4.)

Fig. 4.

participle
>
adjective
>
adverb

4.2.1 Adjectives and adverbs

When a lexical adjective functions syntactically as an adverb in a given context, it was decided to retain the adjective PoS analysis and to add the syntactic information with the **<use-adv>** tag. The only exceptions to this rule are words which have been assigned both an adjective PoS and an adverb PoS in the Orthographic Dictionary, usually because of a significant difference in meaning.

når blodet langsomt kommer i kog
(when the blood slowly begins to boil)
☞ "langsom" A POS INT SG UBEST NOM <use-adv>

desværre dør hun tidligt i filmen
(unfortunately she dies early in the movie)
☞ "tidlig" A POS INT SG UBEST NOM <use-adv>

naboerne er lige ved at ringe til børneværnet
(the neighbours are just about to call the RSPCC)
☞ "lige" U <adv>
"lige" A POS UK UT UB NOM
"lige" N FLS SG UBEST NOM
"lig" A POS UK PL UB NOM
"lig" A POS UK SG BEST NOM

4.2.2 Participles

As Danish lexical present and past participles often appear with an adjectival or adverbial syntactic function on par with lexical adjectives, it was decided to use transcategorization tags for participles as well. Attributive adjectival use of present and past participles is indicated with the **<use-adj>** tag as shown below. Hence no attempt has been made to convert lexical participles to adjectival PoS, even when the (past) participle appears inflected for number, definiteness and (occasionally) gender. Attributive adverbial use of present and past participles is marked with the **<use-adv>** tag.

hele familien overvejer følgende punkter
(the whole family is considering the following points)
☞ "følge" V INF PCP1 NOM <use-adj>

Han var iført meget stramme, slidte cowboybukser
(He was wearing very tight, worn jeans)
☞ "slide" V INF PCP2 UK PL UB NOM <use-adj>
"slide" V INF PCP2 UK SG BEST NOM

min mor klarer sig forbavsende godt
(my mother does surprisingly well)
☞ "forbavse" V INF PCP1 NOM <use-adv>

en stor gruppe formodet raske østeuropæiske børn
(a large group of presumably healthy East European children)
☞ "formode" V INF PCP2 UK SG UBEST NOM <use-adv>
"for|mod" <upl> N INT SG BEST NOM

When both a participle analysis and adjectival analysis are possible, the participle analysis is usually preferred, unless (i) the corresponding verbal lemma does not exist, or (ii) the two analyses differ too greatly in meaning (rarely the case). The former case is true when the DAN-TWOL morphological analyser has productively generated an analysis as the participle form of a non-existing verbal lemma.

kræft i livmoderen eller tilgrænsende organer
(cancer in the uterus or adjacent organs)
☞ "tilgrænsende" A POS UK UT UB NOM
"til|grænse" V INF PCP1 NOM

Man er født med et bestemt antal hårsække
(One is born with a certain number of hair follicles)
"bestemme" V INF PCP2 UK SG UBEST NOM
"be|stemme" V INF PCP2 UK SG UBEST NOM
☞ "bestemt" A POS UK SG UBEST NOM

5. The PAROLE Project

The purpose of the EU-funded LE-PAROLE project is to produce a large-scale harmonised set of 'core' corpus and lexicon resources for the languages of the European Union. One of the two main objectives of the corpus part of the PAROLE project is to produce large monolingual corpora of approx. 20 million running words, which will obey certain common mark-up conventions, namely the PAROLE Corpus Encoding Standard (CES) as presented in [Ridings 96], which is in line with the EAGLES/MULTEXT CES guidelines [Ide et al 95]. The other main objective is to construct a subcorpus of approx. 250,000 running words, which are to be morphosyntactically tagged according to predefined tagsets that are compatible with the PAROLE lexicons. These resources produced by PAROLE will be made available through the European Language Resources Association (ELRA).[4]

Fig. 5.

PoS	Type	3	4	5	6	7	8	9	10	11
Noun		Gender	Number	Case	▨	▨	Definiteness	▨	▨	▨
Verb		Mood	Tense	Person ▨	Number	Gender	Definiteness	TrCat	Voice	Case
Adj		Degree	Gender	Number	Case	▨	Definiteness	TrCat	▨	▨
Pron		Person ▨	Gender	Number	Case	Possessor	Reflexive	Register	▨	▨
Det	▨	Person ▨	Gender ▨	Number ▨	Case ▨	Possessor	▨	▨	▨	▨
Art	▨	Gender ▨	Number ▨	Case ▨	▨	▨	▨	▨	▨	▨
Adv		Degree	Function ▨	Wh-ness ▨	▨	▨	▨	▨	▨	▨
Adpos		Formation	Gender ▨	Number ▨	▨	▨	▨	▨	▨	▨
Conj		Ctype ▨	Coord-posit ▨	▨	▨	▨	▨	▨	▨	▨
Num	▨	Gender ▨	Number ▨	Case ▨	▨	▨	▨	▨	▨	▨
Interj	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Residual		▨	▨	▨	▨	▨	▨	▨	▨	▨
Unique		▨	▨	▨	▨	▨	▨	▨	▨	▨

□ = **categories and features used in the Danish PAROLE corpus**
▨ = **categories and features not used in the Danish PAROLE corpus**

The Danish PAROLE tagged subcorpus will be based on an automatic conversion of the DANTWOL morphosyntactic tags to the Danish PAROLE tagset format (see Fig. 5). The common PAROLE tagset format was specified in the PAROLE *Multilingual Corpus Tagset Specifications* [Volz and Lenz 96], which constitutes an enhanced version of the EAGLES/MULTEXT tagset as presented in [Monachini and Calzolari 96]. In this format, the first position in the tag string contains a character for the PoS, and the second position contains a character representing one of various predefined PoS subcategorization types. The positions 3 - 7 contain morphological features taken from a common PAROLE feature set, while the positions 8 and up contain morphological features that are language-specific. Underspecification and ambiguities are expressed using a combination of various ambiguity markers (and characters), while symbols such as '=' and '-' are used as fillers. In the tagged subcorpora, a morphosyntactic tag is expressed using the following notational format: <W lemma=word_lemma msd="tag_string">word</w>.

The following example (discussed in section 4.2) demonstrates what the conversion to the Danish PAROLE tagset and format looks like:

[4] More information about the PAROLE project in general can be found at http://www-tei.uic.edu/orgs/tei/app/le02.html. Information about the two Danish PAROLE partners in particular is available at http://coco.ihi.ku.dk /~parole/par_eng.html (Det Danske Sprog- og Litteraturselskab) and http://www.cst.ku.dk/home.html (Center for Sprogteknologi).

i foråret 1990 blev Rungsted Gymnasium inviteret:

```
<W lemma="i" msd="SP">i</W>  <W lemma="forår" msd="NCNSU==D">foråret</W>
<W lemma="1990" msd="AC---U=--">1990</W>  <W lemma="blive" msd="VADA=----A-
">blev</W>
<W lemma="Rungsted" msd="NP--U==-">Rungsted</W>
<W lemma="gymnasium" msd="NCNSU==I">Gymnasium</W>
<W lemma="invitere" msd="VAPA=S[CN]I[ARV]-U">inviteret</W>
```

6. Using the Tagged Corpus as a Testing Corpus

As mentioned above, the tagged corpus material produced through the process described here also represents an integral step in the testing of an automatic tagger/parser, the DAN-CG. The Constraint Grammar (CG) is a knowledge-based system in which every piece of linguistic information is hand-coded in the system by a linguist. The system does not make use of statistical information, and does not apply any kind of automatic learning algorithms on the corpus material. The system is considered to be automatic in that it can take any kind of text as input and deliver PoS and a – shallow – syntactic analysis for the words contained therein as output. [Karlsson 94] is an introduction to CG and its principles, and a specific description of the ENG-CG can be found in [Voutilainen 92].

Fig. 6.

```
Nye cykler ruster hurtigt
(New bikes rust quickly)
"<nye>"
      "ny"  A POS UK PL UB NOM
      "ny"  A POS UK SG BEST NOM <= see fig. 7.
"<cykler>"
      "cykle"  V FIN PRS AKT        <= see fig. 8
      "cykel"  N FLS PL UBEST NOM
"<ruster>"
      "ruste"  V FIN PRS AKT
"<hurtigt>"
      "hurtig"  A POS INT SG UBEST
```

The foundation of the CG tagging/parsing process is that one of the analyses in any cohort is the correct one, and this reduces the actual process of tagging/parsing to a choice between the analyses in the cohort, or – in CG terms – to a deletion of the contextually wrong analyses.

The input to the DAN-CG is DAN-TWOL material of the kind used as input in the manual disambiguation process as described above. A set of CG rules is applied to this material in order to remove the analyses in the cohort that are considered to be wrong in the context. In the sentence in Fig. 6 one can see that 'nye' is ambiguous between the A SG and the A PL analyses, and 'cykler' between the N PL and V FIN analyses. The contextually wrong analyses are marked in italics.

The CG rules used to disambiguate at present[5] have the appearance as shown in Figs. 7 and 8.

Fig.7.

```
REMOVE   (A SG NOM)
(1C      N-PL-NOM)
(NOT     *-1 V-TRI/VAL) ;
```

Remove the A SG NOM from the cohort if the next word to the right (1) definitely (C) is a N PL NOM, and no (NOT) trivalent verbal (TRI-VAL) can be found anywhere (*) to the left (-) starting the search from word 1(1).

[5] This sentence was constructed as an example, since real texts most often contain more ambiguity. The rules are taken from the present CG rule file, which is changing constantly.

When applying the standard method of developing a CG, one makes use of the testing mode of the CG software which is designed to return a warning whenever an analysis marked <Correct!> is removed from the cohort. With this mode on, one can quantify the output of the CG tagger/parser by counting how many analyses marked <Correct!> are removed by any given rule in the system, compared to how many analyses *not* marked <Correct!> are removed. Hence, if a rule happens to remove the correct A PL analysis of 'nye' above and leaves the incorrect A SG BEST analysis, this would appear in the log file, and the rule can be located and corrected. These testing procedures are needed since the number of rules in a fully developed CG is quite large (> 1000), and the amount of running words needed for developing it is large as well (often 50,000 to 100,000 words). Hence manually evaluating the output would simply be too time-consuming.[6]

Fig. 8.

```
REMOVE    (V FIN)
 (NOT 0   VERB-NO-CS)
 (*1C     V-FIN BARRIER CLS-BOUND/POS)
 (*-1      (>>>)
BARRIER  V-FIN
OR        CLS-BOUND/POS) ;
```

Remove the V FIN analyses from the cohort if the same (0) word is not (NOT) a verbal of the kind that can make a clause without a clause marker (E.g. 'høre' and 'se') (VERB NO CS), and a word whose only reading is that of a finite verb (V FIN) is found somewhere (*) to the right starting from the first position (1), and not preceded (BARRIER) by anything that can be a clause boundary (CLS BOUN/POS), and – to the left – no finite verbs (V FIN) or possible clause boundaries (CLS-BOUND/POS) are found before (BARRIER) the beginning of the sentence (>>>).

7. Perspective

The co-operative effort between the PAROLE project and DAN-TWOL/Constraint Grammar projects has proved to be very fruitful for both projects. The common tagset and common disambiguation procedure also rendered all the information necessary for both projects, and the parallel performance of the manual disambiguation has greatly improved the quality of the resulting tagged corpus. We hope that the product of our efforts will also be useful for other users in need of a large morphosyntactically tagged Danish corpus in the future.

Bibliography:

Arndt, H. (1992): "Towards a Syntactic Analysis of Danish Computer Corpora" in *Proceedings of the XIIth Scandinavian Conference of Linguistics.*

Bilgram, T. (1994): *Computerstyret analyse af dansk*, Specialeopgave, Dept. of Linguistics, University of Aarhus, Denmark.

Dansk Sprognævn (1986): *Retskrivningsordbogen* (the Orthographical Dictionary), Dansk Sprognævn. [also as a machine-readable text version].

Dansk Sprognævn (1996): *Retskrivningsordbogen* (the Orthographical Dictionary), Aschehoug.

Garside, Leech, Sampson (eds.) (1987): *The Computational Analysis of English, a corpus based approach.* London & New York, Longman.

[6] The making of the CG-tagger/parser for Danish is still in progress, but the present version can be tested at http://ling.hum.aau.dk/~bilgram/CG.html.

Ide, N., D. Durand, G. Priest-Dorman, and J. Veronis (1995) *MULTEXT: Corpus Encoding Standard*, LRE Project 62-050, CNRS. [The most recent version of the CES can be found at http://www.cs.vassar.edu/CES/].

Karlsson, F., A. Voutilainen, J. Heikillä, and A. Anttila (eds.) (1994): *Constraint Grammar: a Language-independent System for Parsing Unrestricted Text.* Berlin and New York: Mouton de Gruyter.

Koskenniemi, K. (1983): *Two-level Morphology. A General Computational Model for Word-form Production and Generation.* Publication No. 11, Dept. of Linguistics, University of Helsinki.

Leidner. J. (1997): "Evaluation Taggers for English: Some Evidence" in *CLUE-TR-971101.* Friederich-Alexander-Universität, Erlangen-Nürnberg, Institut für deutsche Sprache und Literaturwissenschaft. This report is available at: ftp://ftp.linguistik.uni-erlangen.de/pub/reports/CLUE-TR-971101.

Monachini, M. and N. Calzolari (1996): *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages*, EAGLES Document EAG-LSG/IP-T4.6/CSG-T3.2, EAGLES [This is also available at: http://www.ilc.pi.cnr.it/EAGLES96/morphsyn/morphsyn.html].

Ridings, D. (1996):*Text Representation in PAROLE*, MLAP PAROLE 63-386 WP 4.1.3, Göteborg universitet [Also available at: http://svenska.gu.se/~ridings/textrep/textrep.html]. Volz, N. and S. Lenz (1996): *Multilingual Corpus Tagset Specifications*, MLAP PAROLE 63-386 WP 4.1.4, IDS, Mannheim.

Voutilainen, A. (1992): *Constraint Grammar of English, A Performance-Oriented Introduction,* Helsinki University Press.