EN CONSTRAINT GRAMMAR PARSER FOR DANSK

Eckhard Bick
Institut for Sprog og Kommunikation, SDU Odense Universitet
lineb@hum.au.dk, http://visl.hum.sdu.dk

1. Projektet

Det moderne samfunds udstrakte brug af informationsteknologi i almindelighed og elektronisk tekstbehandling i særdeleshed har dels forsynet lingvistikken med et hidtil uset væld af let tilængelige data til kvalitativ og kvantitativ sprogforskning, dels sat den automatiske analyse og generering af natursprog (både tekst og tale) på den forskningspolitiske dagsorden. Det er imidlertid i overvejende grad det engelske sprog der har profiteret mest af denne udvikling, mens "mindre" sprog som dansk til stadighed udviser en udpræget mangel på store og nemt tilgængelige korpora, kvalitets-taggere og frisprogs-parsere.

Som led i et SHF-støttet tværsproglig forskningsprojekt på SDU Odense Universitet har jeg igennem det sidste år udviklet leksika, computerprogrammer og grammatiske regler til automatisk analyse af dansk skriftsprog. Målet har været en robust parser der håndterer almindelig løbende dansk tekst på både det morfologiske og syntaktiske niveau. Det anvendte parsing-paradigme, Constraint Grammar (Karlsson 1990, 1995), er tidligere undersøgt med henblik på dansk af Thomas Bilgram (1994), der tog udgangspunkt i en to-niveau-beskrivelse (TWOL) af dansk morfologi, og opnåede lovende resultater for især nominal-verbal-homonymien.

Den morfologiske analyse i mit eget system bygger ikke på en TWOL-algoritme, men på kildesprogsdelen af et tidligere, MT-orienteret projekt (1986), der siden er blevet udvidet med yderligere leksikografisk materiale fra især DSL (nu i alt ca. 106.000 leksemer), samt et voksende antal "skræddersyede" valens- og semantiske oplysninger (Anders Hougaard, Morten Vestergaard-Lund, Eckhard Bick). Den morfologiske og ordklasse-disambiguering hviler imidlertid - som hos Bilgram - på en grammatik bestående af kontekst-sensitive, såkaldte "Constraint"-regler, der "forbyder" eller "selekterer" bestemte læsninger i bestemte kontekster - hvor kontekstvinduet er perioden, og kontekstkriterier kan omfatte både ords morfologi, placering og valenspotentiale, samt i nogen grad et ords semantiske klassificering.

2. Constraint Grammar

De fleste ord i natursprogstekster er - isoleret set - flertydige med hensyn til ordklasse, bøjning, syntaktisk funktion, semantisk indhold m.m. Det er sætningskonteksten (foruden den indholdsmæssige samenhæng og læserens "viden om verden"), der afgør hvordan ordet skal forstås. *Constraint Grammar* (CG) er en grammatisk metode der søger at gennemføre en sådan éntydiggørelse (disambiguering) ved at opstille regler for hvilken af et ords mulige læsninger der skal vælges og hvilke læsninger der skal forkastes i en given sætningskontekst, eller om der skal tilføjes nye (typisk syntaktiske) læsninger. I selve parseren bliver reglerne kompileret til et computerprogram, der som input tager tekst hvor hvert ord har fået tilføjet en såkaldt *kohorte* (1) af flertydige morfologiske læsninger vha. et leksikon-baseret morfologisk analyseprogram.

(1) hun
    [hun] <femi> PERS UTR 3S NOM
    sælger
    [sælger] N UTR S IDF NOM (1. læsning i kohorten)
    [sælge] V PR AKT (2. læsning i kohorten)
    blomster
    [blomst] N UTR P IDF NOM
    $.

En typisk morfologisk disambigueringregel for 'sælger' i ovenstående eksempel er følgende:

SELECT (VFIN) IF
(-1C (PERS NOM) OR (N NOM)) (*-2 >>> BARRIER NON-PRE-N)
(NOT 0 PRP) ;

Vælg læsningen som finit verbum, hvis ordet til venstre (-1) er et personligt pronomen eller et substantiv i nominativ, og der mellem dette og sætningsstarten (>>>) ikke findes andet end prænominaler, med mindre mål-ordet (0) også kan være en præposition ('fx. ved').

Når Constraint Grammar-modellen bruges til syntaktiske beskrivelse, tilføjes ordbaserede funktions- og dependens-markører (fx. '@<SUBJ' for kernen i et subjekt-syntagme til højre for sætningens finitte verbum), der så igen danner udgangspunkt for yderligere disambigueringsregler.

MAP (@<ACC @<DAT @<SC) TARGET (PERS ACC) IF (NOT –1 PRP) ;

Tilføj læsninger som direkte objekt (@<ACC), indirekte objekt (@<DAT) og subjektsprædikativ (@SC) til personlige pronominer i akkusativ, hvis ordet til venstre ikke er en præposition.

Ca. 40% af de morfologiske, og en mindre andel af de syntaktiske CG-regler stammer direkte fra et lignende system for portugisisk (Bick 2000), resten af de p.t. ca. 2500 regler blev udarbejdet og testet specifikt for dansk.

3. Det deskriptive system

Parserens tag-sæt[1] indeholder 14 ordklasse-kategorier, der kombineres med 20 tags for bøjningsformer, ialt flere hundrede distinkte komplekse tags. I tag-linien 'V PR AKT', for eksempel, alternerer ordklassen 'V' således med 13 andre ordklasser, og indenfor V-klassen alternerer 'AKT' (aktiv) med 'PAS' (passiv), og 'PR' (præsens) med IMPF (imperfektum), samt med INF (infinitiv), IMP (imperativ), PCP1 (præsens participium) og PCP2 (perfektum participium), der alle bortset fra IMP og PCP1 også kombinerer med AKT/PAS (aktiv/passiv). Desuden bøjes én PCP2 klasse, den adjektiviske STA klasse, i numerus, genus, casus og bestemthed. På denne måde beskrives i alt 26 forskellige V-tag-linjer med 17 modulære deltags. Den analytiske karakter af tag-strengene gør dem mere "gennemskuelige", og letter desuden arbejdet for disambiguerings-reglerne. I modsætning til andre systemer (jf., for eksempel, CLAWS-systemet, som beskrevet i Leech, Garside, Bryant, 1994), skelnes der i tag-strengen skarpt mellem grundformer ("ord", leksemer), ordklasser og bøjningskategorier. Desuden etableres ordklasserne næsten udelukkende på morfologisk vis, og holdes dermed adskilt fra de syntaktiske kategorier. Således defineres et substantiv (N) paradigmatisk som *den* ordklasse der udviser genus som (invariant) leksemkategori og numerus som (variabel) ordformkategori. Derimod er både genus og numerus leksemkategorier for propria (PROP), og begge er ordformkategorier for adjektiver (ADJ)[2]. I tabellen (2) er leksemkategorier markeret med *.

(2) Ordklasser og flektionskategorier

word class	**genus**	**nume-rus**	**casus**	**bestemt-hed**	**person**	**tempus modus**	**diatese**	**grad**
	UTR NEU	S P	NOM GEN ACC	DEF IDF	1 2 3	PR IMPF IMP	AKT PAS	(POS) COM SUP
N	+*	+	+	+				
PROP	+*	+*	+					
INDP	+*	+*						
DET	+	+	+	+				

[1] En fuldstændig oversigt over de brugte morfologiske og syntaktiske tags og deres definitioner findes på http://visl.hum.sdu.dk.

[2] Pronominer kan opdeles efter samme skema, i en determiner-klasse (DET) med de samme (variable) kategorier som adjektiver, og en independent (INDP) klasse af "substantiviske" pronominer der udviser de samme (invariante) kategorier som propria-klassen. Personlige pronominer (PERS), som tredje klasse, har 4 ordformkategorier: numerus, genus, casus og person. Alle 3 pronominalklasser adskiller sig fra de "rigtige" nominalklasser ved at de ikke tillader derivation.

PERS	+	+	+		+			
ART	+	+		+				
ADJ	+	+	+	+				+
ADV								(+)
V VFIN					+	+	+	
V INF							+	
V PCP1			+					
V PCP2							+	
V PCP2 STA	+	+	+	+				
NUM	(+)	+*						
PRP								
KS, KC, KP								
IN								

Det syntaktiske tag-sæt råder over godt 20 tags for ord/syntagme-funktion (eller knap det dobbelte hvis dependensmarkører regnes med), der også forekommer i kombination med 3 tags for ledsætningstyper (finitte, infinitte og averbale), i alt ca. 50 distinkte tags.

(3) De vigtigste syntaktiske kategorier

@SUBJ	subjekt	**@ADVL**	frit adverbial
@ACC	direkte (akkusativ-) objekt	**@PRED**	frit prædikativ
@DAT	indirekte (dativ-) objekt	**@APP**	apposition
@PIV	præositionsobjekt	**@>N**	prænominal-dependent
@SC	subjektsprædikativ	**@N<**	postnominal-dependent
@OC	objektsprædikativ	**@>A**	adverbiel præ-dependent
@SA	subjektsrelateret argumentadverbial	**@A<**	adverbiel post-dependent
@OA	objektsrelateret argumentadverbial	**@P<**	præpositions-argument
@MV	hovedverbum	**@INFM**	infinitivmarkør
@AUX	hjælpeverbum	**@VOK**	vokativ

Modulerne for valens og semantik er under udvikling, og det er derfor vanskeligt at angive nøjagtige tal for tag-sættenes størrelse. Kategoriinventaret for både valens og semantik er inspireret af den portugisiske *Palavras*-parser (Bick, 2000) og rummer ca. 100 valenstags og ca. 200 såkaldte semantiske prototyper (fx. <vt> for 'transitivt verbum', <+på> for et substantiv eller adjektiv med præpositionalvalens, <madc> og <mad> for hhv. tællelig og utællelig mad, <drik>, <tøj> etc.). De semantiske kategorier for substantiver er baseret på 16 "atomare" træk (som fx., ±HUM, ±MASS), og genfindes i store træk også i SIMPEL-kategoriinventaret (jf. fx http://www.ub.es/gilcub/SIMPLE/reports/simple/simple_greek.htm).

3.2. Flad dependens-syntaks vs. konstituentgrammatiske træstrukturer

Hver af de ord-baserede CG-tags der bruges i den danske parser, består dels af en markør for syntaktisk funktion (fx. SUBJ, ACC), dels af en dependensmarkør (> <) der peger må syntagmekernen, eller – på sætningsplan – mod verbalet. Syntagmekerner og verbalkædekerner bærer så igen funktionsmærkerne for de komplekse konstituenter.

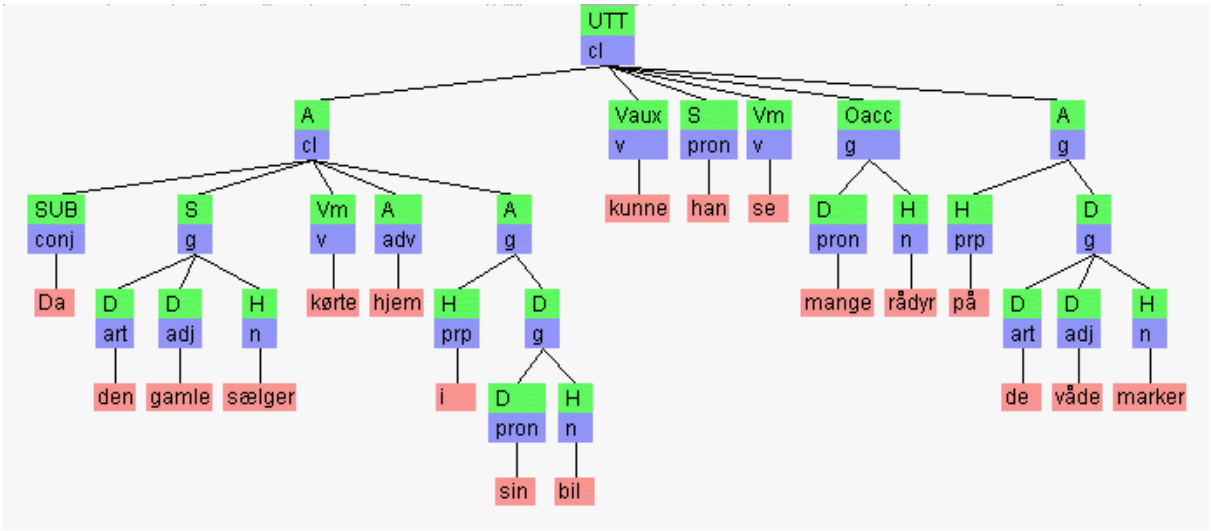
```
(4) Da        [da]       KS                              @SUB
    den       [den]      ART UTR S DEF                    @>N
    gamle     [gammel]   ADJ nG S DEF NOM                 @>N
    sælger    [sælger]   N UTR S IDF NOM                  @SUBJ>
    kørte     [køre]     <mv> V IMPF AKT                  @FS-ADVL>
    hjem      [hjem]     N NEU P IDF NOM                  @<ACC
    i         [i]        PRP                              @<ADVL
    sin       [sin]      <poss> <refl> DET UTR S            @>N
    bil       [bil]      N UTR S IDF NOM                  @P<
    ,
    kunne     [se]       <aux> V IMPF AKT                 @FAUX
    han       [han]      PERS UTR 3S NOM                  @<SUBJ
    se        [se]       <mv> V INF AKT                   @AUX<
```

```
mange    [mange]    <quant> DET nG P NOM        @>N
rådyr    [rådyr]    N NEU P IDF NOM &ACI-SUBJ @<ACC
på       [på]       PRP                       @<OA
de       [den]      ART nG P DEF                @>N
våde     [våd]      ADJ nG P nD NOM             @>N
marker   [mark]     N UTR P IDF NOM            @P<
```

I eksemplet ses hvordan subjektsyntagmet tegnes af kernen ("sælger"), der igen trækker to prænominaler (@>), en artikel og et adjektiv med sig. Ledsætningens verbum ("kørte") bærer et mærke for hele den finitte ledsætnings (FS) funktion som adverbial (@FS-ADVL>).

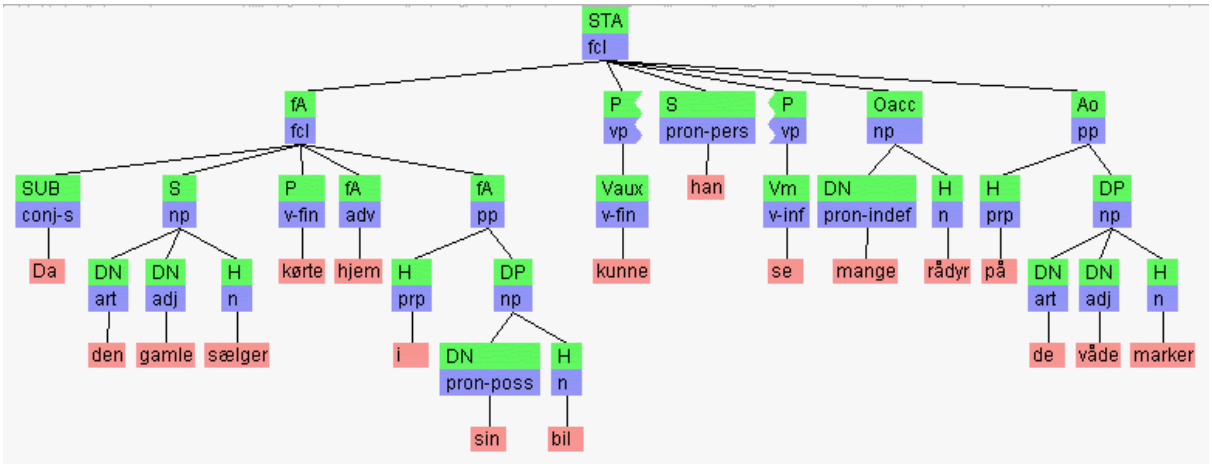
De konsekvente dependensmarkeringer, der mappes og éntydiggøres af CG-reglerne på samme vis som de syntaktiske funktioner, muliggør at generere grafiske konstituenttræer vha. en forholdsvis enkel psg-grammatik, fordi psg-reglerne kan støtte sig til den allerede eksisterende strukturelle information fra CG-niveauet.

(5a)



Det viste træ viser for hver knude både form (blå) og funtion (grøn). De konkret brugte symboler fremkommer ved en notationel filtrering der udvælger fra CG-notationen den pædagogisk/grammatikteoretisk ønskede informationsmængde og forkaster resten. Således ses i det simple træ (5a) kun 2 typer komplekse former, g (gruppe) og cl (clause), og verbalkæden beskrives fladt (Vaux og Vm), mens den mere avancerede træ (5b) samler verbalkæden i en prædikator-gruppe, skelner mellem substantivsyntagmer (np), præpositionssyntagmer (pp) etc., samt differentierer mellem flere underkategorier for verber og pronominer.

(5b)



4. Den danske parsers performans

Under udviklingen af parseren foretages løbende pilotevalueringer for mindre teksprøver, dels kvalitativ for at kunne rette eller tilpasse CG-reglerne, udvide leksikonnet samt bedømme de eksisterende reglers grammatiske dækningsgra, dels kvantitativt for at kunne prioritere bestemte fejltyper og bedømme parserens almindelige robusthed og performans. Korrekthedsprocenter fra tre nylige prøvekørsler med løbende (ukendt) avistekst er vist i nedenstående tabel:

(6) parserens **morfologiske og syntaktiske fejlprocenter** for fri løbende tekst

	korrekt morfologi	**korrekt syntaks**
avistekst (Århus Stiftstidende) ca. 1925 ord	98.3 %	92.5 %
avistekst (Politiken) ca. 1600 ord	98.5 %	92.6 %
Kulturnyt (Internet) ca. 1430 ord	97.6 %	92.0 %

Idet parseren ikke endnu er færdigudviklet, forekommer der en del tilfælde med ufuldstændig disambiguering, hvor både den korrekte læsning og en eller flere forkerte læsninger har "overlevet" CG-disambigueringen. Her blev den første læsning heuristisk anset for at være den endeligt valgte. En alternativ evaluering ville være at arbejde med separate tal for *recall* og *precision*, hvor *recall* (dvs. overlevende korrekte læsninger) så ville være lidt højere end tallet i tabellen, og *precision* (dvs. korrekte læsninger bland alle læsninger) lidt lavere.

5 Parserens tekniske data

Den danske parser består af en række programmoduler, der er skrevet i programmeringssprogene C og Perl. CG-regelsættene er kompatible med Pasi Tapainenens cg2-compiler (Tapainen, 1996) og Martin Carlsens VISL-cg-compiler. Parseren omfatter følgende moduler på det morfologisk-syntaktiske niveau:

- 1. et **morfologisk analyse-program**, som behandler orthografisk præprocessering, ordklasse, bøjning, derivation, komposita og faste udtryk (polyleksikalier). Analyse-modulet støtter sig til et **leksikon** med ca. 106.000 leksemer med angivelse af bøjningspotentiale (svarende til ca. 900.000 fuldformer). Kernen, ca. 11.000 leksemer, består af ord jeg oprindeligt havde udvalgt som danske "kerneord", hvis derivationer og komposita skulle kunne dække ca. 75% af opslagene i Nudansk Ordbog. Denne kerne er nu blevet kontrolleret og udvidet vha. store ordlister fra DSL (Dansk Selskab for Litteratur) og Niels Aage Schmidt (via Mikroværkstedet).
- 2. en **morfologisk disambiguator** med 1400 Constraint Grammar regler, der vha. diverse kontekstbetingelser forkaster eller selekterer blandt de læsninger der bliver foreslået af det morfologiske analyse-program.
- 3. en **syntaktisk "mapper"** med ca. 580 kontekstbaserede regler der "mapper"/tilføjer (flere mulige) syntaktiske funktioner udfra en ordforms disambiguerede morfologiske og ordklasse-tags. Disse regler indeholder også kontekstbetingelser, men er mere simple end disambigueringsreglerne.
- 4. en **syntaktisk disambiguator** med ca. 610 Constraint Grammar regler, der forkaster eller selekterer blandt de mappede syntaktiske funktionstags.
- 5. en **træ-generator** med ca. 170 PSG-regler, der etablerer syntagme- og ledsætningsgrænser udfra CG-notationens funktions- og dependensmarkeringer, og leverer grafiske syntaktiske træer[3].

Herudover findes der 430 CG-regler i separate MT-grammatikker til brug i oversættelsmodulet dansk-esperanto, og knap 100 CG-regler i et eksperimentelt modul til grammatisk stavekontrol. Udover leksemleksikonnet med bøjningsoplysningerne bruger

[3] Programmeringen af en skræddersyet compiler til den her anvendte form for PSG-regler blev udført af VISL's programmør Martin Carlsen.

systemet specialleksika der gør p.t. ca. 14.000 ords valenspotentiale og til dels semantiske klassificering tilgængelig for regelkonteksterne.

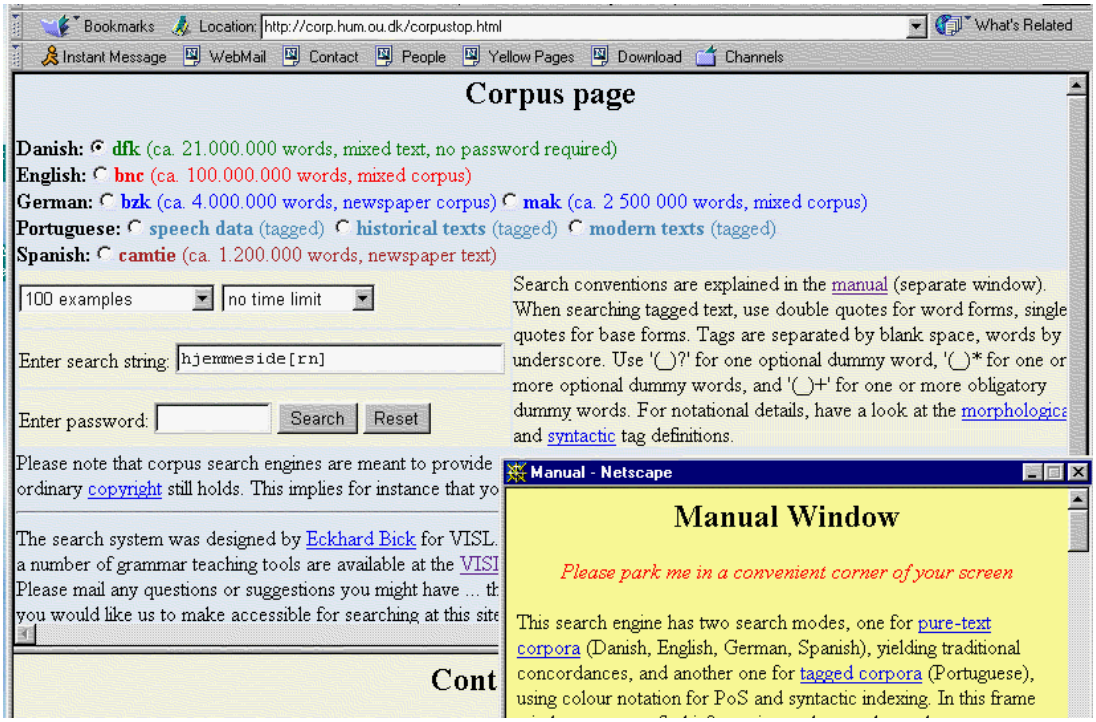
En fuldstændig grammatisk analyse på alle niveauer håndterer ca. 400 ord/sec på en 500 MHz Pentium-baseret Linux-maskine[4]. Den morfologiske analyse alene opnår hastigheder i nærheden af 1000 ord/sec.

6. Perspektivering

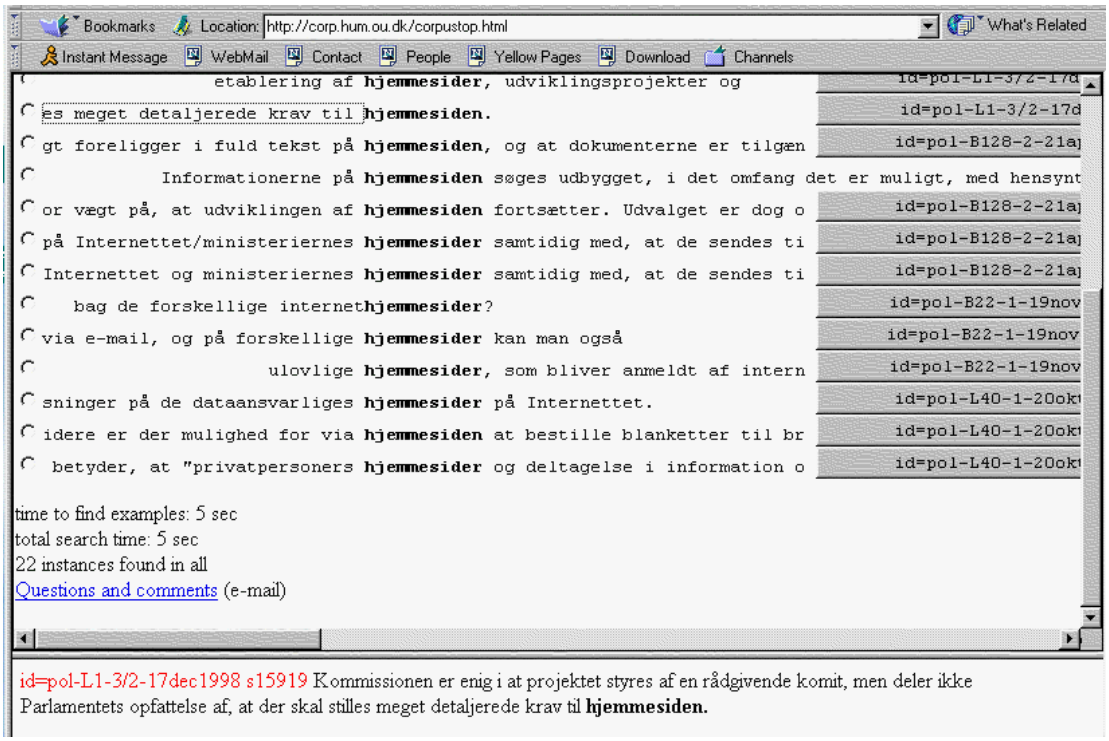
Det færdige system vil blive brugt til at annotere et sideløbende opbygget internet-baseret dansk korpus og udgøre kernen i den åbne del af det undervisningsorienterede VISL-system (http://visl.hum.sdu.dk, jf. Bick 1997). Det danske korpus rummer p.t. ca. 7 millioner ord (målsætningen 21 millioner ord), og består hovedsagelig af udskrifter fra folketingsdebatter, online-avistekster (LOKE), samt enkelte regulære bøger. Nedenstående vises den nuværende indgangsside til korpussøgningen (http://corp.hum.sdu.dk) (7a), samt resultatet af en eksempelsøgning (7b) ('hjemmesider'/'hjemmesiden'):

[4] Tallene gælder for større tekster, idet der skal regnes med en opstarttid på 6-7 sekunder til leksikonindlæsning etc. selv ved analyse af enkeltsætninger.

(7a) **korpus-applikationen**



(7b) **korpus-søgning**



Det færdigtaggede danske korpus vil ud over rene tekststrengssøgninger og såkaldte *regular expressions* også tillade søgning efter ordklasse, bøjningsform og syntaktisk funktion, fx. 'sælger N' til forskel fra 'sælger V' og 'sælger @SUBJ>' til froskel fra 'sælger @<ACC'.

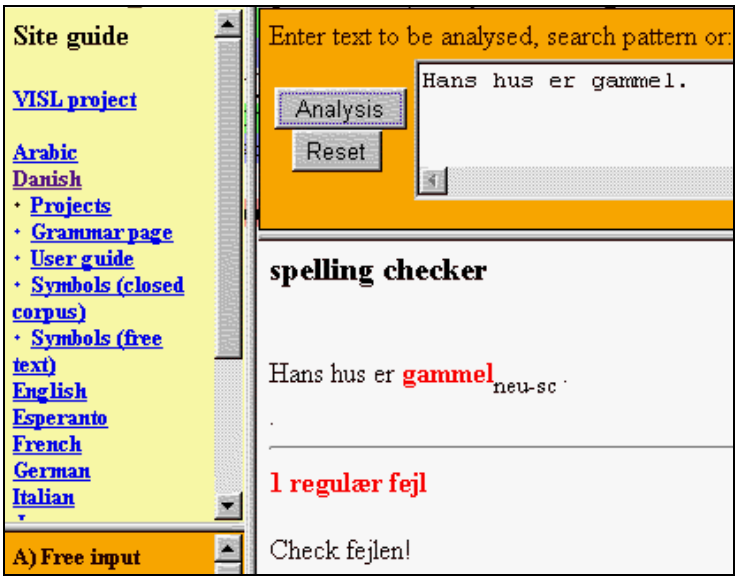
Desuden anvendes systemet allerede, på et eksperimentelt plan, til maskinoversættelse (dansk-esperanto), med sprogparrene dansk-engelsk[5] og dansk-portugisisk som mulige fremtidige applikationsområder. CG-metoden kan her anvendes i progressive trin, idet fx. valenspotentiale kan instantieres vha. CG-disambiguerede funktionstags. Polysemiresolutionen kan baseres dels på eksisterende lav-niveau-information (morfosyntaks), dels på direkte CG-regler der forbyder/selekterer bestemte betydninger (eller direkte: oversættelser) i bestemte kontekster (Bick, 2000). Men allerede den eksisterende syntaktisk-funktionelle analyse og muligheden for kontekstuel disambiguering muliggør maskinoversættelsesapplikationer der byder på mere end simple ord-for-ord leksikonopslag eller streng-matching. Fx tillader syntaktiske objektmarkører som @ACC (direkte objekt) og @DAT (indirekte objekt) en korrekt kasus-markering i deklinerende målsprog, selvom dansk ikke selv markerer akkusativ og dativ morfologisk i substantiver, og prænominalmarkøren @>N muliggør kongruenspropagering af en sådan kasusmarkering (her: **'-n'**) fra np-syntagmehovedet til artikler og attributer:

(8)	*Barnet*	*så*	*mange*	*rådyr.*	(dansk)
	La infano	*vidis*	*multaj**n***	*kapreolo**jn**.*	(esperanto)

Et andet potentielt anvendelsesområde for CG-progressionen i den danske parser er grammatisk motiveret retskrivningskontrol, som også beskrevet for en svensk CG af Juhani Birn (1999). Her formuleres CG-mappingregler for bestemte fejlkategorier udfra en *kun delvis éntydiggjort* morfologisk analyse, idet en forudgående fuldstændig disambiguering (som den foretages i parserens almindelige syntaktiske analyse) måske ville fjerne nogle af de *korrekte* læsninger pga. af stave- eller grammatikfejl i konteksten. P.t. er kun få "fejlgrammatiske" regler blevet implementeret, der fortrinsvis beskæftiger sig med nominalkongruens. Eksemplet viser resultatet af en mappingregel for genuskongruens i subjektsprædikativer:

(8) **grammatisk stavekontrol**

[5] VISL-projektet er i færd med at opbyggeen tilsvarende bilingual dansk-engelsk leksikon-database.


Site guide
VISL project
Arabic
Danish
· Projects
· Grammar page
· User guide
· Symbols (closed corpus)
· Symbols (free text)
English
Esperanto
French
German
Italian
A) Free input
Enter text to be analysed, search pattern or:
Analysis
Reset
Hans hus er gammel.
spelling checker
Hans hus er gammel neu-sc .
1 regulær fejl
Check fejlen!

Bibliografi:

Bick, Eckhard. 1992. Internet Based Grammar Teaching. I: Christoffersen, Ellen & Music, Bradley (eds.), *Datalingvistisk Forenings Årsmøde 1997 Kolding, Proceedings,* pp. 86-106

Bick, Eckhard. 2000. *The Parsing System 'Palavras' - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Århus: Aarhus Universitetsforlag

Bilgram, Thomas. 1994. *Computerstyret analyse af dansk - En praktisk analyse af en væsentlig kilde til homonymi i dansk, med forslag til kontekstuel disambiguering ved hjælp af constraint-regler* (specialeopgave i lingvistik). Århus: Institut for Lingvistik

Birn, Juhani. 1999. Detecting Grammar Errors with Lingsoft's Swedish Grammar Checker. I: *NODALIDA 1999, Trondheim, Proceedings*

Karlsson, Fred. 1990. Constraint Grammar as a Framework for Parsing Running Text. I: Karlgren, Hans (ed.), *COLING-90 Helsinki: Proceedings of the 13th International Conference on Computational Linguistics,* Vol. 3, pp.168-173

Karlsson, Fred & Voutilainen, Atro & Heikkilä, Juka & Anttila, Arto (eds.). 1995. *Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text,* Berlin: Mouton de Gruyter

Tapanainen, Pasi. 1996. *The Constraint Grammar Parser CG-2*, Helsinki: University of Helsinki, Department of Linguistics, Publications no. 27