DEPENDENSSTRUKTURER
I CONSTRAINT GRAMMAR SYNTAKS
FOR PORTUGISISK

Eckhard Bick

Institut for Lingvistik, Århus Universitet, Nordre Ringgade, DK-8000 Århus C
tel: +45 - 89 422170, fax: +45 - 86 281397, e-mail: lineb@hum.aau.dk

Abstract

The paper presents some syntactic aspects of an automatic grammar- and lexicon-based parser for unrestricted Portuguese text, ultimately intended for applications like corpora tagging, grammar teaching and machine translation. Grammatical rules are formulated in the Constraint Grammar formalism (CG) and focus on disambiguation and robustness. In spite of using a highly differentiated tag set, the parser yields correctness rates - for unrestricted and unknown text - of over 99% for morphology/POS and 97-98% for syntax (where work is still in progress), even when geared to full disambiguation. Among other things, argument structure, dependency relations and subclause function are treated in an innovative way, and the author argues that the structural information content of a "flat" CG-based syntactical description can be augmented in such a way that automatic transformation into traditional tree structures (like in DCG and PSG) is made possible. The parser uses valency and semantical class information from the lexicon, but disambiguation on these levels is still experimental.

The system runs at about 100 words/sec on a 100 MHz Pentium based Linux system, when using all levels. Morphological and POS disambiguation alone approach 1000 words/sec.

1. Oversigt

I denne artikel præsenteres en morfologisk-syntaktisk parser for fri portugisisk tekst, hvor der anvendes Constraint Grammar til disambiguering af ikke kun ordklasser og morfologiske tags, men også dependens- og valensforhold, samt ledsætningers funktion. Parseren er udviklet som led i min Ph.D.-forskning om automatisk analyse af portugisisk. Projektet har en leksikografisk baggrund (beskrevet i mit cand.mag.-speciale) og et maskinoversættelses-perspektiv, men i det følgende vil det være det syntaktiske niveau, der står i forgrunden. Især vil jeg diskutere de særlige egenskaber ved en funktionelt udrettet, "flad" dependensgrammatik, og vurdere mulighederne for en automatisk transformation til egentlige træstrukturer. Selv om arbejdet med parseren ikke er afsluttet endnu, vil jeg forsøge en evaluering af nogle foreløbige kvantitative resultater. Endeligt skal en række eksempelsætninger og omfattende taglister gøre det muligt for læseren selv at vurdere parserens notationelle koncept i forhold til andre systemer.

2. Baggrund

De fleste ord i natursprogstekster er - isoleret set - flertydige med hensyn til ordklasse, bøjning, syntaktisk rolle, 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), som den er udviklet af Helsinki-skolen (fx. Karlsson et.al., 1995) 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. I selve parseren bliver reglerne kompileret til et computerprogram, der som input tager tekst hvor hvert ord har fået tilføjet tags for alle dets mulige morfologiske og ordklasse-læsninger af en leksikon-baseret tagger. Som output leveres for hver ordform kun én tag-linie, med den korrekte grundform, ordklasse m.m.

(1)
```
"<nunca>"
        "nunca" ADV
    "<como>"
        "como" <rel> ADV
        "como" <interr> ADV
        "como" KS
        "como" <vt> V PR 1S VFIN
    "<peixe>"
        "peixe" N M S
    "<$.>"
```

[ADV=adverbium, KS=subordinerende konjunktion, V=verbum, N=substantiv, PR=præsens, S=singularis, M=maskulinum, 1=1.person, VFIN=finit verbum, <rel>=relativum, <interr>=interrogativum, <vt>=monotransitiv]

De fire læsninger af ordformern *'como'* kaldes i CG-terminologien en *kohorte*. En typisk CG-regel til disambiguering af denn flertydighed er fx. følgende:

(2) @w =! (VFIN) (NOT *-1 VFIN) (NOT *1 VFIN)
[vælg (=!) for enhver ordform (@w) læsningen VFIN (finit verbum) hvis der ikke (NOT) - hverken til venstre (*-1) eller til højre (*1) - findes et andet ord der kan være VFIN.][1]

Ved først at tilføje ("mappe") alle[2] mulige syntaktiske funktioner til ordformen udfra dens ordklasse, bøjning m.m., og herefter at disambiguere denne syntaktiske flertydighed, kan Constraint Grammar også bruges til syntaktisk parsing, som det fx. er sket i Bank-of-English-projektet (200 millioner ord, Järvinen, 1994).

(3)
```
"<nunca>"
    "nunca" ADV @ADVL
"<como>"
    "como" <vt>  V PR 1S VFIN @FMV
"<peixe>"
    "peixe" N M S @SUBJ @ACC @SC @OC
```

[@ADVL=adveerbial, @FMV=finit hovedverbum, @SUBJ=subjekt, @ACC=akkusativobjekt, @SC=subjektprædikativ, @OC=objektprædikativ]

Tilføjelsen af de mulige syntaktiske tags (@) har i eksemplet resulteret i firedobbelt syntaktisk ambiguitet for *peixe*. Læsningen som direkte objekt (@ACC) kan udvælges positivt med en =! -regel der udnytter verbets transitivitet, men den kan lige så godt fremstå indirekte[3], - ved at være den sidste overlevende læsning, efter at CG-regler har forkastet de andre:

(4) @w =0 (@SUBJ) (0 N) (NOT *-1 V3) (NOT *1 V3)
[forkast (=0) subjektlæsningen hvis ordet (0) er et substantiv (N) og der ikke findes et verbum i 3. person]

@w =0 (@SC) (NOT *-1 <vK>) (NOT *1 <vK>)
[forkast subjektprædikativlæsningen (@SC) hvis der ikke findes et kopulaverbum (<vK>) i sætningen]

@w =0 (@OC) (NOT *-1 @ACC) (NOT *1 @ACC)
[forkast objektprædikativlæsningen (@OC) hvis der ikke findes et direkte objekt (@ACC) i sætningen][4]

[1] Reglen er forenklet, idet den forudsætter at enhver periode indeholder mindst ét finit verbum, hvad der ikke altid er tilfældet i overskrifter, udråb o.l. Reglen kan gøres mere sikker ved at kræve et punktum (*1 PUNKTUM) eller udnytte den mulige valensrelation mellem det transitive *comer* og den 'sikre' NP *peixe* (0 <vt>) (1C NP).

[2] Også i mapping-fasen anvendes constraint-regler, og listen over mulige syntaktiske funktioner for et bestemt ord kan således gøres kontekst-afhængig (og dermed kortere).

[3] Det er denne indirekte disambiguering, der er mest karakteristisk for Constraint Grammar, og her ligger en vigtig årsag til metodens robusthed: selv sjældne eller ufuldstændige konstruktioner vil få mindst én analyse - nemlig den der overlever flest forbudsregler. Parseren foretrække således som regel en struktur, der er "næsten rigtig" frem for en, der er "temmelig forkert".

[4] Alle anførte regler gør brug af *"ubundne"* kontekstbetingelser:
*-1 = kontekstbetingelsen søges opfyldt fra og med det 1. ord til venstre (et eller andet sted til venstre)
*1 = kontekstbetingelsen søges opfyldt fra og med 1. ord til højre (et eller andet sted til højre)
Man kan også bruge *"bundne"* kontekstbetingelser, fx -2 = andet ord til venstre, 3 = tredje ord til højre. De "bundne" kontekstbetingelser kan i princippet gengives som n-gram-regler (som brugt i mange probabilistiske parsere), mens de "ubundne" (*-kontekster) er mere CG-

CG-grammatikker er først og fremmest blevet beskrevet for engelsk (fx. Karlsson et.al., 1991), men der findes - i hvert fald på det morfologiske niveau - projekter for flere andre sprog fra såvel den germanske, romanske og finno-ugriske sprogfamilie (svensk, tysk, fransk, finsk m.m.). En moden CG-grammatik for det morfologiske niveau (ordklasse-disambigueringen m.m.) består typisk af 1.000-2.000 regler. For engelsk opgives fejlprocenter på under 0.3% ved en disambigueringsgrad på 94-97% (Voutilainen, 1992).

3. "Flade" træstrukturer i CG-syntaks

3.1 Syntaktisk form og syntaktisk funktion

I min parser, som i andre Constraint Grammar-systemer, benytter jeg mig af en "flad" repræsentation af syntaktisk struktur. Beskrivelsen indeholder information om både *syntaktisk funktion* (fx argumenter som @SUBJ, @ACC) og konstituentstruktur (*syntaktisk form*). Den sidste bliver markeret ved hjælp af dependensmarkører (<, >) som er rettet mod det pågældende syntagmes hoved og samler konstituenten til en kohærent helhed med implicitte syntagmegrænser. Hvor hovedet ikke er hovedverbet, bliver det anført ved pilespidsen (fx N for nominal-hoved, A for adjekt-hoved[5]). Dependensmarkører bliver enten hæftet til de funktionelle tags (fx @<SUBJ, @ADVL>, @N<PRED), eller står, ved visse bestemmerled, alene (fx @>N for [bestemmer-] prænominal).

Idet hvert ord således kun behøver at "huske" sin umiddelbare dependensrelation (dvs. hvad det selv er dependent til), kan hele den syntaktiske struktur beskrives *lokalt* (som ordrelateret tag), - som i en uro, hvor den enkelte tråd kun "kender" nøjagtig 2 af uroens mange faste dele: i den ene ende den stang den selv hænger i (hovedet, som dependensmarkøren peger på) og i den anden ende det objekt (eller den stang) der hænger i tråden (dependenten, som dependensmarkøren peger væk fra). Hvis bare man skriver ned for hver del i uroen hvilken anden del den skal hænge i, kan man faktisk godt skære den i stykker og gemme den i en skotøjsæske - den strukturelle information bevares[6].

Jeg vil her skelne mellem 3 hovedtyper af konstituentstrukturer, som efterfølgende eksemplificeres:

---

specifikke.

[5] Ved et adjekthoved forstår jeg kernen i et adjektiv- eller adverbialsyntagme. Også attributivt brugte participier tilhører adjektkategorien.

[6] At den strukturelle information både markeres og processeres lokalt (på ordplan) er faktisk kongstanken i CG's syntaktiske filosofi, og jeg vil i afsnit 3 diskutere fordelene (og ulemperne) ved en sådan "flad" beskrivelse, og vise hvordan selv mere komplekse dependenter (ledsætninger m.m.) kan håndteres på denne måde.

[7] Reglen udsiger, at der i et syntagme ikke kan være to argumenter med samme syntaktiske funktion, med mindre der er tale om koordination. Fx kan hovedverbet i et sætningssyntagme

	a) argumenter	**b) frie adjunkter**	**c) bestemmer-adjunkter**
valens	valensbunden	ikke valensbunden	ikke valensbunden
blokeringsregel[7]	gyldig (dog ikke ved koordination)	ikke gyldig	ikke gyldig
fokuserbarhed	kan fokuseres/isoleres ved kløvning	kan fokuseres/isoleres ved kløvning	kan ikke isoleres ved kløvning

a) Argumentstruktur:

Argument	Hoved	Argument
João PROP **@SUBJ>**	*come* V VFIN **<vt>** @FMV	*carne* N **@<ACC**
João	spiser	kød.
	quer V VFIN **<x>** @FAUX	*jogar* V INF **@#ICL-AUX<**
[Han/hun]	vil	spille
	rico ADJ **<+em>** @FMV	*em* PRP **@A<** *ouro* N **@P<**
	rig	på guld

Dependensforholdet er for argumenternes vedkommende også markeret i trådens øvre ende: hovedet bærer en valensmarkør: en tag som <vt> ('monotransitivt verbum'), for eksempel, "forventer" et direkte objekt (@ACC) et eller andet sted i sætningen[8]. *Rico em ouro* er et eksempel på hvordan beskrivelsen håndterer flere hierarkiske niveauer: præpositionen *em* er hovedet for et præpositionssyntagme (dette markeres på dependenten *ouro* som @P<), men samtidigt selv argument for adjektivsyntagmets hoved *rico* (markeret ved @A<). I modsætning til engelsk og dansk kan et pronominalt subjekt på portugisisk inkorporeres i det finitte verbum (fx. *quer jogar*), og skal derfor beskrives ikke som sætningskonstituent, men som (fakultativ og valensstyret) konstituent i verbalsyntagmet.

b) Struktur med frie adjunkter:

Adjunkt	Hoved	Adjunkt
Ontem ADV **@ADVL>**	*ele* PERS @SUBJ> *veio* V VFIN <ve> @FMV	*muito* ADV @>A *tarde* ADV **@<ADVL**
I går	han kom	meget sent.
Zangada PCP **@PRED>**,	*saíu* V VFIN @FMV	*sozinha* ADJ **@<PRED**
Sur	gik [hun] (ud)	alene.

således kun styre ét akkusativ- objekt. Reglen gælder netop kun for argumenter og ikke for andre - frie - konstituenter (her kaldet for adjunkter).

[8] I en rent syntaktisk sammenhæng anses valensmarkører dog for sekundære i forhold til de primære tags (@-tagsene), og et ord kan have en lang række af (potentielle) valensmarkører, og stadigvæk være syntaktisk éntydigt beskrevet igennem sit @-tag. Kun de (syntaktisk) primære tags *skal* disambigueres på det syntaktiske niveau. Valensmarkørdisambiguering kan dog være et nyttig redskab på et højere analyseplan, hvor der tilsigtes polysemidifferentiering (jf. afsnit 4).

Frie adjunkter er ikke valensstyret og dependensen markeres derfor kun ved dependenten: adjunkt- adverbialer (@ADVL[9]) peger på hovedverbet, og frie (adjunkt-) prædikativer peger på en nominalgruppe (ofte subjektet, der igen kan være inkorporeret i det finitte verbum).

c) Struktur med bestemmer- adjunkter:

Prænominal	Hoved	Postnominal
O DET <art> **@>N** *grande* ADJ **@>N**	*poeta* N M S	*fluminense* ADJ **@N<**
Den store	digter	fra Rio.
	caro ADJ M S	*demais* ADV **@A<**
	dyr	for [dyr]
mais ADV <quant> **@>A**	*interessado* PCP M S	
mere	interesseret	

Bestemmere er de dependenter, der er tættest bundet til syntagmehovedet, og man kan argumentere, at en bestemmers syntaktiske funktion netop kun er 'bestemmer', og at en yderligere præcisering af dens funktionelle rolle (attribut, kvantifikator m.m.) allerede fremgår af dens ordklasse- tag og de leksemspecifikke semantiske træk. Jeg nøjes derfor med en ren dependensmarkering (dvs. uden funktionel tag på pilens "åbne" side).

3.2 Syntaktisk informationsindhold og ækvivalente træstrukturer

Spørgsmålet er nu, om det på den ene eller anden måde er muligt, at omdanne den beskrevne flade struktur til en træstruktur, som man kender den fra mere "traditionelle" beskrivelser (DCG[10], PSG[11] m.m.). Og, vil der i så fald være tale om informationstab, eller, modsat, øget ambiguitet?

Begge dele, forekommer det mig. På den ene side ville det koste informationsindhold, at erstatte funktionelle og dependens- tags med en ren konstituentstruktur ("bracketing")[12], især for verbalkomplementeringen og på ledsætningsniveau, hvor mit system er mest differentieret (funktion *og* dependens markeres), og i noget mindre grad ved ad- N- og ad- A- konstruktionerne (kun dependens markeres, og argument- /adjunkt- skellet ekspliciteres normalt kun igennem syntagmehovedets valenstags, og ikke igennem et funktionelt tag for det dependente led, jf. 3.1 c).

På den anden side vil nogle tilfælde af underspecifikation af dependensforhold blive synliggjort, når der opbygges en eksplicit træstruktur, som fx:

i) @N< efter postnominal der selv indeholder et nominalt hoved:

... o gigante Venceslau comedor de gente ***famoso*** *...*

[9] valensbundne circumstantielle adverbialer tagges som @ADV (adverbialobjekt), og præpositionssyntagmer, der ikke kan erstattes med rene adverbier, tagges som @PIV (præpositionalobjekt).

[10] Definite Clause Grammar

[11] Phrase Structure Grammar

[12] Selvfølgelig kan man så vælge - når der skal transformeres fra "flad" til træ- beskrivelse - at "berige" den klassiske phrase- structure- notation ved at bevare de funktionelle og dependens- markører fra CG- notationen.

@>N @NPHR @N< @N< @N< @P< @N<

I denne sætning kan *famoso* syntaktisk relateres til:

- *gente* (forbydes dog af genus- kongruenskrav)
- *comedor* (den rigtige løsning udfra princippet om syntagmekohæsion eller "close attachment")
- *Venceslau* (usandsynligt hoved for et adjektiv- postnominal, fordi der er tale om et egennavn)
- *gigante* (kernen i den hierarkisk øverste NP- parentes i udtrykket)

ii) koordinationsproblemer, fx enkelt- vs. dobbelt- tilhæftning af et @N< (postnominal):

... cinco homens e quatro mulheres ***do Rio*** *...*
(@>N @NPHR @CO @>N @NPHR) @N< @P<
eller @>N @NPHR @CO (@>N @NPHR @N< @P<)

iii) @<PRED (prædikativ) efter mellemstående nominale grupper (som fx. @<ACC):

... viu o amigo, ***exausto*** *... - não viu nenhuma solução, exausta*
@FMV @>N @<ACC @<PRED @ADVL> @FMV @>N @<ACC @<PRED

Her er den første sætning flertydig ved han- køns- subjekt, hvor prædikativet kunne pege på både subjektet eller objektet; den anden sætning er ligeledes syntaktisk flertydig (denne gang ved hun- køns- subjekt), men *ville* kunne udredes semantisk (løsninger kan ikke være udmattede).

Som det fremgår, ville mulige løsningsstrategier involvere fx. kongruens [i (i), i det mindste for portugisisk], minimal attachment [eller minimal coordination], og semantisk kongruens mellem hoved og bestemmer [i (iii)]. Samtidigt er det dog vanskeligt at se hvordan nogen som helst primært syntaktisk beskrivelse skulle kunne håndtere denne type ambiguitet, - hvorfor den "flade" beskrivelses "elegante underspecifikation" måske endda burde betragtes som den bedste eller i hvert fald mest pragmatiske løsning.

3.3 Automatisk transformation

Følgende skridt kan udgøre skelettet af en algoritme til opbygning af konstituenttræer udfra en "flad" dependensbeskrivelse som den er foreslået i min parser:

1. Alle adnominaler (@>N, @N<) og adverbielle adjekter (@>A, @A<) hæftes til deres hoveder, defineret som det første ord af egnet ordklasse, der mødes i den af dependensmarkørerne (>, <) angivne retning. De i skridt (1) etablerede syntagmer vil herefter flyttes og koordineres som helheder [skridt (4) og (2)].

2. Koordinatorer betragtes som intra- syntagmatisk, hvor et skridt- 1- element skal krydse dem for at finde et hoved af egnet ordklasse. Ellers koordineres funktionsækvivalente ord/synagmer på det højest mulige niveau der ikke bryder uniqueness princippet.

3. Ledsætningsgrænser introduceres, hvor 2 ukoordinerede argumenter kommer i konflikt pga. blokeringsreglen, og mellem dependensmarkører (af samme niveau) der peger i modsatte retninger.

4. Argumenter og adjunkter på (led)sætningsniveau (hvad enten de er ord, syntagmer eller hierarkisk laverestående ledsætninger) hæftes til nærmeste @MV (hovedverbum) i den udpegede retning, medmindre der er en mellemstående ledsætningsgrænse - i dette tilfælde vælges det førstkommende @MV efter den anden (komplementære) ledsætningsgrænse, osv.

Den gradvise introduktion af hierarkiske parenteser (eller træforgreningssektioner) kunne tænkes at foregå som i følgende eksempel:

a) rå "flad" notation:

```
O pai de o menino que venceu comprou dez cervejas.
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>- @#FS-N<- @SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC
```

b) tilhæfting af prænominaler:

```
(O pai ) de ( o menino) que venceu comprou (dez cervejas).
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>- @#FS-N<- @SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC
```

c) tilhæftning af postnominal relativ og etablering af PP:

```
(O pai ) ( de (( o menino) que venceu )) comprou (dez cervejas).
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>- @#FS-N<- @SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC
```

d) afsluttet dependensbeskrivelse af sætningens førest NP:

```
((O pai ) ( de (( o menino) que venceu ))) comprou (dez cervejas).
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>- @#FS-N<- @SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC
```

e) (led)sætningsgrænser (markeret ved '-') introduceres pga. sammenstød af ukoordineret 2x @SUBJ> og 2x @FMV (uniqueness princip), samt subjekt/objekt- tilhæftning:

```
(((O pai ) ( de (( o menino) - (que venceu)- ))) comprou (dez cervejas)).
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>- @#FS-N<- @SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC
```

Jeg har skrevet et Perl[13]-program, der realiserer nogle af disse principper[14] og transformerer den flade ordbaserede syntaktiske struktur til en mere traditionel træstruktur, hvor konstituenthierarkiet og syntagmegrænserne synliggøres, og hvor de komplekse konstituenter med deres form og funktion indføjes som selvstændige størrelser imellem ordene. Nedenstående en analyseret sætning før og efter transformationen:

(5) anlyseret tekst, i "flad", ordbaseret CG-notation:

[13] Et programmeringssprog, der er særligt velegnet til strengmanipulation og sproglige opgaver.

[14] Programmet er eksperimentelt, og opbygger - hvor CG-beskrivelsen underspecificerer - kun én ud af flere mulige træstrukturer.

ordform	*grundform*	*valens & semantik*	*ordklasse & bøjning*	*syntaks*
*a	[a]	<art>	DET F S	@>N
crise	[crise]	<sit>	N F S	@SUBJ>
apura	[apurar]	<vt> <sN>	V PR 3S IND VFIN	@FMV
o	[o]	<art>	DET M S	@>N
paladar	[paladar]	<anost> <fh>	N M S	@<ACC
de	[de]	<sam->	PRP	@N<
o	[o]	<-sam> <art>	DET M S	@>N
consumidor	[consumir]	<DERS -or>	N M S	@P<
e	[e]		KC	@CO
valoriza	[valorizar]	<vt> <sN>	V PR 3S IND VFIN	@FMV
o	[o]	<art>	DET M S	@>N
dono	[dono]	<H>	N M S	@<ACC
de	[de]		PRP	@N<
restaurante	[restaurante]	<inst>	N M S	@P<
que	[que]	<rel>	SPEC M/F S/P	@SUBJ> @#FS-N<
pilota	[pilotar]	<vt> <vH>	V PR 3S IND VFIN	@FMV
a	[a]	<art>	DET F S	@>N
própria	[próprio]	<jn>	ADJ F S	@>N
cozinha	[cozinha]	<ejo>	N F S	@<ACC

[**ordklasser:** DET=determiner, N=noun, V=verb, PRP=preposition, KC=coordinating conjunction, SPEC=specifier-pronoun, ADJ=adjektiv; **bøjning:** S=singular, P=plurar, M=male, F=female, PR=present, 3S=third person singular; **derivation:** <DERS -or>=suffiksderivation på '-or'; **syntaks:** @>N=prenominal, @SUBJ>=subject, @FMV=finite main verb, @<ACC=accusative object, @N<=postnominal, @P<=argument of preposition, @CO=coordinator, @#FS-N<=finite subclause functioning as postnominal; **valens:** <art>=article, <rel>=relative, <vt>=monotransitive verb; **semantik:** <H>=human, <sit>=situation, <ejo>=functional place, <inst>=institution, <anost>=anatomical bone structure; **selektionsregler:** <fh>=human feature, <sN>=has non-human subject, <vH>=has always human subject, <jn> has non-human head; **ortografi:** <sam->&<-sam>=first and second part of fused expression]

(6) samme tekst, efter transformation til træstruktur, med indføjede syntagme- tags og hierarkisk indrykning:

@SUBJ>:np		
|- @>N:DET F S	*a	[a] <art>
|- @H:N F S	crise	[crise] <sit>
@FMV:V PR 3S IND VFIN	apura	[apurar] <vt> <sN>
@<ACC:np		
|- @>N:DET M S	o	[o] <art>
|- @H:N M S	paladar	[paladar] <anost> <fh>
|- @N<:pp		
|- @H:PRP	de	[de] <sam- >
|- @P<:np		
|- @>N:DET M S	o	[o] <- sam> <art>
|- @H:N M S	consumid or	[consumir] <DERS - or>
@CO:KC	e	[e]
@FMV:V PR 3S IND VFIN	valoriza	[valorizar] <vt> <sN>
@<ACC:np		
|- @>N:DET M S	o	[o] <art>
|- @H:N M S	dono	[dono] <H>
|- @N<:pp		
| |- @H:PRP	de	[de]
| |- @P<:N M S	restaurant e	[restaurante] <inst>
|- @N<:fcl		
|- @SUBJ>:SPEC M/F S/P	que	[que] <rel>
|- @FMV:V PR 3S IND VFIN	pilota	[pilotar] <vt> <vH>
|- @<ACC:np		
|- @>N:DET F S	a	[a] <art>
|- @>N:ADJ F S	própria	[próprio] <jn>
|- @H:N F S	cozinha	[cozinha] <ejo>

[@H=head, np=noun phrase, pp=prepositional phrase, fcl=finite clause, ':'=separator for function and form]

4. Hvilken slags Constraint Grammar

I princippet er CG en robust disambigueringsfilosofi, der ikke på én gang genererer en hel analyse med en bestemt "lovlig" sætningsstruktur, men accepterer ethvert input og mejsler bort hvad der ikke kan være del af nogen (tilladt) struktur. Her er hverken mejsleteknikken (regelsættet) eller mejsleredskaberne (regel- compilerne) bestemt af CG- idéen som sådan. Hver billedhugger bestemmer selv hvordan han vil forme sit værk. Og dog ...

Historisk set udspringer CG fra morfologisk analyse, de fleste systemer benytter sig af en morfologisk toniveau- analyse (TWOL, jf. Koskenniemi,

1983) som præprocessor, og fokuserer på morfologiske træk og ordklasser. Den grammatiske beskrivelse er derfor i høj grad ordbaseret og implementeres ved at hæfte tags til ordformer. "Flad" syntaks er en naturlig konsekvens af dette. Men uden specielle dependensforbindelser kan en sådan flad beskrivelse kun fungere tilfredsstillende, hvor et enkelt ord bærer hele vægten af et syntagmes funktion. Der vil uvægerligt være problemer med dependensforhold der involverer flere forskellige syntaktiske niveauer. Således løber en CG-beskrivelse uden (funktionelle) ledsætningstags ind i vanskeligheder som følgende:

? (Led)sætningsgrænser, selv hvor de synliggøres, er ikke udlagt hierarkisk, hvorfor der kan være problemer med uklare sætningstilhørsforhold (fx efter indskudte relativsætninger).

? Visse valenstræk "udfyldes" ikke altid, som fx. i tilfælde af "manglende" subjekt på engelsk *('Visiting the Louvre was not his only reason for coming to Paris')*, eller manglende akkusativ-objekter ('that/que/at'-sætninger efter "kognitive" verber).

? "Overskydende" argumenter pga. uklare tilhørsforhold mht. sætningshierarki, som *i 'O perigo de os inimigos atacarem à noite era imanente.'*, hvor både *perigo* og *inimigos* er subjekter, noget der er imod blokeringsregel, og kun kan løses ved at løfte det andet subjekts hovedverbum *(atacarem)* ud af matrixsætningen og beskrive det som (ledsætnings-)argument til den forudgående præposition 'de'.

? Nedsat informationsindhold i sammenligning med en træstruktur (jf. ovenfor).

Jeg mener at det ved at skelne mellem CG som disambigueringsteknik på den ene side, og den udmejslede grammatiske beskrivelse på den anden side, er muligt at skabe en form for flad repræsentation der er funktionelt ækvivalent til træstrukturer og som kan håndtere argument- og valensstrukturer på en hierarkisk måde.

Min metode har været (a) at forsyne *alle* de syntaktiske tags med "rettede" dependensmarkører (jf. ovenfor), og (b) at hæfte 2 tags til de centrale forbinderord ("complementizer" som: subordinerende konjunktioner, relativer og interrogativer) i finitte og absolutte ledsætninger, samt til infinitiver, gerundier og participier i infinitte ledsætninger[15]. Disse ord vil så bære både en "indadvendt" tag (@...) der beskriver deres funktion i ledsætningen, og en "udadvendt" tag (@#...) der beskriver ledsætningens egen ledfunktion i sætningens dependenshierarki. Teknisk set håndteres @-tags og @#-tags som to adskilte lister, således at "indadvendte" og "udadvendte" tags kan disambigueres uafhængig af hinanden, af distinkte regelmoduler.

(7)

Sabe	[saber] <vq> V PR 3S IND	@FMV
que	[que] KS	**@#FS-<ACC** **@SUB**
os	[o] <art> DET M P	@>N
problemas	[problema] N M P	@SUBJ>

[15] En anden metode til funktionel tagging af ledsætninger beskrives af Voutilainen (1994). Her er det hovedverbet, der bærer ledsætningens tag (...@), mens dependensforholdene gøres mere eksplicitte ved at indsætte markører for ledsætningsgrænser, og ved at skelne mellem argumenter af henholdsvis finitte og infinitte verbaler.

são	[ser] <vK> V PR 3P IND	@FMV
graves	[grave] ADJ M/F P	@<SC
$.		

[@FMV = finite main verb, @#FS-<ACC = finite subclause, functioning as direct (accusative) object attached to a main verb to the left, @SUB = subordinator, @>N = prenominal modifier, @SUBJ> = subject for a main verb to the right, @<SC = subject complement for a (copula) verb to the left, V = verb, KS = subordinating conjunction, DET = determiner, N = noun, ADJ = adjective, PR = present tense, IND = indicative, 3S = third person singular, 3P = third person plural, M = male, F = female, S = singular, P = plural, <art> = article, <vq> = cognitive verb, <vK> = copula verb]

5. Et teleologisk bedømmelsesperspektiv

Når man sammenligner forskellige syntaktiske beskrivelser, udgør informationsindhold og konstituentstruktur kun to af de mulige bedømmelsesperspektiver, og begge må ses i lyset af et bestemt teoretisk bagland, som fx funktionel (FG) eller generativ grammatik. Det kan imidlertid være interessant at se på hvilke praktiske anvendelser en bestemt beskrivelsesmodel retter sig imod.

Her er mit eget perspektiv maskinoversættelse (MT), og aspekter som de følgende vil derfor få tillagt særlig vægt:

? Detaljerede ordfølgeuafhængige funktionelle tags gør det nemmere at transformere kildesprogsstruktur (SL) til målsprogsstruktur (TL), uden at skulle introducere for mange indviklede transformationsregler. Således kan den danske hovedsætnings- ordfølge SVO med subjekt i forfeltet etableres direkte på trods af den mere frie portugisiske ordstilling:

(8a) O rei @SUBJ> queria @FMV mais terra @<ACC.
(8b) Queria @FMV o rei @<SUBJ mais terra @<ACC.
(8c) -> kongen @SUBJ> ønskede @FMV mere land @<ACC.

? Det er af stor betydning for polysemidifferentieringen at vide, hvilket af et ords potentielle valensmønstre der er blevet realiseret i en given (led) sætningskontekst, og hvilken semantisk klasse udfylder en given valensplads (slot). I denne forbindelse får valenstags (og selektionsrestriktioner) betydning ikke kun som *sekundære* tags (som udelukkende bruges til at disambiguere morfologiske/syntaktiske tags), men også som selvstændige *primære* tags, der kan og skal disambigueres:

(9aa) rever <vt> 'gense' realiseret valens: transitiv <vt>
(9ab) rever <vi> 'sive igennem' realiseret valens: intransitiv <vi>
(9ba) revista <+n><rr> 'avis' realiseret valens: titel <+n>, semantisk klasse: læsestof <rr>
(9bb) revista <CP> 'inspektion' realiseret semantisk klasse: +CONTROL, +PERFEKTIV

? De ovenfor omtalte problemer med underspecifikation af postnominaler, koordination og frie nominaladjunkter bliver til et gode, når man betragter dem udfra et MT-perspektiv: - for det første er mange af disse tilfælde eksempler på "ægte flertydighed", der kun kan tydes af den fuldt kontekstualiserede - menneskelige - lytter/læser (og under alle omstændigheder er der tale om ægte *syntaktisk* flertydighed). - Og for det andet er en række af disse strukturelle ambiguiteter (især koordination (11a) og "kort" (10b) vs. "lang" (10a) tilhæftning af postnominale præpositionssyntagmer) forholdsvis universelle, dvs. sproguafhængig, således at de kan bevares i oversættelsen, der baseres direkte på den "flade" beskrivelse (10c).

(10a) Han hentede **((**manden @<ACC med @N< cyklen @P<**)** fra @N< Kina @P<**)**.
(10b) Han hentede **(**manden @<ACC med @N< **(**cyklen @P< fra @N< Kina @P<**))**.
(10c) Foi buscar o homem @<ACC com @N< a bicicleta @P< de @N< a China @P<

At gøre en sådan flertydighed eksplicit (for et sprogpar der ellers håndterer den éns) ville kun belaste oversættelsesmodulet med irrelevant ballast. Adjektiviske bestemmere, enten postnominal eller som frie adjunkter, er derimod mere problematiske, idet der kan være kongruensrelationer (11b) mellem hoved og bestemmer:

(11a) gifte @>N kvinder @NPHR og @CO mænd @NPHR
(11b) homens @NPHR e @CO **mulheres** @NPHR casad **as** @N<

6. Statistisk evaluering

For at kunne afprøve nye og kontrollere gamle regler i min parser har jeg udarbejdet et "bench mark"- corpus (i alt ca. 33.000 ord), hvor der for hver flertydige kohorte markeres med en <Correct!> -tag hvilken læsning der er korrekt. Pga. de mange gentestninger har reglerne efterhånden kunnet opnå fuld disambiguering og fejlprocenter på under 0.1% for disse arbejdstekster. For ukendt tekst er tallene selvfølgelig lavere; alligevel er resultatet ikke irrelevant. Det viser nemligt, at CG-metoden ikke lider under systemimmanente interference-problemer i samme grad som fx. en probabilistisk tagger baseret på en ren trigram-HMM[16], hvor der (så vidt jeg ved) selv ved gentræning og -måling på samme corpus sjældent opnås fejlprocenter på under 3%, end ikke for ordklasse-tags[17].

[16] Hidden Markov Model, hvor de mulige sætningsanalyser udtrykkes som (oftest ordklasse-) tagsekvenser og siden vurderes for deres respektive sandsynlighed: at en ordform skulle bære en given tag beregnes som produktet af a) den leksikale sandsynlighed (ord/ordklasse) og b) n-gram-sandsynligheden (for bigrammer fx. ordklasse$_n$/ordklasse$_{n-1}$), og hele sekvensen sandsynlighed igen er produktet af de "individuelle" sandsynligheder for de i sekvensen realiserede tags.

[17] I en probabilistisk tagger vil "manuelle" indgreb (håndlavede regler, bias eller priming), designet til at håndtere uregelmæssigheder eller sjældne strukturer, ofte resultere i skadelige interferencer, fordi de probabilistiske regler er "majoritetsdrevne", og en lille "gevinst" for minoritetstilfældene vil tit føre til tilsvarende større "tab" mht. majoritetstilfældene, idet

For at opnå maksimal præcision, har jeg også arbejdet med et større utagget tekstmateriale (170.000 ord fra Borba- Ramsey- corpuset[18]), både på det morfologiske og det syntaktiske niveau. Dette var muligt, fordi *precision* (defineret *som overlevende korrekte læsninger : overlevende læsninger i alt*) kan approksimeres ved at nedbringe ambiguiteten, i hvert fald så længe lejlighedsvise bench mark- kørsler sikrer at nye regler kun forkaster få korrekte tags, og så længe ambiguiteten stadigt er høj. Ambiguiteten kan så måles nemt med automatiske midler (fx. programmet grep) på en hvilken som helst tekst. Derimod kan *recall* (defineret *som overlevende korrekte læsninger : alle korrekte læsninger*) kun kvantificeres ved optælling i mindre testtekster (der findes mig bekendt ikke noget stort analyseret portugisisk corpus til sammenligning). Indstiller man parseren til fuld disambiguering (hvor der med undtagelse af de få tilfælde af ægte ambiguitet kun er én overlevende læsning per ordform), kan man her betragte recall tallene som et direkte mål for parserens præstation, og jeg vil i det følgende bruge det mere generelle udtryk *correctness* i betydningen af *recall ved 100% disambiguering*.

En optælling af fejltyperne under test- kørslen af en mindre ("ukendt") prosa- tekst på ca. 2.500 ord ("O tesouro" af Eça de Queiroz) gav følgende resultat:

fejl i:	*antal fejl:*	
ordklasser	16	
grundformer	1	
Alle morfologiske	**17**	**(99.3 %correctness)**
verbalfunktion	3	
verbers argumenter	25	
præpositioners argumenter	2	
Argumentstruktur	**30**	
bestemmere	13	
Bestemmere	**13**	
adjunkter	11	
Adjunkter	**11**	
finitte ledsætninger	6	
infinitte ledsætninger	3	
absolutte ledsætninger	1	
Ledsætninger	**10**	
Alle syntaktiske	**64**	**(97.4 %correctness)**
"lokale" syntaktiske fejl pga. morfologiske/ordklasse- fejl	-27	
Rent syntaktiske	**37**	**(98.5% correctness)**

opprioriteringen af undtagelserne går ud over de "normale" statistiske regler (jf. Chanod & Tapanainen, 1994).

[18] Corpuset indeholder mest brasiliansk materiale, og er i alt på 5 millioner ord. Over 600.000 ord er offentliggjort på CD som led i ECI-projektet (European Corpus Initiative).

Man kunne formode at fejlene var fordelt jævnt over hele teksten, hvad der - ved en gennemsnitlig sætningslængde på 15 ord - ville svare til en "fejltæthed" af ca. 1 morfologisk fejl i hver tiende sætning, og en syntaktisk i hver tredje. Dette er imidlertid ikke tilfældet. Fejlene optræder ofte i grupper: indlysende nok, vil de fleste ord med ordklassefejl også kunne findes på listen over syntaktiske fejl, og mange syntaktiske fejl vekselvirker med læsninger i naboordene, pga. regler der involverer sætningsgrænse- ord, uniqueness- princippet osv. Således kan en N-V-ordklassefejl afføde 2 eller 3 syntaktiske fejl omkring sig. Denne "ophobningstendens" for syntaktiske fejl har en gavnlig sideeffekt på parserens robusthed (mange sætninger er således helt fejlfrie), og letter desuden grammatikerens arbejde: en korrektur ét sted kan "helbrede" en hel kæde af sekundære interferens- fejl. Fejlinterferencen betyder også at den syntaktiske parser alene, dvs. når den forsynes med morfologisk fejlfri tekst som input, kan opnå endnu bedre resultater (forskellen er typisk på 0.5- 1 procentpoint).

For at undersøge, om fejlprocenterne varierer i afhængighed af teksttypen, har jeg også testet parseren på aktuelle avistekster[19] (VEJA-magasinet). Der er igen tale om (for parseren) ukendt, løbende tekst. Artiklerne repræsenterer henholdsvis underholdnings- og kunst- genrerne.

Tekst:	**"VEJA" (videogames)** 2412 ord		**"VEJA" (kunst)** 1837 ord		**ialt** 4249 ord	
Fejltyper:	antal fejl	% korrekt	antal fejl	% korrekt	antal fejl	% korrekt
Morfologi (alle)	29	**98.8 %**	7	**99.6 %**	36	**99.2 %**
ukendte engelske ord i overskrifter	- 10 - 3		- 1 - 0		- 11 - 3	
Morfologi (ren)	16	**99.3 %**	6	**99.7 %**	22	**99.5 %**
Syntaks (alle)	66	**97.3 %**	46	**97.5 %**	112	**97.4 %**
syntaks pga. morfologi	- 37		- 7		- 44	
Syntaks (ren)	29	**98.8 %**	39	**97.9 %**	68	**98.4 %**

En nærmere gennemgang af fejltyperne viser, at de valgte avisteksterne adskiller sig fra fiktionsprosa både leksikalsk og syntaktisk. For det første møder man en stor andel af komplekse egennavne (fx. 'Massachusets Institute of Technology'), forkortelser ('MIT') og engelske modeord (således er det ét enkelt ord, *console*, der - brugt som ukendt engelsk substantiv ['spillekonsol'], og ikke som portugisisk verbum ['trøster'] - tegner sig for en tredjedel (!) af fejlene i teksten om video- spil). For det andet er teksterne - på det syntaktiske plan - meget rige på frie prædikativer (typisk oplysninger

[19] Tal for yderligere 2 avistekster fra VEJA (genremæssigt placeret indenfor politik og sundhed), viser nogenlunde de samme fejlprocenter (jf. Bick, 1996).

om personer, institutioner eller forkortelser, som alder, sted, definition m.m.) og indskudte "overflødige" finitte verber i form af citationsrammer.

Fejlprocenterne skal desuden ses i lyset af det meget differentierede tag-set (jf. 7.1). Således kan parserens detaljerede dependens- og funktionsoplysninger for præpositional-syntagmerne (som fx. post-nominal @N<, adverbielt postadjekt @A<, adverbielt adjunkt @<ADVL, @ADVL>, @ADVL, adverbielt objekt @<ADV, @ADV>, præpositionelt objekt @<PIV, @PIV>, subjektsprædikativ @<SC, frit prædikativ, @<PRED, argument for forbinderled @AS<) give anledning til en lang række potentielle "indbyrdes" fejl, der ville være "usynlige" i en beskrivelse, der smelter disse tags sammen til en simpel "syntagmatisk" tag 'PP' (præpositionssyntagme), eller et rudimentært "funktionelt" 'ADVL' (adverbial). Indbyrdes "forvekslinger" inden for PP-gruppen står således for 15 tilfælde, eller hele 22%, af de 68 rent syntaktiske fejl i VEJA-teksterne.

7. Parseren

7.1 Tag-sættet

Parserens tag-sæt indeholder 13 ordklasse-kategorier, der kombineres med 24 tags for bøjningsformer, ialt flere hundrede distinkte komplekse tags. I tag-linien 'V PR 3S IND VFIN', for eksempel, alternerer ordklassen 'V' således med 12 andre ordklasser, og indenfor V-klassen alternerer 'PR' (præsens) med 5 andre tider, der hver igen findes i 6 forskellige person-numerus former for både 'IND' (indikativ) og 'SUBJ' (konjunktiv). På denne måde beskrives 6x6x2=72 finitte verbalformer ved hjælp af kun 6+6+2=14 deltags. Denne 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"), 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. Det modsatte gælder for numeralia (NUM), mens både genus og numerus er leksemkategorier for propria (PROP), og ordformkategorier for adjektiver (ADJ)[20].

[20] Pronominer kan opdeles efter samme skema, i en determiner-klasse (DET) med de samme (variable) kategorier som adjektiver, og en "specifier"-klasse (SPEC) 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. Pronominer som 'o' og 'este', der både kan forekomme "adjektivisk" og "substantivisk", er efter dette system entydige medlemmer af DET-klassen. Artikel-klassen får heller ikke særstatus: 'o' er altid DET, uanset om det bruges som "artikel", "adjektivisk demonstrativ" eller "substantivisk demonstrativ". Tagsene <art> og <dem> optages på taglisten, men de er *ikke* ordklasse-kategorier, og disambigueres først på et senere tidspunkt (valens-niveauet), til brug ved MT.

Participiet (V PCP), ordklasssernes enfant terrible, er morfologisk markeret som ('-id/-ad'); men udenfor verbalkæden overtager det adjektivets ordformkategorier, og parseren vælger i dette tilfælde at "fusionere" PCP/ADJ-ambiguiteten: <ADJ> V PCP.

Det syntaktiske tag-sæt råder over 40 tags for ord/syntagme-funktion og ca. 30 tags for sætningsfunktion (der dækker over tre slags ledsætninger: finitte, infinitte og absolutte [=verballøse]). Også her er det virkelige antal af distinkte tag-strenge meget højere, fordi det ord der bærer ledsætningens tag, jo også skal markeres for dets ledsætnings-interne funktion.

Systemerne for valens og semantik er under udvikling, og det er derfor vanskeligt at angive nøjagtige tal for tag-sættenes størrelse. Omtrentlige tal er ca. 100 for valensklasser (især for verber), og ca. 200 for semantiske klasser (især for substantiver). De semantiske klasser er baseret på 16 "atomare" træk (som, fx., ±HUM).

7.2 Parserens tekniske data

Den portugisiske parser består af en række programmoduler, der - bortset fra lingsofts sproguafhængige compiler for CG-regler - er skrevet af mig selv i programmeringssprogene C og Perl. Parseren omfatter følgende moduler på det morfologisk-syntaktiske niveau[21]:

- 1. et **morfologisk analyse-program** (beskrevet i Bick, 1995), som behandler orthografisk præprocessering, ordklasse, bøjning, derivation, faste udtryk (polyleksikalier) og inkorporerende verber. Analyse-modulet støtter sig til et håndbygget **leksikon** med 70.000 enheder, der dækker over ca. 50.000 leksemer og udgør en tilpasset elektronisk version af ordbogsmateriale fra forfatterens cand.mag.-speciale om leksikografi (Bick, 1993)
- 2. en **morfologisk disambiguator** med 1700 Constraint Grammar regler
- 3. en **syntaktisk "mapper"** med 400 kontekstbaserede regler der "mapper" (alle mulige) syntaktiske funktioner udfra en ordforms morfologiske/ordklasse-tags
- 4. en **syntaktisk disambiguator** med 1500 Constraint Grammar regler
- 5. en **disambiguator for valens og semantiske klasser** (med 2200 Constraint Grammer regler, eksperimentel)

En fuldstændig grammatisk analyse på alle niveauer håndterer ca. 100 ord/sec på en 100 MHz Pentium-baseret Linux-maskine. Den morfologiske/ordklasse-disambiguering alene opnår hastigheder i nærheden af 1000 ord/sec.

Systemet kan afprøves igennem en interaktiv brugerflade på følgende web-adresse: *http://ling.hum.aau.dk/~eckhard/Linguistics.html.* Større prøvetekster til automatisk analyse (i ISO Latin-1 format) kan også sendes via e-mail til *eckhard@ling.hum.aau.dk* under emnet *portpars* (teksten skal begynde med ordet *parsmail* på første og afsenderens returadresse på anden linie, og afsluttes med ordet *parsslut* på en linie for sig).

8. Perspektiv

[21] Hertil kommer eksperimentelle moduler for portugisisk-dansk MT: polysemidisambiguering, oversættelse af disambiguerede grundformer, portugisisk-dansk syntaktisk transformation og en generator for dansk morfologi.

Parseren kan sammenfattende beskrives som et leksikon- og grammatikbaseret system, der beskriver ord og sætninger med hensyn til både form og funktion, hvor dens notationelle særpræg ligger i dens ordbaserede "flade" gengivelse af syntaktisk struktur. Den bagvedliggende formalisme, Constraint Grammar, har vist sig også for Portugisisk at muliggøre lave fejlprocenter, en høj hastighed samt en meget robust håndtering af fri tekst. Selve metoden synes i øvrigt til en vis grad at være "niveauneutral", idet jeg med success har kunnet anvende den på stadigt "højere" analyse- niveauer: ledsætningsfunktion, valens- og semantisk disambiguering (samt herigennem polysemiresolution).

Systemets formelle og indholdsmæssige egenskaber må formodes at have stor betydning for mulige anvendelsesområder, og jeg vil afsluttende diskutere systemet udfra denne synsvinkel. Nedenstående tabel viser hvilke af parserens egenskaber jeg tillægger betydning ved anvendelsen indenfor bestemte opgaveområder (af hvilke nogle for øvrigt kan afprøves på ovennævnte web- site).

EGENSKABER	**corpus - arbejde**	**grammatisk stavekontrol**	**grammatik - formidling**	**maskin- oversættelse**	**informations - ekstraktion**
indholdsmæssige:					
1. leksikon- baseret		+++		+++	+++
2. beskriver form og dependens	++	++	+++	++	+
3. beskriver funktion	++	+	+++	+++	+++
4. flad syntaks	+++	+	+++	+	+
formelle/metodiske:					
5. lav fejlprocent	+	+++	++	+++	+
6. høj hastighed	+++	++		++	+++
7. robust	+++	++	+	++	++
8. metodemæssigt niveauneutral	+			+++	+++
9. ordbaseret notation	+++	+	++	+	+
REALISERING	alle CG-systemer	engelsk	VISL[22] (portugisisk)	portugisisk - dansk	?

Mens både stavekontrol, maskinoversættelse og informationsekstraktion profiterer af detaljerede leksikalske oplysninger (1), har kun de sidste to applikationer brug for en metode der tillader også en vis semantisk analyse

[22] VISL står for *'visual interactive syntax learning'* og er et CTU- støttet projekt på Odense Universitet (Institut for Sprog og Kommunikation). Projektsprog er engelsk, tysk og fransk, med portugisisk som foreløbig "modelsprog" i opstart- fasen.

(8). Grammatikformidlingen udskiller sig fra de andre felter ved at lægge mere vægt på parserens indholdsmæssige (2, 3) end dens formelle egenskaber, og især tidsfaktoren (6) og robustheden (7) spiller en mindre rolle, idet der typisk vil arbejdes med korte fejlfrie tekster (enkelte sætninger). Den flade ordbaserede notation (4, 9) har den pædagogiske fordel, at kategorier, funktion m.m. kan markeres direkte i den løbende tekst, fx. igennem farvenotation, understregning, sub- /superscript- indices, eller i form af en slags "meta- tekst"[23]- linie. Ved det lingvistiske corpus- arbejde værdsættes ligeledes den flade notation, omend på en anden baggrund: strengsøgningsoperationer lettes betydeligt og gøres mere fleksible. Blandt de formelle egenskaber er corpus- arbejdet og informationsekstraktion de områder der har størst gavn af parserens høje hastighed (6) , mens fejlprocenten (5) er vigtigst for de tekstproducerende - og dermed læserkontrollerede - systemer (stavekontrol og maskinoversættelse).

[23] En horisontal metatekstnotation kunne fx se ud som i følgende sætning:

Having read the letter from Italy she called her Swedish friend immediately.

x:A> v:X< d:>N s:<O prp:N< n:P< pe:S> v:V d:>N adj:>N s:<O adv:<A

[hvor notationen er ordklasse:funktion, og x=hjælpeverbum, v=verbum, d=determiner, s=substantiv, prp=præposition, n=proprium, pe=personligt pronomen, adj=adjektiv, adv=adverbium, A>/A<=adverbial, X<=ikke- første led i verbalkæde, >N=prænominal, <O=direkte objekt, N<=postnominal, P<=styrrelse af præposition, S>=subjekt, V=finit hovedverbum]

Appendiks I: Sætningseksempler

a) halvgrafisk træ- notation med eksplicit syntagme- markering

```
@ADVL>:ap                                *depois          [depois] <+de>
 |- @H:ADV
 |- @A<:pp
   |- @H:PRP                              de               [de]
   |- @P<:np
     |- @>N:NUM F S                       uma              [um] <card>
     |- @H:N F S                          década           [década]
     |- @N<:PP                            em=vigor         [em=vigor] <adj>
$\,
@SUBJ>:np
 |- @>N:DET M S                           o                [o] <art>
 |- @H:N M S                              estilo           [estilo] <ak>
 |- @N<:ADJ M S                           gastronômic      [gastrônomo] <DERS -ico>
                                          o                <jn>
 |- @N<:fcl
   |- @SUBJ>:SPEC M/F S/P                 que              [que] <rel>
   |- @FMV:V IMPF 1/3S IND VFIN           supervaloriz     [supervalorizar] <vt> <vH>
                                          ava
   |- @<ACC:np
   |   |- @>N:DET M S                     o                [o] <art>
   |   |- @H:N M S                        requinte         [requinte] <am>
   |   |- @N<:pp
   |     |- @H:PRP                        de               [de] <sam- >
   |     |- @P<:np
   |       |- @>N:DET F S                 a                [a] <- sam> <art>
   |       |- @H:N F S                    decoração        [decoração] <CP> <ac>
   |- $\,
   |- @<ADVL:pp
     |- @H:PRP                            sob              [sob]
     |- @P<:np
       |- @>N:DET M S                     o                [o] <art>
       |- @H:N M S                        comando          [comando] <s> <CI>
       |- @N<:pp
         |- @H:PRP                        de               [de] <+hum>
         |- @P<:np
           |- @H:N M P                    profissionai     [profissional] <prof>
                                          s
           |- @N<:fcl
             |- @SUBJ>:SPEC M/F S/P       que              [que] <rel> <que- hum>
             |- @ADVL>:ADV                não              [não] <dei> <setop>
             |- @FMV:V IMPF 3P IND        eram             [ser] <vK> <sH>
VFIN
             |- @<SC:pp
             |   |- @H:PRP                de               [de] <sam- > <+top>
             |   |- @P<:np
             |     |- @>N:DET M S         o                [o] <- sam> <art>
             |     |- @H:N M S            ramo             [ramo] <anbo> <stok> <fag>
             |- $\,
```

@FMV:V PR 3S IND VFIN	vive	[viver] <vt> <va+STED> <vH>
@<ACC:np		
|- @>N:DET F S	a	[a] <art>
|- @H:N F S	hora	[hora] <dur> <temp>
|- @N<:pp		
|- @H:PRP	de	[de] <sam- >
|- @P<:np		
|- @>N:DET F S	a	[a] <- sam> <art>
|- @H:N F S	verdade	[verdade] <feat> <ss> <am> <sh>
$.		

b) "flad notation" uden eksplicit syntagme- markering

(For overskuelighedens skyld er grundformerne og de fleste sekundære tags fjernet i eksemplerne; desuden er der foretaget hierarkisk indrykning ved de syntaktiske tags for at vise dependens- markørernes funktion.)

Não	ADV	@ADVL>		
existe	V PR 3S IND <vi>	@FMV		
em	PRP <sam- >	@<ADVL		
a	DET F S <art> <- sam>	@>N		
história	N F S	@P<		
de	PRP <sam- >	@N<		
o	DET <art> <- sam>	@>N		
mundo	N M S	@P<		
ocidental	ADJ M/F S	@N<		
nada	SPEC M S	@<SUBJ		
comparável	ADJ M/F S <+a>	@N<		
a	PRP <sam- >	@A<		
o	DET M S <+rel> <- sam>	@P<		
que	SPEC M/F S/P <rel>	@#FS-N<	@ACC>	
os	DET M P <art>		@>N	
brasileiros	N M P		@SUBJ>	
realizaram	V PS/MQP 3P IND <vt>		@FMV	
em	PRP <sam- >		@<ADVL	
esse	DET M S <- sam>		@>N	
ano	N M S		@P<	
$.				

O	DET M S <art>	@>N			
cientista	N M/F S$	@SUBJ>			
político	ADJ M S	@N<			
diz	V PR 1S IND <+interr>	@FMV			
por	PRP		@ADVL>		
que	SPEC M/F S/P <interr>	@#FS- <ACC	@P<		
os	DET M P <art>		@>N		
brasileiros	N M P		@SUBJ>		
devem	V PR 3P IND		@FAUX		
se	PERS M/F 3S/P ACC <refl>		@ACC>		
orgulhar	V INF 0/1/3S <de^vrp>		@#ICL- AUX<	@IMV	
de	PRP			@<PIV	
$1992	NUM M/F P <card>			@P<	
e	KC	@CO			

pede	V PR 3S IND <vq>	@FMV	
que	KS	@#FS- <ACC	@SUB
tenham	V PR 3P SUBJ <vt>		@FMV
mais	DET M/F S/P <KOMP>		@>N
paciência	N F S <+com>		@<ACC
com	PRP		@N<
Itamar	PROP M/F S/P		@P<
Laurentino	PROP M/F S/P		@N<
Gomes	PROP M/F S/P		@N<
$.			

Sabe	V PR 3S IND <vq>	@FMV		
que	KS	@#FS-<ACC	@SUB	
os	DET M P <art>		@>N	
problema	N M P		@SUBJ>	
s				
são	V PR 3P IND <vK>		@FMV	
graves	ADJ M/F P		@<SC	
$\,				
que	KS	@#FS-<ACC	@SUB	
há	V PR 3S IND <vt>		@FMV	
	<vUK>			
gente	N F S			@SUBJ>
morrendo	V GER <de^vp>		@ICL-<ACC	@IMV
de	PRP			@<PIV
fome	N F S <de+>			@P<
em	PRP <sam- >			@N<
o	DET M S <art> <-			@>N
	sam>			
Brasil	PROP M S			@P<
$.				

Entre	PRP	@ADVL>		
os	DET M P <art>	@>N		
anos	N M P <+num>	@P<		
$60	NUM M/F P	@N<		
e	KC	@CO		
$80	NUM M/F P	@N<		
$\,				
ele	PERS M 3S NOM/PIV	@SUBJ>		
passou	V PS 3S IND <vt+TID>	@FMV		
vinte	NUM M/F P	@>N		
anos	N M P <num+>	@<ADV		
em	PRP <sam- >	@<ADVL		
		@N<		
a	DET F S <art> <- sam>	@>N		
França	PROP F S	@P<		
$\,				
estudand	V GER <vt>	@#ICL-	@IMV	
o		<ADVL		
Ciência	N F S		@<ACC	
política	ADJ F S		@N<	
e	KC		@CO	
História	N F S		@<ACC	
de	PRP <sam- >		@N<	
o	DET M S <art> <- sam>		@>N	
Brasil	PROP M S		@P<	
$.				

Appendiks II: Tag-sæt

ORDKLASSE-TAGS

N	*Substantiver*
PROP	*Propria*
SPEC	*Specifiere* (defineret som ikke-flekterende pronominer, der ikke kan forekomme prænominalt): fx indefinitte pronominer, substantiviske kvantifikatorer og relativer
DET	*Determinere* (defineret som genus-/numerus-flekterende pronominer, der *kan* forekomme prænominalt): fx artikler, adjektiviske
PERS	*Personlige pronominer* (defineret som person-bøjede pronominer)
ADJ	*Adjektiver* (inkl. ordinaler, ekskl. participier, der tagges V PCP)
ADV	*Adverbier* (både 'primære' adverbier og deriverede '-mente'-adverbier)
V	*Verber* (fuldverber, hjælpeverber)
NUM	*Numeralia* (kardinalier)
PRP	*Præpositioner*
KS	*Subordinerende konjunktioner*
KC	*Koordinerende konjunktioner*
IN	*Interjektioner*
EC	Morfologisk "synlige" *affikser* ('elemento composto') (fx "anti-gás")

Såfremt sekundære tags (markeret som <...>) bevares i analysen, vil adberbier (ADV) og pronominer (SPEC, DET, PERS) yderligere blive differentieret i underklasser, af hvilke to (<rel> [relativer] og <interr> [interrogativer]) ofte deles om den samme ordform - og derfor må disambigueres inden for den samme ordklasse. Dette gælder til dels også for <setop> [set-operator] underklassen af adverbier. Andre sekundære tags (fx. valenstags som <vt> for transitive verber) hjælper med at disambiguere de primære (morfologiske og syntaktiske) tags, men disambigueres ikke selv på dette analyseniveau. Som også også for de rent semantiske tags vedkommende (fx. <prof> for professionel) kan en disambiguering her imidlertid være ganske effektiv mht. til en senere polysemiresolution.

FLEKTIONS-TAGS

Genus:	M (maskulinum), F (femininum), M/F [for: N', PROP', SPEC', DET, PERS, ADJ, V PCP, NUM]
Numerus:	S (singularis), P (pluralis), S/P [for: N, PROP', SPEC', DET, PERS, ADJ, V PCP, V VFIN, INF, NUM]
Casus:	NOM (nominativ), ACC (akkusativ), DAT (dativ), PIV (præpositiv), ACC/DAT, NOM/PIV [for: PERS]
Person:	1 (første person), 2 (anden person), 3 (tredje person), 1S, 1P, 2S, 2P, 3S, 3P, 1/3S, 0/1/3S [for: PERS, V VFIN, V INF]

Tid: PR (præsens), IMPF (imperfektum), PS (præteritum), MQP (pluskvamperfektum), FUT (futurum), COND (konditionalis) [for: V VFIN]
Modus: IND (indikativ), SUBJ (konjunktiv), IMP (imperativ) [for: V VFIN]
Finithed: VFIN (finit form), INF (infinitiv), PCP (participium), GER (gerundium) [for: V]

(I denne opstilling betyder " ' ", at den pågældende kategori her er en leksemkategori uden fleksion, mens ingen " ' " betyder, at kategorien for denne ordklasse optræder som ordformkategori, dvs. tillader fleksion.)

<*> Asterisken betyder at ordet har været skrevet med stort (<*>), eller at det har et indledende (<*1>), henholdsvis afsluttende (<*2>) citationstegn.

$ Dollar- tegnet markerer grafiske enheder, der ikke selv er ord, især tegnsætning ('$.' for et punktum, '$\,' for et komma) og tal ('$1947')

SYNTAKTISKE TAGS

@SUBJ> @<SUBJ subjekt
@ACC> @<ACC akkusativ- (direkte) objekt
@DAT> @<DAT dativ- objekt (kun pronominalt, ellers @PIV)
@PIV> @<PIV præpositionelt objekt
@ADV> @<ADV adverbielt objekt (sted, tid, varighed, mængde)
@SC> @<SC subjektprædikativ
@OC> @<OC objektprædikativ
@ADVL> @<ADVL adverbial
(Alle ovennævnte argumenter [@SUBJ, @ACC, @DAT, @PIV, @ADV, @SC, @OC] og de adverbielle adjunkter [@ADVL] refererer til det nærmeste fuldverbum til venstre [<] eller højre [>].)
@ADVL 'fri' adverbialsyntagme (uden styrende verbum)
@NPHR 'fri' nominalsyntagme (uden styrende verbum)
@VOK 'vokativ' (dvs. 'fri' tiltale- proprium i direkte tale)
@>N prænominalt adjekt [typisk DET/ADJ/PCP før N]
(refererer til det nærmeste NP-hoved til højre, der ikke selv er et adnominal)
@N< postnominalt adjekt [typisk ADJ/PCP efter N]
(refererer til det nærmeste NP-hoved til venstre, der ikke selv er et adnominal)
@>A adverbielt præadjekt [typisk ADV før ADV/ADJ/PCP]
(refererer til nærmeste ADJ/PCP/ADV eller attributivt brugt N til højre)
@A< adverbielt postadjekt (sjældent, fx 'caro demais')
@APP referentiel apposition (altid efter NP + komma)
@PRED> 'foranstillet' frit prædikativ
(refererer til efterfølgende @SUBJ, selv hvor denne er inkorporeret i VP'en)
@<PRED 'efterstillet' frit predikativ
(refererer til nærmeste NP-hoved til venstre, *eller* til nærmeste @SUBJ til venstre)
@N<PRED prædikat i 'small clause' indledt med 'com/sem'
(sjælden, fx 'com a mão *na bolsa*', 'sem o pai *ajudando*, não conseguiu')
@P< argument af præposition
@S< sætningskomplement ('não venceu *o que* muito o contrariou')
@FAUX finit hjælpeverbum (cp. @#ICL-AUX<)
@FMV finit hovedverbum
@IAUX infinit hjælpeverbum (cp. @#ICL-AUX<)
@IMV infinit hovedverbum
@PRT-AUX< partikel i verbalkæde (præposition eller "que" efter hjælpeverbum)
@CO koordinerende konjunktion
@SUB subordinerende konjunktion
@KOMP< argument af komparativ (fx "do que" efter 'melhor')
@COM direkte komparator uden forudgående komparativ (fx 'forte *como* um urso')
@PRD rolle- prædikator (fx "arbejde *som*", "fungere *som*")
@#FS- finit ledsætning

(kombineres med tags for ledsætningsfunktion samt ordets sætningsinterne funktion,
fx @#FS-<ACC @SUB for "não acredito *que* seja verdade")

@#ICL- infinit ledsætning
(kombineres med tags for ledsætningsfunktion samt ordets sætningsinterne funktion,
fx @#ICL-SUBJ> @IMV i "*consertar* um relógio não é fácil")

@#ICL-AUX< argument- verbum i verbalkæde, refererer til forudgående hjælpeverbum
(verbalkædenotationen @FAUX- @#ICL-AUX< anvendes hvor begge verber har samme subjekt, notationen @FMV- @#ICL-<ACC bruges hvor subjekterne er forskellige)

@#AS- 'absolut' (i.e. verballøs) ledsætning ("small clause")
(kombineres med tags for ledsætningsfunktion samt ordets sætningsinterne funktion,
fx @#AS-<ADVL @ADVL> i "ajudou *onde* possível")

@AS< argument af complementiser i 'absolut' (dvs. verballøs) ledsætning

@#FS-S< sætningsanafor (sætningskomplement, der selv er en finit sætning)
(refererer tilbage til hele den forudgående sætning '...., *o que* era novo para mim')

VALENS-TAGS

("Sekundære" leksikon-tags, der benyttes af parseren under disambigueringen af de "primære" tags, i.e. ordklasse og syntaktisk funktion. På det syntaktiske plan disambigueres kun nogle få af valenstaggene, herunder relativ/interrogativ-skellet ved adverbier og pronominer. Kun de mest almindelige valenstags nævnes her.

<vt>	monotransitivt verbum akkusativ-objekt
<vi> (<ve>)	intransitivt verbum (ergativt verbum)
<vtd>	ditransitivt verbum akkusativ- og dativ-objekt
<PRP^vp>	monotransitivt verbum med præpositionelt objekt (med PRP som hoved)
<PRP^vtp>	ditransitivt verbum med akkusativ- og præpositionelt objekt
<vK>	copula-verbum med subjektprædikativ
<vtK>	copula-verbum med objektprædikativ
<va>	transitivt verbum med adverbielt argument: <va+LOC>, <va+DIR>, <vta+LOC>, <vta+DIR>
<vt+QUANT>	transitivt verbum med NP som kvantitativt adverbialobjekt (fx. "pesar")
<vt+TID>	transitivt verum med NP som temporalt adverbialobjekt (fx. "durar")
<vU>	"upersonligt" verbum (normalt i 3.person singularis, fx "chove")
<x>	hjælpeverbum med infinitiv (tagsekvens @(F)AUX- @#ICL-AUX<)
<x+PCP>	hjælpeverbum med participium (tagsekvens @(F)AUX- @#ICL-AUX<)
<x+GER>	hjælpeverbum med gerundium (tagsekvens @(F)AUX- @#ICL-AUX<)
<PRP^xp>	hjælpeverbum med (præpositions-) partikel og infinitiv (tagsekvens @(F)AUX- @PRT-AUX< - @#ICL-AUX<)
<xt>	"hjælpe-" verbum ved sanse-verber (ACI) og kausative konstruktioner, (tagsekvens @(F)MV- @SUBJ> - @#ICL-ACC)
<PRP^xtp>	"hjælpe-" verbum med akkusativ-objekt, der fungerer som subjekt i efterfølgende præpositionelt indledt infinitiv objektsætning (tagsekvens @(F)MV- @<ACC - @<PIV - @#ICL-P<)
<vr>	refleksive verbs (also <vrp>, <vaux-r>, <vaux-rp>)
<vq>	"kognitivt" verbum der styrer 'que'-sætning
<PRP^vpq>	"kognitivt" verbum der styrer præpositionssyntagme med 'que'-sætning
<qvK>	"upersonligt" verbum med 'que'-sætning som subjektprædikativ ("*parece* que")
<+interr>	"diskurs-" verbum eller nominal der styrer interrogativ-sætning
<+n>	substantiv der styrer egenavn (PROP) (fx "o *senhor* X")
<+num>	substantiv der styrer talord (fx "*cap.* 7", "no *dia* 5 de dezembro")
<num+>	"måleenheds-" substantiv (fx "20 metros")

<attr>	attributivt substantiv (fx "um presidente *comunista*")
<mass>	mængdesubstantiv (fx "leite", "água")
<+INF>	nominal der styrer infinitiv (N, ADJ)
<+PRP>	styrer præpositionssyntagme indledt af præpositionen PRP, fx <+sobre>
<PRP+>	(typisk) argument af præpositionen PRP
<+que> <+PRP+que>	nominal der styrer 'que'- sætning (N, ADJ)
<art>	bestemt artikel (DET)
<quant0/1/2/3>	kvantifikator (DET: <quant1>, <quant2>, <quant3>, SPEC: <quant0>)
<dem>	demonstrativt pronomen (DET: <dem>, SPEC: <dem0>)
<poss>	possessivt pronomen (DET)
<refl>	refleksiv ("se" PERS ACC, "si" PERS PIV)
<diff>	differentiator (DET) (fx "outro", "mesmo")
<rel>	relativt pronomen (DET, SPEC)
<interr>	interrogativt pronomen (DET, SPEC)
<post- det>	typisk placeret som post- determiner (DET @N<)
<post- attr>	typisk efterstillet typisk foranstillet adjektiv (ADJ @>N)
<adv>	adverbiel brug af adjektiv eller PP (ADJ @ADVL)
<KOMP> <igual>	"sidestillende" komparativ (ADJ, ADV) (fx "tanto", "tão")
<KOMP> <corr>	"korrelativt" komparativ (ADJ, ADV) (fx "mais velho", "melhor")
<komp> <igual>	"sidestillende" partikel der refererer til komparativ (fx "como", "quanto")
<komp> <corr>	"korrelativ" partikel der refererer til forudgående komparativ (e.g. "do=que")
<quant>	kvantificerende/intensitets- adverbium (fx "muito")
<setop>	"operationelt" adverbium (fx "não", "nunca", "ja'", "mais" in "não mais")
<dei>	deiktiske diskurspartikler (fx "aqui", "ontem")
<k>	konjunktionelle adverbier (fx "pois")
<card>	kardinalier (NUM)
<NUM- ord>	ordinaler (ADJ)
<NUM- fract>	brøk (N)
<cif>	ciffer (<card> NUM, <NUM- ord> ADJ)
<sam- >	første del af fusioneret flerords- udtryk ("de" i "dele")
<- sam>	sidste del af fusioneret flerords- udtryk ("ele" i "dele")
<*>	1. bogstav med stort
<hyfen>	ord med bindestreg
<ABBR>	forkortelse

SEMANTISKE TAGS

Der er ca. 200 forskellige ("sekundære") semantiske tags (især ved substantiver og adjektiver), som fx. <prof> for 'professional'. Disse tags er introduceret udfra et maskinoversættelses- perspektiv, og polyseme ord vil således have flere semantiske tags. Det semantiske niveau er stadigt eksperimentelt, men forsøg med enkelte tags tyder på at CG semantisk og valens- disambiguering er et effektivt redskab til polysemiresolution, ikke mindst i en maskinoversættelsessammenhæng. For substantivers vedkommende, er de semantiske tags afledt af 16 hierarkisk ordnede "atomare" træk. Verber tagges for ±HUM- subjektselektion, og adjektiver for ±HUM- nominalselektion.

Litteratur

Eckhard Bick, *Portugisisk - Dansk Ordbog,* Mnemo, Århus, 1993, 1995

Eckhard Bick, *The Parsing System "Palavras", Documentation,* upubliceret Ph.D. projektevaluering, 1995

Eckhard Bick, *Automatic Parsing of Portuguese,* i *Proceedings of the Second Workshop on Computational Processing of Written Portuguese,* Curitiba, 1996

Jean- Pierre Chanod & Pasi Tapanainen, "Tagging French - comparing a statistical and a constraint- based method", adapted from: *Statistical and Constraint- based Taggers for French,* Technical report MLTT- 016, Rank Xerox Research Centre, Grenoble, 1994

Timo Järvinen, "Annotating 200 million words: The Bank of English project", i *Proceedings of The 15th International Conference on Computational Linguistics Coling- 94,* Kyoto, Japan, 1994 (citeret fra: Pasi Tapanainen, *The Constraint Grammar Parser CG- 2,* Publications No. 27, Department of Linguistics, University of Helsinki, 1996)

Fred Karlsson, Atro Voutilainen, Juka Heikkilä, Arto Anttila (eds.), "Constraint Grammar, A Language- Independent System for Parsing Unrestricted Text, with an application to English", i: *Natural language text retrieval. Workshop notes from the Ninth National Conference on Artificial Intelligence,* Anaheim, CA, American Association for Artificial Intelligence, 1991

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

Fred Karlsson, "Robust parsing of unconstrained text", pp. 97- 121, i: Nellike Oostdijk & Pieter de Haan, *Corpus- based research into language,* Amsterdam, 1994

Kimmo Koskenniemi, *Two- Level Morphology: A General Computational Model for Word- Form Recognition and Production,* Publication No. 11, Department of Linguistics, University of Helsinki, 1983

Geoffrey Leech, Roger Garside, Michael Bryant, "The large- scale grammatical tagging of text", pp. 47- 64, in: Nellike Oostdijk & Pieter de Haan, *Corpus- based research into language,* Amsterdam, 1994

Atro Voutilainen, Juka Heikkilä, Arto Anttila, *Constraint Grammar of English, A Performance- Oriented Introduction,* Publication No. 21, Department of General Linguistics, University of Helsinki, 1992

Atro Voutilainen, *Designing a Parsing Grammar,* Publications No. 22, Department of Linguistics, University of Helsinki, 1994