AUTOMATISK ANALYSE
AF PORTUGISISK SKRIFTSPROG

Eckhard Bick

Institut for Lingvistik, Århus Universitet, Nordre Ringgade, DK-8000 Århus C
tel: +45 - 89 422152, fax: +45 - 86 281397, e-mail: lineb@hum.aau.dk

Abstract

The paper describes an automatic grammar- and lexicon-based parser for unrestricted Portuguese text. The parser has been developed as a three-year Ph.D.-project and is ultimately intended for applications like corpora tagging, grammar teaching and machine translation, which all have been made accessible in the form of internet based prototypes. Grammatical rules are formulated in the Constraint Grammar formalism (CG) and focus on robust disambiguation, treating several levels of linguistic analysis in a related manner. In spite of using a highly differentiated tag set, the parser yields correctness rates - for unrestricted and unknown text - of over 99% for morphology (part of speech and inflection) and 97-98% for syntactical function, even when geared to full disambiguation. Among other things, argument structure, dependency relations and subclause function are treated in an innovative way, that allows automatic transformation of the primary, "flat" CG-based

syntactic notation into traditional tree structures (like in DCG and PSG). The parser uses valency and semantical class information from the lexicon, and a pilot study on disambiguation on these levels has been conducted, yielding encouraging results.

The system runs at about 200 words/sec on a 200 MHz Pentium based Linux system, when using all levels. Morphological and POS disambiguation alone approach 2000 words/sec.

1. Oversigt

I denne artikel evalueres 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 applikativt perspektiv involverende bl. a. maskinoversættelse og grammatik-formidling (automatic tutoring), men i det følgende vil jeg koncentrere mig om at præsentere parserens notationelle system, især på det syntaktiske niveau, samt redegøre for hvordan man indenfor samme parsing-formalisme kan tackle en bilingual motiveret polysemi-resolution. Endeligt skal en kvantitativ evaluering samt en række eksempelsætninger 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[1] af ordformern *'como'* kaldes i CG-terminologien en *kohorte.* En typisk CG-regel[2] til disambiguering af denne flertydighed er fx. følgende:

[1] Forskellen mellem <rel> ADV og <interr> ADV er strengt set ikke morfologisk eller ordklassemotiveret, men udtryk for en semantisk-funktionel distinktion (den danske oversættelse ville i det første tilfælde som regel være 'som', men i det andet 'hvordan'. Som det beskrives sidst i artiklen, er det 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

(2) SELECT (VFIN) IF (NOT *-1 VFIN) (NOT *1 VFIN)
[vælg for enhver ordform læsningen VFIN (finit verbum) hvis der ikke (NOT) - hverken til venstre (*-1) eller til højre (*1) - findes et andet ord der kan være VFIN.][3]

---

som *sekundære* tags (som udelukkende bruges til at disambiguere morfologiske/syntaktiske tags), men også som selvstændige *primære* tags, der kan og skal disambigueres, som i ordformen *'revista'*, hvor den enkelt ordklasseambiguitet (V-N) bliver til firedobbelt leksemambiguitet.

rever <vt> V 'gense'	realiseret valens: transitiv <vt>
rever <vi> V 'sive igennem'	realiseret valens: intransitiv <vi>
revista <+n><rr> N 'avis'	realiseret valens: titel <+n>, semantisk klasse: læsestof <rr>
revista <CP> N 'inspektion'	realiseret semantisk klasse: +CONTROL, +PERFEKTIV

[2] Jeg anvender her konventionen fra Pasi Tapanainens cg2-compiler, der bl.a. erstatter de ældre operatorer '@w=0' og '@w=!' med de almindelige engelske ord 'REMOVE' med ' SELECT'.

[3] 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).

Ved først at tilføje ("mappe") alle[4] 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=adverbial, @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 'SELECT'-regel der udnytter verbets transitivitet, men den kan lige så godt fremstå indirekte[5], - ved at være den sidste overlevende læsning, efter at CG-regler har forkastet de andre:

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

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

REMOVE (@OC) IF (NOT *-1 @ACC) (NOT *1 @ACC)
[forkast objektprædikativlæsningen (@OC) hvis der ikke findes et direkte objekt (@ACC) i sætningen][6]

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

[4] 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).

[5] 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".

[6] 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-specifikke.

3.1 Syntaktisk form og syntaktisk funktion

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, og også i min parser 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[7]). Dependensmarkører bliver enten hæftet til de funktionelle tags (fx @<SUBJ, @ADVL>, @N<PRED), eller står, ved visse bestemmerled, alene (fx @>N for [bestemmer-] prænominal).

(5)	Temos	[ter] <vt> V PR 1P IND VFIN	@FMV
	em	[em] <sam-> PRP	@<ADVL
	este	[este] <-sam> <dem> DET M S	@>N
	país	[país] <top> N M S	@P<
	uns	[um] <art> DET P S	@>N
	castelos	[castelho] <hus> N M P	@<ACC
	muito	[muito] <quant> ADV	@>A
	velhos	[velho] ADJ M P	@N<

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[8].

I eksemplet befinder 'muito' sig langt nede i uroen, men kender sin 'adverbial-adjekt'- (@>A) snor til 'velho'. Denne igen fastgøres til venstre som postnominal (@N<) til 'castelo'. 'Castelo' selv ved, at det er direkte objekt (@<ACC) til et venstre-(<)stående hovedverbum, 'temos', som er roden i dependens-uroen.

Men uden mere komplekse 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, som det fx er tilfældet når en infinitivsætning fungerer som subjekt i hovedsætningen*('Visiting the Louvre was not his only reason for coming to Paris')*, eller når en infinitivsætnings eget subjekt efter blokeringsreglen konkurrerer med matrixsætningens subjekt *('O perigo de os inimigos atacarem o*

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

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

castelo era imanente'). Også hierarkisering af ledsætningsgrænser (fx ved indskudte ledsætninger) kan være et problem.

Min løsning 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[9]. 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.

(6)

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>			
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]

Lad os se på et mere komplekst eksempel: *O baque foi atenuado pelo fato de sua mulher ter um emprego que garante as despesas básicas da família.* Nedenstående analyse gør det tydeligt hvordan dependensrelationerne samler sætningens byggeklodser i en hierarkisk struktur. Kasserne markerer (udefra indad) hovedsætningen, et passivkomplement, en infinitiv ledsætning (der fungerer som præpositionskomplement) og en finit ledsætning (der fungerer som et postnominalt attribut). Nominalsyntagmer er skygget, og den syntaktiske makrostruktur er tilføjet til venstre.

(7)

SUBJ	o	[o] <art> DET M S **@>N** 'den'
	baque	[baque] <cP> N M S **@SUBJ>** 'fald'
VP	foi	[ser] <x+PCP> V PS 3S IND VFIN **@FAUX** 'blive'
	atenuado	[atenuar] <vt> <sN> V PCP M S **@IMV @#ICL-AUX<** 'svække'
PP-PASS	por	[por] <sam-> <+INF> <PCP+> PRP **@<PASS** 'af'

[9] 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. Tapanainen (1997) har udviklet en egentlig dependensgrammatik som overbygning for en CG-baseret morfologisk disambiguering. Her arbejdes der med nummertilordning af head og dependenter.

P<	o	[o] <-sam> <art> DET M S **@>N** 'den'
	fato	[fato] <ac> <+de+INF> N M S **@P<** 'kendsgerning'
PP-N<	de	[de] PRP **@N<** 'af'
SUBJ	sua	[seu] <poss 3S/P> DET F S **@>N** 'hans'
	mulher	[mulher] <H> N F S **@SUBJ>** 'kvinde'
VP & **ICL-P<**	ter	[ter] <vt> <sH> V INF 0/1/3S **@IMV @#ICL-P<** have'
ACC	um	[um] <quant2> <arti> DET M S **@>N** 'en'
	emprego	[emprego] <stil> <ac> N M S **@<ACC** 'stilling'
SUBJ & **FS-N<**	que	[que] <rel> SPEC M/F S/P **@SUBJ>** **@#FS-N<** 'som'
	garante	[garantir] <vt> <v-cog> V PR 3S IND VFIN **@FMV** garantere'
ACC	as	[a] <art> DET F P **@>N** 'den'
	despesas	[despesa] <ac> N F P **@<ACC** 'udgift'
	básicas	[básico] <jn> ADJ F P **@N<** 'basal'
PP-N<	de	[de] <sam-> PRP **@N<** '(genitiv)'
P<	a	[a] <-sam> <art> DET F S **@>N** 'den'
	família	[família] <HH> N F S **@P<** 'familie'

Nedenstående ordkæde viser hvordan en dependensgrammatisk "attachment sequence" ser ud hvis man fører den op fra laveste niveau (her fra artiklen 'a') til højeste niveau, verbalkernen i hovedsætningen ('>' betyder "hæfter til", ':' betyder "danner") :

a > família:NP > de:PP > despesas:NP > garante:FS > emprego:NP > ter:ICL > de:PP > fato:NP > por:PP > atenuado:ICL > foi:S a > família:NP > de:PP > despesas:NP > garante:FS > emprego:NP > ter:ICL > de:PP > fato:NP > por:PP > atenuado:ICL > foi:S

3.2. Transformation af flad dependenssyntaks til træstrukturer

I lyset af den store popularitet som konstituentgrammatikkerne nyder i nutidens lingvistik, er det således nærliggende at spørge: kan en flad (CG) syntaktisk beskrivelse på denne måde bevare så megen strukturel information, at der kan opretholdes en vis ækvivalens og "transformerbarhed" i forhold til klassiske træ-notationer?

For at vise at dette godt kan lade sig gøre, har jeg skrevet et computerprogram, der identificerer syntagme- og ledsætningsgrænserne i en flad CG-beskrivelse, markerer dem som *form* (np, pp, icl m.m.) og tildeler dem som*funktion* deres kernes syntaktiske CG-tag. Transformationen vises ved følgende sætning: *O crise apura o palador do consumidor e valoriza o dono de restaurante que pilota a própria cozinha.*

(8a) anlyseret tekst, i "flad", ordbaseret CG-notation:

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

(8b) teksten transformeret til træstruktur, med indføjede syntagme-tags og hierarkisk indrykning:

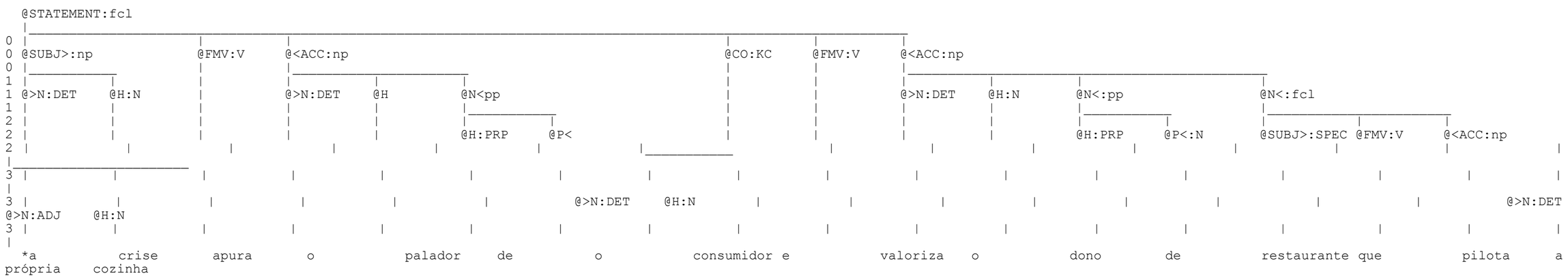
@SUBJ>:np		
|-@>N:DET F S	*a	[a] <art>
|-@H:N F S	crise	[crise] <sit>
@FMV:V PR 3S IND VFIN	apura	[apurar] <vt> <sN>
@<ACC:np		
|-@>N:DET M S	o	[o] <art>
|-@H:N M S	paladar	[paladar] <anost> <fh>
|-@N<:pp		
|-@H:PRP	de	[de] <sam->
|-@P<:np		
|-@>N:DET M S	o	[o] <-sam> <art>
|-@H:N M S	consumidor	[consumir] <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	restaurante	[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>

[**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]

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

(8c) Samme sætning, automatisk transformeret til vandret trænotation



En vigtig forskel mellem den flade CG-notation og træ-notationen er, at denne *skal* opløse visse flertydigheder, som den flade syntaks underspecificerer, fx. i forbindelse med tilhæftningen af postnominaler (især præpositionssyntagmer), koordination og frie nominaladjunkter. Denne underspecifikation bliver imidlertid til et gode, når man betragter den udfra et MT-perspektiv: - for det første er mange af tilfældene 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 (10a) og "kort" (9b) vs. "lang" (9a) tilhæftning af postnominale præpositionssyntagmer) forholdsvis universelle, dvs. sproguafhængig, således at de kan bevares i oversættelsen, der baseres direkte på den "flade" beskrivelse (9c).

(9a) Han hentede **((**manden @<ACC med @N< cyklen @P<**)** fra @N< Kina @P<**)**.
(9b) Han hentede **(**manden @<ACC med @N< **(**cyklen @P< fra @N< Kina @P<**))**.
(9c) 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:

(10a) gifte @>N kvinder @NPHR og @CO mænd @NPHR
(10b) homens @NPHR e @CO **mulheres** @NPHR casad**as** @N<

4. 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[10], 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[11].

For at opnå maksimal præcision, har jeg også arbejdet med et større utagget tekstmateriale (170.000 ord fra Borba-Ramsey-corpuset[12]), 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)**

[10] 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.

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

[12] 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).

verbalfunktion	3			
verbers argumenter	25			
præpositioners argumenter	2			
Argumentstruktur		**30**		
Bestemmere		**13**		
Adjunkter		**11**		
Ledsætninger		**10**		
Alle syntaktiske			**64**	**(97.4 % correctness)**
"lokale" syntaktiske fejl pga. morfologiske/ordklasse-fejl	-27			
Rent syntaktiske			**37**	**(98.5% correctness)**

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[13] (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 %**

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

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 avistekster 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 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. 5.1). 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.

5. Parseren

5.1 Tag-sættet[14]

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 disambigueringsreglerne. 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)[15].

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).

5.2 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

[14] En fuldstændig oversigt over de brugte morfologiske og syntaktiske tags og deres definitioner findes i [Bick, 1997], eller kan hentes via internet på *http://ling.hum.aau.dk/~eckhard/Linguistics.html.*

[15] 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), ordklassernes 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.

programmeringssprogene C og Perl. Parseren omfatter følgende moduler på det morfologisk-syntaktiske niveau[16]:

- 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. 200 ord/sec på en 200 MHz Pentium-baseret Linux-maskine. Den morfologiske/ordklasse-disambiguering alene opnår hastigheder i nærheden af 2000 ord/sec. Systemet kan afprøves igennem en interaktiv brugerflade på følgende web-adresse: *http://ling.hum.aau.dk/~eckhard/Linguistics.html.*).

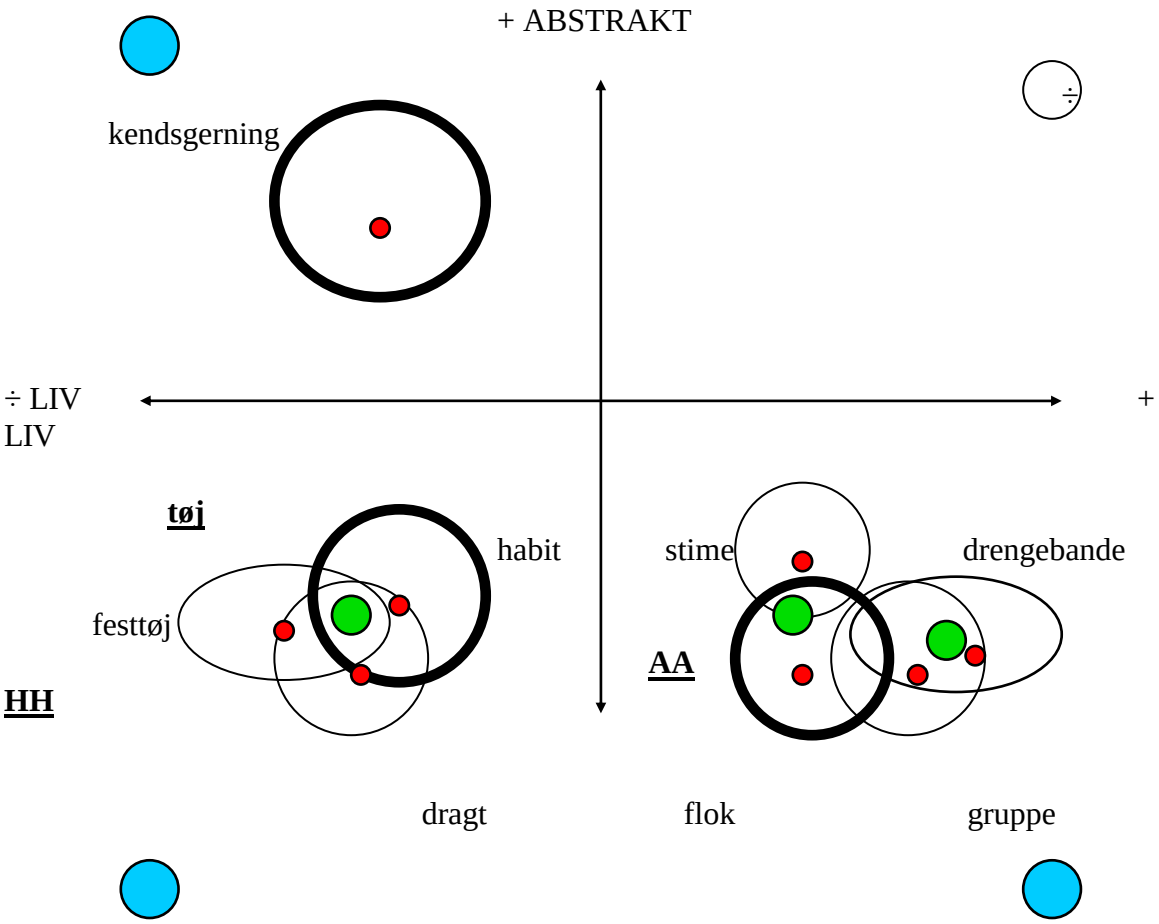
6. Det semantiske perspektiv: "Incremental Semantical Parsing" (ISP)

Det er almindeligt at niveaudele sproglig analyse (såvel manuel som automatisk), fx i et morfologisk, syntaktisk, semantisk og pragmatisk niveau, hvor forskellige applikationer kræver en analyse på forskellige niveauer. Således er en morfologisk analyse tilstrækkelig for en forsker der arbejder med corpusbaserede frekvensanalyser, mens den internetbaserede grammatikformidling i projektet VISL kræver en syntaktisk analyse og maskinoversættelse en semantisk.

Det ser imidlertid ud som om det samme redskab - Constraint Grammar - kan "presses" til stadigt højere analyseniveauer (Bick, 1996, 1997) - forudsat, der samtidigt udvikles en tilsvarende leksikografisk database. Man kan sige at analysens finkornethed her som andetsteds ikke er teknikken iboende, men snarere formålsdreven, og kan forbedres "inkrementelt". Således er det måske principielt umuligt databasemæssigt at *definere* det brasiliansk portugisiske ord *fato,* men i et bilingualt (dvs. praktisk orienteret MT-) perspektiv kan man udmærket adskille de tre *danske* oversættelser "kendsgerning", "habit" og "flok" ved hjælp af *atomic semantic features* som henholdsvis *abstrakt ikke levende* ("kendsgerning"), *ikke abstrakt ikke levende* ("habit") og *ikke abstrakt levende* ("flok"). Disse træk er ovenikøbet tilstrækkelige til at afgrænse (ikke definere!) større prototypfamilier mod hinanden, som *"tøj"* og *"dyrisk flerhed"* eller *"menneskeflerhed"* (i skemaet henholdsvis AA og HH). I en Constraint Grammar parser kan et hierarki af

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

leksikon- og kontekst-drevne grammatiske regler "forbyde" eller "selektere" disse træk eller prototypiske trækfamilier[17] i den konkrete sætning.



Diagrammet placerer en række ord i et semantisk felt, i forhold til hianden og i forhold til prototypiske begreber (halvstore grønne cirkler) eller trækkombinationer (store blå cirkler). Ordenes kernebetydninger er symboliseret ved små røde punkter, og deres semantiske muligheder med cirkler af mere eller mindre vilkårlig størrelse. Det fremgår at 'festtøj', 'dragt' og 'habit' er vanskelige at adskille, siden de alle tilhører prototypen 'tøj'. Derimod er et enkelt atomisk træk - ±LIV - nok til at distancere alle tre fra ord som 'flok' eller 'drengebande'. Vil man skelne mellem ord indenfor samme LIV/ABSTRAKT-kvadrant, skal der yderligere træk til, fx. ±DYR til at afgrænse AA-ordet 'stime' fra HH-ordet 'drengebande' ('flok' og 'drengebande' har et semantisk overlap, der kommer til udtryk i 'en flok drenge' og bedst kan beskrives som metaforisk: 'flok'-semet projicerer sit træk +DYR på det valensbundne komplement 'drenge'). Trækkombinationen +ABSTRAKT/+LIV udgår iøvrigt, idet ±LIV er en hierarkisk binær underopdelng af ÷ABSTRAKT.

En særlig elegant og "inkrementel" løsning for polysemireduktion af indholdsmæssigt flertydige ord er den semantiske udnyttelse af "lavere parsing-

[17] I alt anvendes ca. 200 forskellige tags for semantiske prototyper. For substantivers vedkommende, er de semantiske tags afledt af 16 hierarkisk ordnede "atomare" træk. Verber tagges for ±HUM-subjektselektion, og adjektiver for ±HUM-nominalselektion.

information" (morfologisk form eller syntaktisk funktion), som systemet allerede *er* i stand til at slå fast. Ordet "saber" fx. betyder 'vide' når det er bøjet i imperfektum, men 'få at vide' i perfektum. Her kan morfologisk information kapitaliseres til semantiske formål. Også ordklassen kan bruges: er "saber" brugt som hjælpeverbum (AUX), betyder det 'kunne'. Endelig kan man udnytte syntaktisk information fra sætningens andre led til at instantiere et af flere mulige valensmønstre for "saber": mens både 'vide' og 'få at vide' kræver direkte objekter, skal betydningen 'smage' vælges før adverbiale komplementer (godt/dårligt), og 'smage af' før et præpositionalobjekt indledt af præpositionen 'a'.

Leksikografisk kan denne fremgangsmåde implementeres ved hjælp af såkaldte (polysemi-) diskriminatorer:

(11) **saber V**

@MV, IMPF, <vq><vt>	'vide'
@MV, PERF, <vq><vt>	'få at vide'
@AUX, <+INF>	'kunne'
@MV, <va>	'smage'
@MV, <a^vp>	'smage af'
@MV, <de^vp>	'kende til'

[@ = syntaktisk funktion: MV =hovedverbum, AUX=hjælpeverbum; <> =valens: <vt> =transitiv, <+INF> efterfulgt af infinitiv, <va> =med adverbialobjekt, <vp> =med præpositionalobjekt, a^ =præposition "a", de^ =præposition "de"; morfologi: IMPF =imperfektum, PERF =perfektum]

Endeligt kan de semantisk éntydige (eller allerede disambiguerede) ord hjælpe ved analysen af de flertydige. Således skal den portugisiske præposition "de" oversættes med 'fra', når præpositionens argument er et sted (+LOC), men 'af', hvis der følger et materialeord (fx. *de ouro* af guld) og med genitiv, hvis komplementet er et menneske (+HUM: *o cachorro do homem* - mandens hund). Igen skal tilsvarende diskriminatorer optages i leksikonnet, i form af semantisk beriget valensinformation (såkaldte selektionsrestriktioner).

(12) Følgende sætning illustrerer mulighederner:

apesar=**de**	[apesar=de] <sam-> PRP @ADVL> 'på trods **af**'	*
a	[a] <-sam> <art> DET F S @>N 'den'	
advertência	[advertência] <s> N F S @P< 'råd'	
de	[de] <sam-> <+hum> PRP @N< '**(genitiv)**'	*
o	[o] <-sam> <art> DET M S @>N 'den'	
meu	[meu] <poss 1S> DET M S @>N 'min'	
pai	[pai] <fam> N M S @P< 'far'	
,		
que	[que] <rel> SPEC M/F S/P @#FS-N< 'som'	
não	[não] ADV @ADVL> 'ikke'	
gosta	[gostar] <de^vp> <vH> <ink> V PR 3S IND VFIN @FMV 'kunne lide'	
de	[de] <sam-> PRP @<PIV '**af**'	*
a	[a] <-sam> <art> DET F S @>N 'den'	
minha	[meu] <poss 1S> DET F S @>N 'min'	

```
nova            [novo] <ante-attr> <jn> ADJ F S @>N 'ny'
vida            [vida] <feat> <per> N F S @P< 'liv'
,
comprei         [comprar] <vt> <vH> <ink> V PS 1S IND VFIN @FMV 'købe'
uma             [um] <quant2> <arti> DET F S @>N 'en'
carroçada       [carroçada] <qus> N F S @<ACC 'læs'
de              [de] <quant+> PRP @N< '(partitiv)'                    *
coisas          [coisa] <cc> <ac> N F P @P< 'ting-1'
$,
por=exemplo     [por=exemplo] <adv> <+NP> PP @<ADVL 'fx'
um              [um] <quant2> <arti> DET M S @>N 'en'
fato            [fato] <tøj> <AA> N M S @ACC< 'habit'
de              [de] <+mat> PRP @N< 'af'                              *
lã              [lã] <cm> <stof> N F S @P< 'uld'
preta           [preto] <col> <jn> ADJ F S @N< 'sort'
que             [que] <rel> SPEC M/F S/P @SUBJ> @#FS-N< 'som'
veio            [vir] <va+DIR> <sN> V PS 3S IND VFIN @FMV 'komme'
de              [de] <sam-> <+top> PRP @<ADV 'fra'                    *
a               [a] <-sam> <art> DET F S @>N 'den'
*argentina      [Argentina] <top> PROP F S @P< 'Argentina'
de              [de] <+V> PRP @N< 'med'                               *
avião           [avião] <fly> N M S @P< 'fly'
em=menos=de     [em=menos=de] <c> PRP @<ADVL 'på mindre end'          *
uma             [um] <card> NUM F S @>N 'een'
semana          [semana] <dur> <num+> N F S @P< 'uge'
.
```

Den tilsvarende leksikonartikel oplister først en række valensmæssige og semantiske kontekstualiseringsmuligheder for præpositionen 'de', og angiver så hvilken oversættelse der skal vælges hvis den ene eller anden polysemi-diskrimintor instantieres (dvs. overlever disambiguerings-constraints'ene). Også information om syntaktisk funktion - fra det "næstlavere" parsingniveau - (her @KOMP< for komparativkomplement) kan bruges som diskriminator:

de PRP <komp><corr><+hum><+mat><+top><+V><+feat><+il><+tøj><quant+>

__	af	*(default-oversættelse)*
__ <quant+>	(partitiv)	*(efter quantitiva)*
__ <+mat>	af	*(før materiale-ord)*
__ <+hum>	(genitiv)	*(før egennavne og ord for mennesker)*
__ <+V><+feat><+il><+tøj>	med	*(før køretøjer, træk, værktøjer eller tøj)*
__ <+top>	fra	*(før toponymer og andre stedbetegnelser)*
__ <komp> @KOMP<	af, blandt	*(som komparativkomplement: "den største af ..."*
__ <komp><corr> @KOMP<	end	*(som korrelativ komparativkomplement: "større end"*

For substantivet *'fato'* foreligger følgende polysemidiskriminatorer i leksikonnet, hvoraf nogle er valensinstantieringer (<+que>, <+de+que>, <+de+INF>), nogle semantiske prototyper (<ac><tøj><AA>) og én en oplisting af alle de atomare semantiske træk, prototyperne tilsammen dækker over (fx. A = +ANIM, a = ÷ANIM).

fato N M <ac><tøj><AA><+que><+de+que><+de+INF><=EecIiJjAahmNnvpsdxflt=>	
__ <ac><+de+que><+de+INF>	kendsgerning
__ <tøj>	habit, kostyme
__ <AA>	flok {fx geder}
fato=de=banho N M	badedragt
fato=de=macaco N M	kedeldragt

I sætningen *'Um fato de ovelhas corria no campo'* skal parseren bruge 8 regler for at disambiguere polysemien i *'fato'*, - ikke medregnet de regler for sætningens *øvrige* ord, der skulle til for at skabe de nødvendige éntydige kontekstbetingelser.

(12a)

*um	[um] <quant2> <arti> DET M S @>N 'en' 51227
fato	[fato] **<AA>** N M S @SUBJ> '**flok**' UTR 22490
de	[de] <quant+> PRP @N< '(partitiv)' 14502
ovelhas	[ovelha] **<zo>** N F **P** @P< 'får' NEU 36477
corria	[correr] <vi> V IMPF 1/3S IND VFIN @FMV '**løbe**' PCP-ER 13442
em	[em] <sam-> <+top> PRP @<ADVL 'i' 18125
o	[o] <-sam> <art> DET M S @>N 'den' 35367
campo	[campo] <BB> <top> <topabs> N M S @P< 'mark-2' UTR 8784

En prøvekørsel af parseren med regel-tracing viser at der først anvendes 3 valensinstantieringsregler:

*REMOVE (<+de+que>) (*1 CLB/SB LINK NOT 0 QUE-KS) ;*
... hvis den næstfølgende (led)sætningsgrænse ikke er konjunktionen 'que'.
*REMOVE (<+que>) (*1 NON-ADV LINK NOT 0 QUE-KS) ;*
... hvis det førstfølgende ikke-adverbielle ord ikke er konjunktionen 'que'.
*REMOVE (<+de+INF>) (*1 CLB/SB OR <+PRP+INF> BARRIER @#ICL-P<) ;*
... hvis der ikke forekommer en præpositionskomplementerende infinitiv før den næstfølgende sætningsgrænse eller infinitivvalente præposition.

Herefter fjernes positive (store bogstaver) eller negative (små bogstaver) semantic features. Den eneste virkelige regel er den første, der slår fast at *'fato'* i denne sætning kan bevæge sig; de andre er bare "reflekskonklusioner" ud fra trækket +MOVE.

*REMOVE (<i>) (0 @SUBJ> AND <I>) (*1 @MV BARRIER CLB-ORD LINK 0 V-MOVE);*
hvis det er subjekt og der følger et bevægelses-hovedverbum uden sætningsgrænse imellem, så kan det bevæge sig.
REMOVE (<j>) (NOT 0 <i>) ;

hvis det kan bevæge sig (aktiv bevægelighed, + =I, ÷ =i), kan man også bevæge det (passiv bevægelighed, + =J, ÷ =j).

REMOVE (<tøj>) (NOT 0 <i>) ;
det kan ikke være prototypen tøj hvis den kan bevæge sig.

REMOVE (<e>) (NOT 0 <i>) ;
det kan ikke være abstrakt hvis den kan bevæge sig.

REMOVE (<ac>) (NOT 0 <e>) ;
det kan ikke være prototypen 'abstrakt ting', hvis det ikke er abstrakt.

I udtrykket *'Um fato de lã preta'* bruges 4 af de samme regler plus en regel der fastslår den postnominale materialekontekst (af sort uld) til højre.

(12b)

*um	[um] <quant2> <arti> DET M S @>N 'en'
fato	[fato] **<tøj>** <AA> N M S @NPHR '**habit**'
de	[de] <+mat> PRP @N< 'af'
lã	[lã] <cm> **<stof>** N F **S** @P< '**uld**'
preta	[preto] <col> <jn> ADJ F S @N< 'sort'

*REMOVE (<+de+que>) (*1 CLB/SB LINK NOT 0 QUE-KS) ;*
*REMOVE (<+que>) (*1 NON-ADV LINK NOT 0 QUE-KS) ;*
*REMOVE (<+de+INF>) (*1 CLB/SB OR <+PRP+INF> BARRIER @#ICL-P<) ;*
*REMOVE (<e>) (*1 PRP-DE BARRIER NON-POST-N LINK 0 @N< LINK 1 @P< LINK 0 <M> AND <E>);* det kan ikke være abstrakt, hvis der uden andet end postnominaler imellem følger præpositionen 'de' brugt som postnominal og med et direkte efterfølgende (dvs. artikelløs) argument af typen +MASS og +KONKRET (dvs. fx. stof, materiale).
REMOVE (<ac>) (NOT <=e>) ;

Selv hvor ingen af de semantiske regler griber, kan det stadigt være en valensinstantiering, der afgør polysemidifferentieringen[18]. Her er det <+de+que> der overlever constraints'ene.

(12c)

*o	[o] <art> DET M S @>N 'den'
fato	[fato] **<ac>** <tøj> <AA> **<+de+que>** N M S @NPHR '**kendsgerning**'
de	[de] PRP @N< '(af)'
que	[que] KS @SUB @#FS-P< '**at**'
sua	[seu] <poss 3S/P> DET F S @>N 'hans'
namorada	[namorada] <title> N F S @SUBJ> 'kæreste'
tem	[ter] <vt> <sH> V PR 3S IND @FMV 'have'
um	[um] <quant2> <arti> DET M S @>N 'en'
emprego	[emprego] <stil> <ac> N M S @<ACC 'stilling'

*REMOVE (<+qu>) (*1 NON-POST-N LINK NOT 0 QUE-KS) ;*

[18] Parseren vælger det oversættelsesalternativ, der har flest overlevende diskriminatorer. Er dette kriterium utilstrækkelig, vælges på heuristisk vis den første oversættelse i listen.

hvis det første ord efter eventuelle postnominaler ikke er konjunktionen 'que'.
*REMOVE (<+de+INF>) (*1 CLB/SB OR <+PRP+INF> BARRIER @#ICL-P<) ;*

7. Konklusion

Constraint-Grammar-baserede parsere er robuste og kan opnå meget lave fejlprocenter på fri, løbende tekst. Metoden lægger op til en deskriptivt elegant, ordbaseret notation, der inden for samme formalisme kan håndtere flere grammatiske analyseniveauer. På det morfologiske plan synes portugisisk, et stærkt inflekterende sprog med relativ fri ordstilling, at udvise den samme grad af ambiguitets- og regelkompleksitet som engelsk, et inflektionsfattigt sprog med fast ordstilling, et faktum der underbygger Constraint-Grammer-skolens påstand om formalismens universalitet og sproguafhængighed[19]. På det syntaktiske plan er det lykkedes for portugisisk at behandle også ledsætningers mere komplekse form og funktion, samt at muliggøre automatisk transformation fra en detaljeret flad dependensnotation til konstituentgrammatiske træstrukturer. Endeligt viser forsøg med det semantiske plan[20] at formalismen også er egnet til en bilingual motiveret polysemiresolution, på den ene side ved at udnytte morfologisk-syntaktisk information (herunder instantieret valens) fra "lavere" analyseniveauer, på den anden side ved at disambiguere semantisk ambiguitet ved hjælp af tags for semantiske prototyper og atomare semantiske træk.

Resultaterne peger på at parseren vil kunne integreres i applikative kontekster som fx. maskinoversættelse, grammatiske tutoring systemer[21] og grammatiske filtre til corpussøgning.

[19] Sproguafhængighed gælder formalismen og compiler-implementeringen, *ikke* de enkelte regler, der ikke kan overføres fra et sprog til et andet.
[20] Det semantiske plan er ellers ikke omfattet af Ph.D.-projektet, ligesom en del af det syntaktiske arbejde ligger udenfor projektrammen.
[21] Parseren er blevet forsynet med en tilsvarende (prototypisk) brugerflade i forbindelse med VISL-projektet ved Institut for Sprog og Kommunikation, OU *(Visual Interactive Syntax Learning).*

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

Eckhard Bick, *Dependensstrukturer i Constraint Grammar Syntaks for Portugisisk,* i *Datalingvistisk forenings årsmøde nr.6,* Aalborg 1997

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)

Timo Järvinen & Pasi Tapanainen, *A Dependency Parser for English,* Helsinki, 1997

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