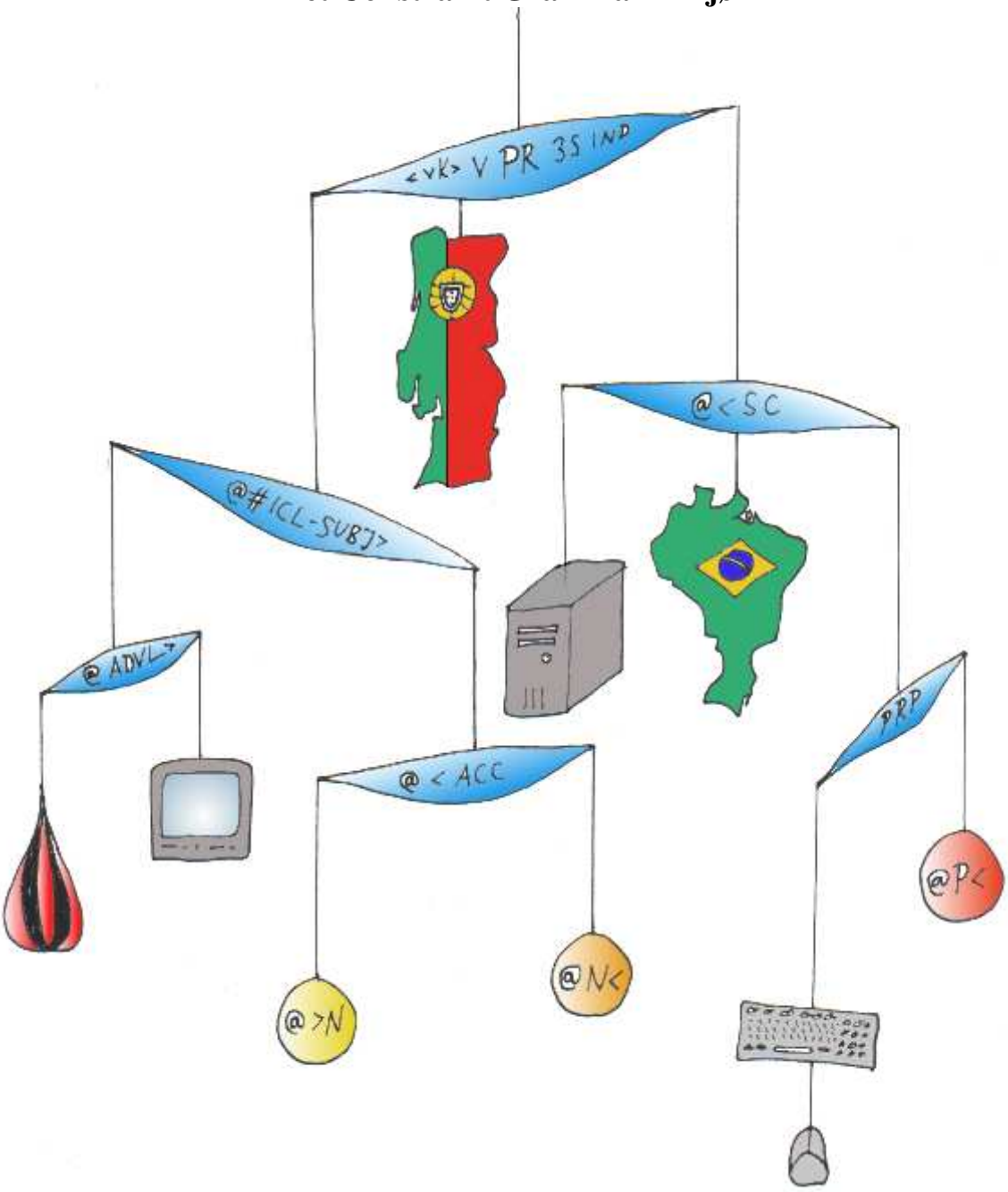
Eckhard Bick

♠

Parsing-systemet "Palavras"

Automatisk Grammatisk Analyse af Portugisisk
i et Constraint Grammar-Miljø



Institut for Lingvistik, Aarhus Universitet
e-mail: lineb@hum.au.dk, web: http://visl.hum.sdu.dk

1. Introduktion

1.1. Projektet

Jeg vil idag forsvare en afhandling der beskriver udviklingen og opbygningen af en morfologisk-syntaktisk Constraint Grammar-parser for fri portugisisk tekst. Foruden systemets struktur og performans evalueres en række lingvistiske og metodologiske implikationer, herunder vekselvirkningerne mellem parsing-teknik, korpusdata og grammatisk system. Projektet har en leksikografisk baggrund (beskrevet i Bick 1993) og et applikativt perspektiv, der involverer bl.a. korpus-annotation, maskinoversættelse og grammatik-formidling (Bick 1997-3).

Forskningsforløbet har hele tiden rummet to gensidigt kompletterende aspekter, det ene teoretisk, det andet praktisk. Det fysiske resultat, parseren selv samt en række applikationer, er tilgængelige på http://visl.hum.sdu.dk.

Det tekstuelle resultat, afhandlingen, sammenfattes i det følgende, hvor der dog kun redegøres for afhandlingens mere generelle afsnit, først og fremmest for at kontekstualisere forskningsresultaterne. Hvad angår en morfologisk-syntaktisk beskrivelse af portugisisk (fx. produktiv derivation, ledsætningstyper, komparation, verbalkæder, adverbialer), samt diskussionen af de mere tekniske parsing-problemer (fx. håndtering af ortografisk variation, navne, forkortelser, leksiko-morfologisk heuristik[1]), henvises der til afhandlingens engelske hovedtekst. En pædagogisk fremstilling af den implementerede portugisiske syntaks findes desuden i (Bick 1999).

1.2. Constraint Grammar

De fleste ord i natursprogstekster er - isoleret set - flertydige med hensyn til ordklasse, bøjning, syntaktisk funktion, semantisk indhold m.m. I en automatisk analyse er det først og fremmest sætningskonteksten der skal afgøre hvordan et ord skal forstås, - snarere end den indholdsmæssige tekstuelle sammenhæng eller menneskelig "viden om verden", der begge er langt vanskeligere at repræsentere i et computerprogram.

Constraint Grammar (CG), som den er udviklet af prof. Fred Karlsson og Helsinki-skolen (Karlsson, 1990 og Karlsson et.al., 1995), er en grammatisk metode der beskriver sproglig struktur ved først at formulere og så at opløse grammatisk flertydighed. Disambigueringen er kontekstuel, og foregår ved at opstille regler for hvilken af et ords mulige læsninger der skal vælges, forkastes, tilføjes eller ændres[2].

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 morfologisk analysator. Ideelt set leveres som output for hver ordform kun én tag-linie[3], med den korrekte grundform, ordklasse, syntaktisk funktion m.m.

---

[1] For en diskussion af systemets leksiko-morfologiske heuristik, jf. desuden (Bick, 1998).

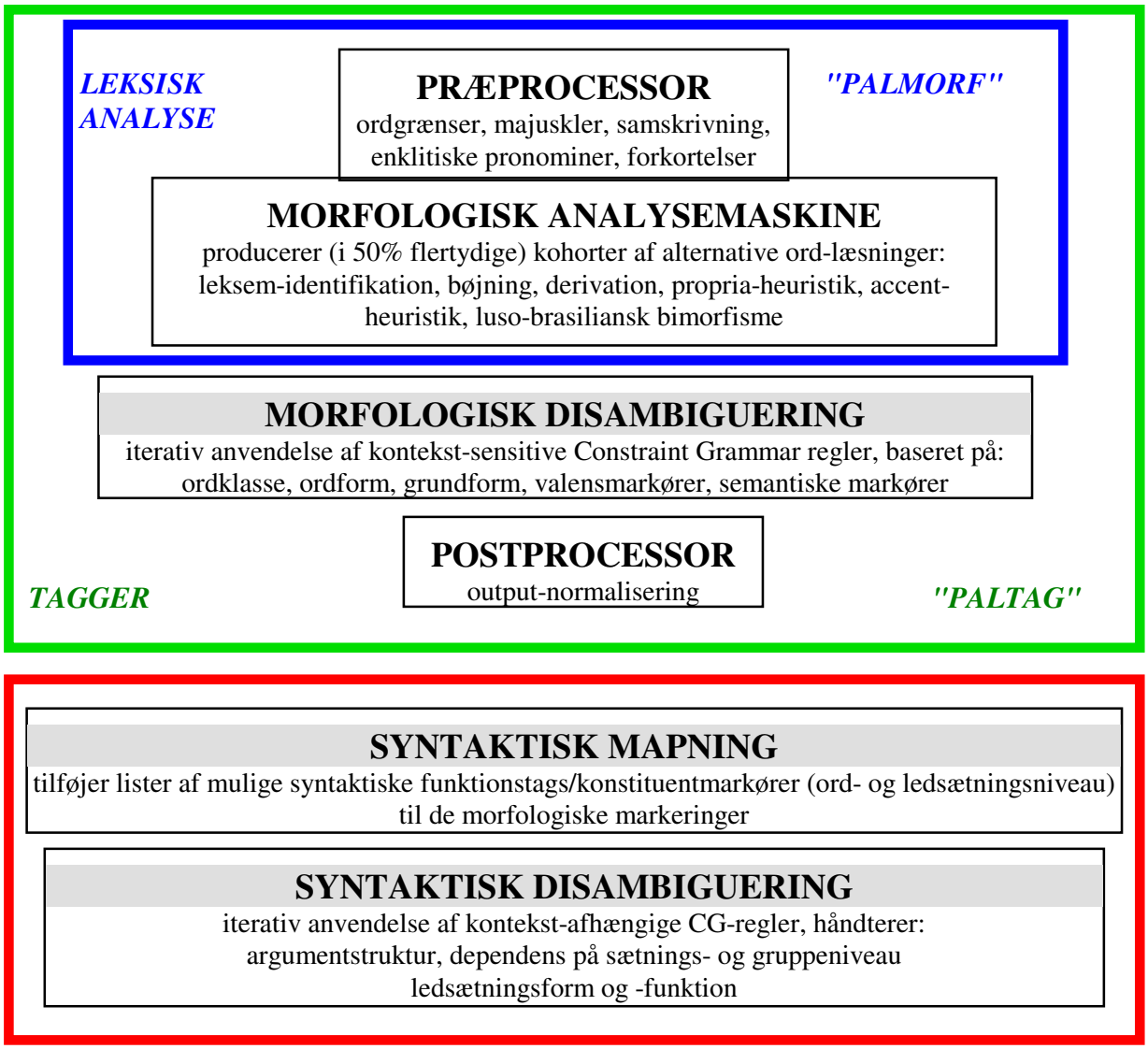
[2] Disse 4 operationer (vælge, fjerne, tilføje, ændre) kan betragtes som strengoperationer i stil med substitutionskommandoer for regular expressions i unix'verdenens sed, gawk og perl. En fleksibel ændring af tags er dog med de nuværende kompilere kun mulig ved netop at gøre supplerende brug af fx. perl-baserede filterprogrammer før og efter de egentlige CG-operationer. Selvom jeg i en periode har eksperimenteret med min egen kompiler, er systemet p.t. optimeret til Pasi Tapanainens noget hurtigere og mere effektive cg2-compiler (Tapanainen, 1996).

[3] Med undtagelsen af ægte flertydigheder, samt flertydigheder der vanskeligt kan opløses vha. information fra sætningsvinduet alene. Alt efter hvilken heuristicitetsgrad der tillades i regelsættet, vil der i disse tilfælde kunne være flere overlevende læsninger.

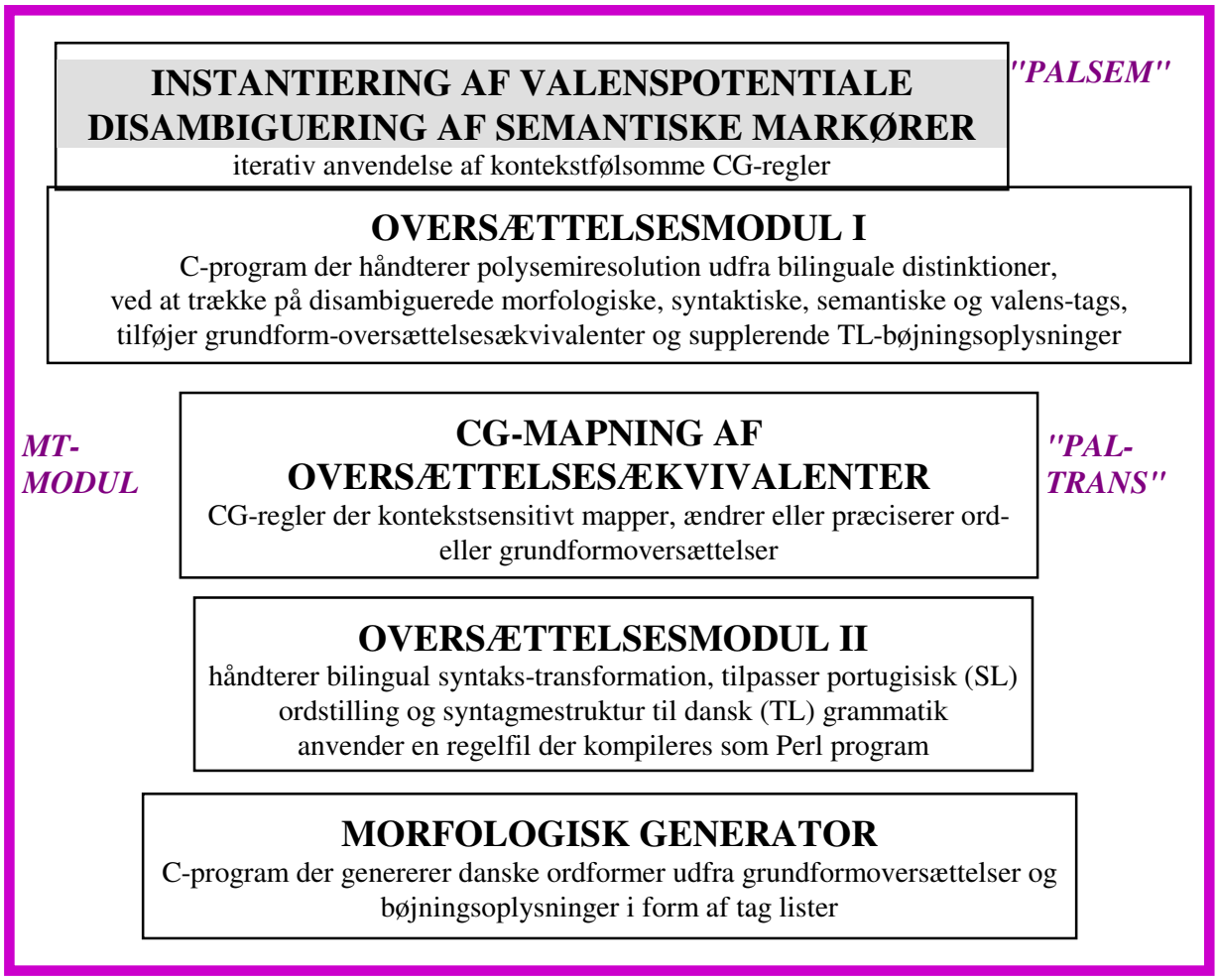
2. Modulprogredient Constraint Grammar parsing

Udfra et ønske om lingvistisk stringens skelnes der i CG normalt mellem et morfologisk (tagging) og et syntaktisk (parsing) niveau, der behandles successivt-progredient[4], med et parsing-leksikon som det logiske udgangspunkt. Jeg har i mit system bibeholdt denne modulære teknik, og søgt at vise at metoden er særdeles velegnet også til yderligere analyseprogression, ikke alene hen imod en stadig mere fintmasket syntaks, men også men henblik på notationel filtration (herunder konstituent- eller dependensbaserede træstrukturer) og, i sidste ende, semantiske distinktioner.

Fig.2: Tagging/Parsing moduler i PALAVRAS-systemet



[4] Distinktionen forekommer også "teknisk" naturlig, idet behovet for tilføjelsen af supplerende kontekstafhængig information (ud over de rent leksiko-morfologiske oplysninger) vokser med analysens tiltagende kompleksitet. Dette opnås med den for Constraint Grammar typiske alternerende succession af disambiguering- og mapping-regler, hvor sidstnævnte leverer nyt (i.e. ikke primært leksiko-morfologisk) input til førstnævnte.



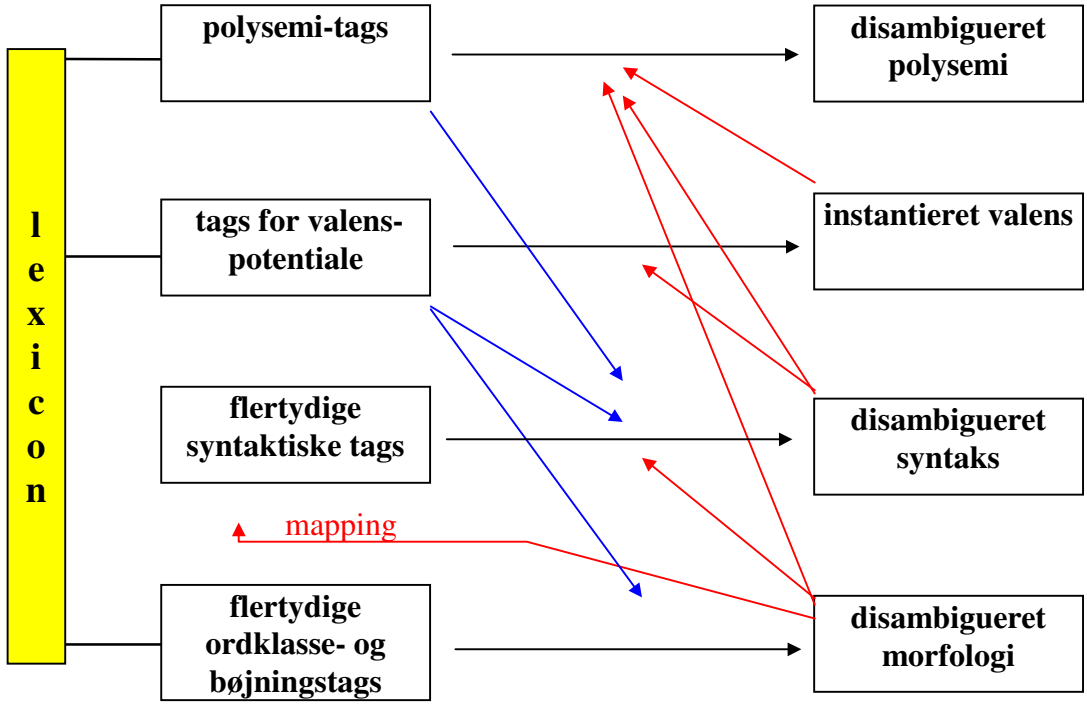
Den portugisiske parsers progression på det syntaktiske område - herunder ledsætningsfunktion, udvidet dependensmarkering og træstrukturer - frembringer kvantificerbare resultater der kan måle sig med hvad der er opnået for "benchmark"-syntaksen i engelske ENGCG og FDG, og selvom en egentlig semantisk CG kun er realiseret som pilotprojekt, er der tale om et fungerende system for fri løbende tekst, der viser at Constraint Grammar i princippet kan løfte parsing-opgaver på dette niveau, og mange af de brugte distinktionstræk er implementeret for *hele* leksikonet. De semantiske CG-regler selv håndterer dels en omfattende valensinstantiering, dels polysemiresolution på udvalgte områder.

Selv om valensinstantiering teknisk ville kunne realiseres på mange andre måder, viser systemet at CG formalismen kan håndtere hvad der grundliggende svarer til en unifikationsproces, hvor der trækkes på tilstedeværelsen (eller fraværet) af bestemte - førdisambiguerede - syntaktiske tags. Mens disse "valensunifikationsregler" er forholdsvis simple og kun har sekundær semantisk værdi, er de egentlige polysemiregler mere komplekse og inddrager - foruden syntaksen - også potentiel flertydig semantisk kontekst. Endeligt har jeg - som sidste niveau - lanceret et CG niveau til mapping (og ændring) af portugisisk-danske oversættelsesækvivalenter, hvor reglerne raffinerer og korrigerer allerede valgte oversættelser (der er fremkommet ved at trække på morfologiske, syntaktiske, valens- og semantiske tags som polysemi-diskriminatorer).

Er man indstillet på at betale prisen i form af regelkompleksitet og -volumen, synes der således ikke at være *principielle* begrænsninger mht. hvilke niveauer af *grammatiske* distinktioner der kan håndteres ved brug af Constraint Grammar formalismen. Afgørende for CG-reglernes potentiale synes snarere at være kvaliteten

og mængden af informationen i systemets leksikon, samt kvaliteten og mængden af disambiguerede tags på forudgående (lavere) analyseniveauer, og tilgængeligheden af sekundære (ikke-disambiguerede) tags fra senere (højere) analyseniveauer:

Fig.1: Vekselvirkninger imellem parsing-niveauer



For eksempel kan det være vanskeligt at identificere ordet 'a' som direkte objekt (@ACC) alene på baggrund af morfologisk og flertydig kontekst, men opgaven lettes betydeligt hvis verbers valenspotentiale er kendt, og efter at verberne er blevet ordklasse-disambigueret. Tilsvarende forudsætter disambigueringen af et flertydigt valens- eller semantisk potentiale gennemførelsen af den *syntaktiske* analyse. Således vil <+HUM> blive valgt i tagstrengen af et substantiv, der - af de syntaktiske regler - er blevet identificeret som subjekt af et kognitivt eller tale-verbum. Dette hindrer dog ikke <+HUM> i at være en nyttig *sekundær* tag allerede på det syntaktiske niveau. Faktisk er progressionen fra sekundær til primær tag[5] typisk og essentiel for parserens progressivt modulære opbygning. Samtidigt gør denne progression det muligt at udskyde vanskelige disambiguerings-opgaver til et senere, mere informationsrigt, niveau.

En generel lingvistisk fordel ved modulprogredient parsing (Progressive Level Parsing) er at forskellige lingvistiske systemer og klassifikationer kan holdes adskilt. Således var det muligt at definere ordklassekategorier primært morfologisk (igennem inventaret af bøjningskategorier[6]), uden at miste den syntaktiske og semantiske information indeholdt i de traditionelle ordklassedefinitioner (der - hvor ønsket - kan

[5] Ved primære tags forstås her tags der - på det aktuelle niveau - er genstand for regelbaseret disambiguering, mens sekundære tags kun indgår som mulige kontekstbetingelser, uden - endnu - selv at blive disambigueret.
[6] Inspireret af (Arndt, 1992).

"re-hybridiseres" ved at lade et filterprogram trække på også sekundære og syntaktiske tags).

Transformationen af pronomen-subkategorier kan tjene som eksempel for et teoriafdrevent tag-filter. I mit system skelnes der mellem 3 pronomenklasser der alle defineres morfologisk: Personlige pronominer (PERS), numerus-genus-bøjelige pronominer (DET) og ubøjelige pronominer (SPEC). Syntaktisk funktionelt kan alle erstatte hele NP'er i rollen som subjekt, objekt etc., men kun DET optræder prænominalt (@>N). De 3 klasser svarer i traditionel portugisisk grammatik til 6 "pseudosemantiske" pronomenklasser[7] og en "funktionel-syntaktisk" artikel-klasse:

(2) **Table: pronoun subclass filtering**

Traditionel pronomen-klasse	**CG tags**
personligt pronomen	PERS
possessivt pronomen	DET <poss>
demonstrativt pronomen	DET/SPEC <dem>
interrogativt pronomen	DET/SPEC <interr>
relativt pronomen	DET/SPEC/ADV <rel> @#FS/AS-N<
indefinit pronomen	DET <quant1/2/3> SPEC <quant0> DET/SPEC <rel> ¬ @#FS/AS-N<
artikel	DET <art>/<arti>

Som det fremgår, bruges både ordklasse-tags, sekundær-leksiske tags og syntaktiske tags i filteret, svarende til den ret hybride og idiosynkratiske karakter af de traditionelle pronominalklasser.

Traditionelt beskrives de portugisiske ordformer *a, as, o, os* enten som bestemte artikler, eller - i forbindelse med en PP-modifikator eller en relativsætning - som demonstrative pronominer. I mit system bruges derimod kun én ordklasse (DET), der så suppleres med 2 leksisk bundne sekundære tags, <art>[8] og <det> der funktionelt disambigueres af GC-regler på valens-niveau'et, hvor både syntaktisk funktion og ordklassekontekst er sikkert etableret.

Tilsvarende betragtes sætningsindledende relative adverbier ikke som konjunktioner på det morfologiske niveau, men opnår en implicit "konjunktionalitet" på det syntaktiske niveau igennem sætningsindledertags (@#FS).

Også leksemafgrænsningen håndteres først og fremmest morfologisk, uden at inddrage semantiske kriterier, - der først introduceres som sekundære semantiske tags af ét og samme leksem, og senere disambigueres på det semantiske niveau.

3. Kappelystne parsing-paradigmer

Automatiske grammatiske annotationssystemer kan iflg. Karlsson (1995) klassificeres efter i hvilken grad det enkelte system er enten regelbaseret eller

[7] Sometimes the category of *reflexive pronoun* is added, which would have to be filtered as a syntactic subclass of personal pronouns: <refl> PERS.

[8] For these words, <art> implies <artd> (definite article), which is the icon used as a set definition and in output filtering.

probabilistisk (eller hybrid). Som en yderligere parameter kunne man tilføje leksikaliseringsgraden, idet der også indenfor "lejrene" er stor forskel på mængden og granulariteten af den leksiske information der ligger til grund for analysen. Probabilistiske HMM[9]-baserede systemer har typisk haft gode resultater på de lavere analyseniveauer (ordklasse-tagging, fonetisk struktur i talegenkendelse), mens regelbaserede systemer, især PSG[10] og beslægtede genskrivningsgrammatikker (HPSG, DCG), har en lang tradition på det syntaktiske område.

I takt med fremkomsten af store korpora med mange millioner ord er manuel korrektur af automatisk annoterede tekster blevet mere og mere problematisk, og både processeringshastighed, robusthed og fejlmarginaler har fået stadigt større betydning, og selv om regelbaserede systemer i teorien burde have et større potentiale rent sprogbeskrivningsmæssigt, synes de store korpora at favorisere de probabilistiske systemer, dels ved at levere tiltrængt træningsmateriale, dels fordi mange regelbaserede systemer er udviklet kun for små leksiske eller grammatiske subsegmenter af sproget. I dette dilemma kan Constraint Grammar som reduktionistisk metode forene de probabilistiske systemers robusthed ved fritekst-analyse med de regelbaserede systemers lingvistiske finmaskethed.

3.1. Probabilistisk tagging

Den store fordel ved probabilistiske taggere er analysehastigheden, den korte træningstid og netop det faktum at det ikke er nødvendigt at skrive en egentlig regelgrammatik. At introducere en regelbaseret bias i en probabilistisk tagger ved at introducere regler for ord med en uregelmæssig adfærd kan endda føre til negative interferencer mht. systemets håndtering af de regelmæssige ("majoritets-") tilfælde (Chanod and Tapanainen, 1994). Men et HMM-system har kun et meget begrænset syntaktisk vindue (typisk bigrammer eller trigrammer), et faktum der skyldes at sandsynligheden $p(t_n|t_1 \ldots t_{n-1})$ sættes lig med $p(t_n|t_{n-1})$ (for bigrammer), eller med $p(t_n|t_{n-1}t_{n-2})$ (for trigrammer). Desuden fortyndes systemets leksiske kollokationsviden som regel af at der benyttes rene ordklasse-tags uden fleksionsoplysninger. I en generativ grammatik håndteres den syntaktiske struktur derimod på en eksplicit måde, og fungerer parallelt som beskrivelsens objekt og disambigueringsværktøj. I en CG, endeligt, *er* der muligt at håndtere syntaktisk struktur i hele sætningen, men analysen fremkommer som en slags sidegevinst af sekventielle kontekstuelle disambigueringsregler, og processen er derfor mere robust end i en generativ grammatik. På den anden side råder CG-formalismen - som den bruges i dag - kun over forholdsvis grove redskaber til udnyttelsen af statistiske tendenser i kollokationsmønstre, - som at markere leksikonopslag som <Rare>, eller at ordne regler i successive heuristiske grupper, uden mulighed for en mere matematisk nøjagtig tildeling af sandsynlighedsværdier til enkelte regler.

Probabilistiske PoS-taggere

System	*korrekt*

[9] Hidden Markov Model, en matematisk model hvor en overflade-sekvens af symboler genereres stokastisk af en underliggende (skjult) Markov-proces med en tilstands- eller transitionsstyret symbolgenerator.

[10] De her nævnte grammatiktraditioner (PSG = Phrase Structure Grammar, HPSG = Head Driven Phrase Structure Grammar, DCG = Definite Clause Grammar) udtrykker i modsætning til Constraint Grammar strukturel information direkte, ved at genskrive syntaktiske enheder som sekvenser af (terminale eller non-terminale) konstituenter. Konsekvensen er en større gennemsigtighed af det grammatiske regelsystem, men også en større følsomhed i tilfælde af ufuldstændige regler eller normafvigende tekstinput.

HMM trænet på taggede corpora	Church 1988	97%
HMM trænet på taggede corpora	Garside 1987	97%
HMM trænet på utaggede corpora	Cutting 1992	96%
Maksimum-entropi-tagger	Ratnaparki 1996	96.5-97%
probabilistiske selvlærende algoritmer	Brill 1997	96.8%
probabilistisk	Lezius et.al. 1996	

Probabilistiske syntaktiske parsere

System		*recall*	*præcision*
LR-parser (Susanne-korpus) (3 højest prioriterede læsninger)	Carrol & Briscoe 1995	73.56%	39,82%
Kerne-dependent-relationer (Wall Street J.)	Collins 1996	85%	85%

Moderne PoS-taggere opnår i dag korrekthedsprocenter på 96-97%, - nærmest uanset metode. HMM-systemer trænet på taggede korpora (97%, Church 1988 og Garside 1987) har således kun en marginalt bedre performans en systemer trænet på utaggede korpora (96%, Cutting et.al. 1992), og lignende fejlprocenter rapporteres for en maksimum-entropi-tagger (96.5-97%, Ratnaparkhi 1996 og http://www.hd.uib.no/corpora/1997-3/0086.html, 22.10.97) samt for systemer baseret på selvlærende algoritmer (96.8%, Brill 1997). Resultater med andre sprog end engelsk synes at bekræfte 97%-asymptoten som en slags øvre performansgrænse for probabilistiske morfologiske/PoS-taggere (fx. 95.9% for tysk, Lezius et. al. 1996).

Resultaterne for probabilistisk *syntaktisk* parsing er væsentlig dårligere. Carrol & Briscoe (1995), for eksempel, beskriver en probabilistisk LR-parser med en recall (for data fra Susanne-korpus'et) på 73.56% og en præcision på 39.82% for de 3 højestprioriterede læsninger for hver sætning. Collins (1996) rapporterer 85% recall og præcision for en probabilistisk beskrivelse af kerne-dependent-relationer i WSJ[11]-tekst.

CG vs. Probabilistisk analyse

Samuelsson & Voutilainen 1999	*recall*	
	ENGCG	**statistisk tagger**
ved 93.5% *præcision*	99.9 %	97.2 %
ved 97.5% *præcision*	99.57 %	96.28 %

I en nyere direkte sammenligning[12] mellem en opdateret ENGCG og en statistisk tagger[13] konstaterer Samuelsson & Voutilainen (1999), at Constraint Grammar systemet havde en mindst 10 gange lavere fejlrate end det probabilistiske system på tilsvarende ambiguitetsniveauer. ENGCG havde således en fejlprocent på 0.1% ved en tag/ord-kvotient på 1.07, og en fejlprocent på 0.43% ved en tag/ord-kvotient på 1.026, mens fejlprocenterne for den statistiske tagger var 2.8% hhv. 3.72%.

[11] Wall Street Journal

[12] Begge systemer brugte det samme tag-sæt (de modulære CG-tagsene blev filtreret til syntetiske tags) og begge systemer blev testet på den samme blandede benchmark-tekst (50.000 ord).

[13] Taggeren blev trænet på 357.000 ord fra Brown-korpus'et, - den træningstest-størrelse hvor taggerens indlæringskurve blev asymptotisk.

3.2. Generativ parsing

De generative grammatikker der anvendes i syntaktisk parsing (fx. GB, HPSG), postulerer traditionelt et sprogligt system der er stabilt og har klare grænser for hvad der er korrekt, hvor målet er at generere "alle og kun de" sætningslæsninger der er kompatible med det opstillede sprogsystem. Det inkorporerede forholdsvis præskriptive sprogsyn indebærer en risiko for at bevisbyrden i tilfælde af (for parseren) uanalyserbart sprogligt input flyttes fra grammatikken til korpusset, hvorimod probabilistiske og CG-baserede systemer nærmest aksiomatisk accepterer at "korpusset altid har ret", og leverer analyser også for ufuldkomment input, leksiske huller etc.

I en sammenligning mellem Constraint Grammar og Generativ Grammatik, kan man skelne mellem konceptuelle og implementeringsmæssige forskelle. I et konceptuelt perspektiv er CG - i modsætning til en PSG - parsing-orienteret og systemets grammatiske viden kan ikke bruges til at *generere* sætninger. I CG defineres ambiguitet i et metasprog der ikke er direkte strukturbunden, og disambigueringsprocessen er derfor mere fleksibel. CG er reduktionistisk snarere end generativistisk, hvad der gør grammatikken mere tolerant (eller robust) over for hvad chomskyansk grammatik ville kalde "performance failures": Ufuldstændige ytringer, dialektal variation etc. Implementeringsmæssigt er en væsentlig forskel mellem CG- og PSG-paradigmerne, at ambiguiteten, i en CG, reduceres *graduelt* og uden *retracing*, ved at fjerne ordbaseret information i stedet for direkte at opbygge strukturel information.

Det er som regel vanskeligt at sammenligne performansen af grammatikbaserede systemer med den af probabilistiske systemer, idet systemets deskriptive potentiale vægtes højere end den praktiske værdi i analysen af *fri* tekst, f.eks. ved at øge den grammatiske kompleksitet på bekostning af leksikonets dækningsgrad. Blandt undtagelserne er ANLT-værktøjerne (Phillips & Thompson 1987, baseret på GPSG) eller TOSCA-projektet (Oostdijk 1991, Extended Affix Grammar), der begge arbejder med fri løbende tekst. Men selv ved fuld leksisk og grammatisk dækning er det et systemimmanent problem ved *lange* sætninger at en generativ grammatik overgenererer - eller udspecificerer ægte syntaktiske ambiguiteter - i en sådan grad at den resulterende "analyseskov" består af hundred- eller tusindvis af træer. En løsningsmulighed er hybridsystemer med en probabilistisk indeksering af genskrivningsreglerne (og efterfølgende prioritering af træerne). Wauschkuhn (1996) anvender i en tysk parser manuelt tildelte såkaldte "sikkerhedsfaktorer" eller "plausibilitetsfaktorer" til regelindeksering.

3.3. CG-regler i hybridsystemer

Også CG-lignende regler er blevet anvendt i hybridsystemer. Lindberg (1998) trænede en selvlærende algoritme ("Inductive Logic Programming") på et svensk tagged korpus med det formål at formulere eller selektere REMOVE-regler for en lokal kontekst på ±2 ord. Systemet opnåede 98% recall med en resterende ambiguitet af 1.13 læsninger pr. ord, og et grammatikvolumen på 7000 regler. Et andet hybridsystem, baseret på Relaxation Labelling, beskrives i Padró i Cirera (1997) for engelsk og spansk. Her integreredes morfologiske CG-regler i en HMM-tagger for at opbygge en statistisk model over distributionen af *tag targets* og kontekstbetingelser. CG-reglerne blev dels automatisk akvireret fra et træningskorpus vha. statistiske beslutningstræer, dels skrevet "i

hånden" på baggrund af output-fejl i probabilistiske HMM-taggere[14]. Begge typer CG-regler forbedrede performansen i forhold til både HMM-tagging og relaxation labelling isoleret set. Det samlede system opnåede en recall på 97.35% for fuldt disambigueret WSJ-tekst.

Hybrid-systemer (probabilistisk + CG)

System		recall	præcision
Inductive Logic Programming (Automatisk formulering/selektering af REMOVE-regler)	Lindberg 1998	98%	88.5%
Relaxation Labelling (WSJ text) (Automatiske *og* håndskrevne regler)	Padró i Cirera 1997	97.35%	100%

Mens regelbaserede hybridsystemer således synes at kunne forbedre performansen af probabilistiske HMM-baserede ordklasse-taggere, er de opnåede fejlprocenter stadigvæk langt højere end i de publicerede rene CG-systemer. En mulig forklaring er begrænsningen til regler uden global (sætnings-) kontekst (dvs. regler med fx. trigram-kontekst), og det faktum at lingvist-formulerede regler (i modsætning til automatisk akvirerede regler) kun har fundet begrænset anvendelse.

4. Den Portugisiske Parser

4.1. Typologi

I den voksende familie af CG-baserede taggere/parsere er portugisiske PALAVRAS p.t. det eneste fuldt udviklede og dokumenterede system for et romansk sprog, og det har derfor en vis typologisk interesse at sammenligne den portugisiske parser med de andre eksisterende CG-systemer, ikke mindst det veldokumenterede engelske system (Karlsson et.al., 1995).

Rent notationelt viser en tag-oversigt (kapitel 8.1) over de grammatiske kategorier i CG'erne for indoeuropæiske sprog (portugisisk, engelsk, svensk, norsk, tysk) et stort overlap hvad angår ordklassekategorier, mens forskellene er større på det syntaktiske niveau. Her udskiller den portugisiske parser sig ved dels at tilføje dependensmarkører også på sætningsniveauet, dels ved en konsekvent tagging af ledsætningsfunktion[15].

[14] Der brugtes kun 20 lingvist-formulerede regler, mod 8473 automatisk genererede constraints.

[15] I den morfologiske del anvendes 13 ordklasse-kategorier, der kombineres med 24 tags for bøjningsformer, hvad der resulterer i flere hundrede distinkte komplekse tags. Denne analytiske karakter af CG-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). Det syntaktiske tag-sæt råder over 40 tags for ord/syntagme-funktion og ca. 30 tags for ledsætningsfunktion (der dækker over tre slags ledsætninger: finitte, infinitte og averbale). Der arbejdes med ca. 100 valensklasser (især for verber), og ca. 200 semantiske prototyper (især for substantiver).

Også ud fra et CG-teknisk perspektiv er foskellene mellem det engelske og det portugisiske system større på de syntaktiske end det morfologiske område. Således udviser begge sprog en morfologisk ambiguitet på omkring 2 (læsninger pr. ordform) før disambiguering, og V/N-ambiguiteten er et problem ikke kun for det fleksionsfattige engelske sprog, men også for portugisisk, hvor 3 af de 4 typiske nominale bøjningsmorfemer ('-o', '-a', '-as') samtidigt forekommer i verbale bøjningsparadigmer. Men selv om begge sprog mangler morfologiske subjekt- og objektmarkeringer for substantiver, kan en engelsk CG udnytte ordstillingsregler (fx. SVO-rækkefølge i fremsættende sætninger) i højere grad end et portugisisk system. Til gengæld er det en strukturel fordel for den portugisiske CG at sætningsindledere her er obligatoriske for finitte ledsætninger.

En kvantitativ regeltypologi for PALAVRAS' CG-regelsæt synes at underbygge ENGCG-resultater der peger på at den venstre sætningskontekst har større betydning for disambigueringen end højre-konteksten, svarende til den lineære og sekventielle struktur af naturligt sprog. For portugisisk er andelen af venstre-kontekster 60%, både for bundne og frie kontekster[16], mens ENGCG-tallene er 81% for frie og 42.6% for bundne kontekster (Karlsson, 1995, p. 352). En vigtig forskel mellem de morfologiske og de syntaktiske regler i PALAVRAS er at proportionen mellem frie og bundne kontekstbetingelser er 10 gange højere i syntaktiske regler (2.0 resp. 0.2), og at regelkompleksiteten er højest for de syntaktiske regler (gennemsnitlig 5.28 kontekster) og lavest for de morfologiske regler (3.37 kontekster).

Ser man på disambigueringsgevinsten ved brugen af regler med forskellige targets, synes det for portugisisk at være substantivdisambigueringen der giver den største disambigueringsgevinst, siden substantiv-ordformer udviser en høj ambiguitet, og ingen særlig disambiguerings-bias (i modsætning til infinitiverne der også er ambiguøse, men har en kraftig bias imod de alternative læsninger).

4.2. Performans

Fejlprocenterne for tag-baserede parsere (typisk probabilistiske og CG-systemer) udtrykkes som regel ved at beregne annotationens *præcision* og *recall,* der måler proportionen mellem antallet af *overlevende korrekte læsninger* og henholdsvis (a) *overlevende læsninger i alt (præcision)* eller (b) *alle korrekte læsninger (recall).* I et CG-system er *præcisionen* et mål for den overlevende ambiguitet (efter en eller flere regelbaserede disambigueringsrunder) og kan approksimeres ved automatisk optælling i analyseret fritekst, så længe ambiguiteten stadigt er stor - og antallet af fejlforkastede læsninger lille. Derimod kan *recall* kun kvantificeres for korrigerede benchmark-korpora eller ved manuel optælling i mindre testtekster. Efterhånden som disambigueringen nærmer sig 100% (hvor der med undtagelse af de få tilfælde af ægte ambiguitet kun er én overlevende læsning per ordform), vil tallene for *alle korrekte læsninger* og *overlevende læsninger i alt* nærme sig hinanden og blive lig ordtallet. *Præcision* og *recall* vil derfor begge være afhængige alene af antallet *af (overlevende) korrekte læsninger* - med andre ord, *præcision* nærmer sig *recall.* I denne situation kan *recall*-tallene derfor betragtes som et direkte mål for parserens performans, og jeg vil i det følgende bruge det mere generelle udtryk *correctness* ("korrekthed") i betydningen af *recall ved 100% disambiguering.*

I det følgende vises og diskuteres resultaterne fra analysen af en række forskellige testtekster.

[16] Ved bundne kontekster forstås kontekstbetingelser i CG-regler hvor en præcis position i forhold til target-ordet angives, mens kontekstbetingelserne i frie kontekster kan instantieres overalt mellem target-ordet og sætningsgrænsen.

Fig.3: PALAVRAS' performans (fri løbende tekst)

Teksttype	**Tekststørrelse**	**Morfologi**	**Syntaks**	**Syntaks alene**
(1) "O tesouro" (novelle)	ca. 2500 ord	99.3 %	97.4 %	98.5 %
(2) "VEJA" (nyhedsmagasin, tema: politik)	ca. 4800 ord	99.7 %	97.3 %	97.8 %
(3) "VEJA" (nyhedsmagasin, tema: politik)	ca. 3140 ord	99.2 %	96.4 %	97.3 %
(4) "VEJA" (tema: videospil)	2412 ord	98.8 %[17]	97.3 %	98.8 %
(5) "VEJA" (tema: kunst)	1837 ord	99.6 %	97.5 %	97.9 %
(6) "Globo rural" (landbrugsavis)	ca. 2380 ord	99.7 %	97.5 %	97.8 %
(7) "Isto é" (nyhedsmagasin, tema: samfundsstof)	ca. 2920 ord	99.6 %	98.0 %	98.6 %
(8) Didaktisk fagtekst (tema: biologi)	ca. 6020 ord	99.7 %[18]	98.0 %	98.5 %
Samlet	ca. 26000 ord	99.5 %	97.5 %	98.2%

Oversigten viser, at parseren typisk opnår korrekthedsprocenter på over 99% i den morfologiske (ordklasse-) analyse, og 97% - 98% i den syntaktiske del.

Det skal bemærkes, at fejlene ikke er jævnt fordelt over hele teksten, men snarere optræder i grupper i nogle sætninger, med helt fejlfrie sætninger herimellem. Årsagen er interdependensen mellem morfologiske og syntaktiske fejl. Således kan fx. en N/V-ordklassefejl afføde 2 eller 3 syntaktiske fejl omkring sig. Fejlinterferensen betyder også at det syntaktiske parser-modul alene, dvs. når det forsynes med morfologisk fejlfri tekst som input, ville kunne opnå noget bedre resultater end i den integrerede analyse (forskellen er typisk på 0.5-1 procentpoint, jf. sidste kolonne i tabellen).

En nærmere gennemgang af fejltyperne viser, at de valgte avistekster (2-7) adskiller sig fra fiktions- (1) og fagprosa (8) både leksikalsk og syntaktisk. For det første møder man blandt fejlkilderne en stor andel af komplekse egennavne, forkortelser og tidstypiske engelske låneord, og 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.) samt indskudte "syntaktisk overflødige" finitte verber i form af citationsrammer. I det store og hele forekommer performansen imidlertid stabil på tværs af forskellige tekstgenrer.

Et pilotprojekt med transskriberet talesprog (fra det brasilianske NURC-korpus, Castilho 1989) viste ved en sammenligning med skriftsprogs-resultater en stor forskel på hhv. den morfologiske og den syntaktiske robusthed. Således var de

[17] Dette relativt lave tal kan forklares ved en række fejl i overskrifter, samt ved at en tredjedel af de morfologiske fejl skyldes et enkelt ord ('console'), et engelsk ord der ikke figurerede i taggerens leksikon, og undslap den heuristiske analyse fordi ordet også kan tolkes som bøjningsform af det portugisiske verbum 'consolar' ('trøste').

[18] Den i forvejen høje morfologiske korrekthedsprocent for denne testtekst stiger til svimlende 99.9% hvis tekstindscanningsfejl og overskriftsfejl fratrækkes.

morfologiske fejlprocenter kun marginalt lavere end for skriftsprogskilder, selv ved en umodificeret grammatik, mens den syntaktiske korrekthedsprocent faldt til ca. 91% før, og 95% - 96% efter en regel-tilpasning af systemet[19].

Fejlprocenterne i tabellen skal desuden ses i lyset af det ret differentierede tag-sæt. Således kan parserens detaljerede dependens- og funktionsoplysninger for præpositional-syntagmerne (som fx. postnominalled @N<, adverbielt postadjekt @A<, adverbielt adjunkt @<ADVL, @ADVL>, @ADVL, adverbielt objekt @<ADV, @ADV>, præpositionelt objekt @<PIV, @PIV>, subjektsprædikativ @<SC, frit prædikativ, @<PRED eller forbinderledsargument @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 (4) og (5).

CG-performans (morfologisk)

	recall	*precision*
ENGCG (engelsk)	99.7%	93-97%
SWECG (svensk)	99.7%	95%
PALAVRAS (portugisisk)	99.5%	ca. 100%

Også performansmæssigt viser tallene for de enkelte CG-systemer en større variation på det syntaktiske end på det morfologiske niveau. Således oplyses der både for engelske ENGCG og svenske SWECG morfologiske fejlprocenter på 0.3%, ved en disambigueringsgrad på hhv. 93-97% og 95%. Den gennemsnitlige fejlprocent for PALAVRAS (kap. 3.9 i afhandlingen) er 0.5%, dvs. lidt højere, men er til gengæld målt ved en nær fuldstændig (dvs. næsten 100%) disambiguering. På det syntaktiske niveau rapporterer Voutilainen & Tapanainen (http://www.conexor.fi) morfosyntaktiske successrater på 94.2-96.8% for ENGCG (ved 11.3-13-7% ambiguitet) og 96.4-97% for FDG[20] (ved 3.2-3.3% ambiguitet). PALAVRAS opnår for portugisisk de samme korrekthedsprocenter som FDG for engelsk (i.e. ca. 97%), selv med en ambiguitetsprocent nær 0%. Kun for 2 CG-baserede parsere, FDG og PALAVRAS[21], foreligger performansdata efter transformation til syntaktiske træstrukturer.

CG-performans (syntaktisk)

	recall	*precision*
ENGCG (engelsk)	94.2 - 96.8 %	86.3 - 88.7 %
FDG (engelsk)	96.4 - 97 %	96.7 - 96.8 %
PALAVRAS (portugisisk)	97.5 %	ca. 100%

[19] Et vigtigt led i tilpasningen til talesprogsdata var introduktionen og disambigueringen af såkaldte *dishæsions-markører* i transskriptionen (fx. pauser, interjektioner og andre småord der kan bruges til at definerer mere præcise analysevinduer for den syntaktiske CG-analyse).

[20] Functional Dependency Grammar

[21] Et spansk CG-trægenererings-system og en engelsk overbygnings-CG med et trægenererende modul er under udvikling og kan afprøves på http://visl.hum.sdu.dk.

En direkte sammenligning med den engelske FDG tyder på at PALAVRAS opnår en noget højere recall for (portugisiske!) subjekter og objekter, og en lidt dårligere recall for subjektsprædikativer. Sammenligningen må dog siges pga. notationelle forskelle at være langt mere problematisk end for almindelige CG-analyser.

En tag-for-tag evaluering[22] af en testtekst på 5000 ord viser for portugisisk et gennemsnitlig fald på 0.3% i både præcision og recall, hvis fejl i træstrukturen lægges oven i fejlprocenten for syntaktiske tags alene.

5. "Flade" træstrukturer i CG-syntaks

5.1 Syntaktisk form og syntaktisk funktion

De fleste CG-systemer benytter sig af en morfologisk toniveau-analyse som præprocessor (TWOL, jf. Koskenniemi 1983), og fokuserer på morfologiske træk og ordklasser. Den grammatiske beskrivelse er derfor i høj grad ordbaseret og implementeres ved at hæfte tags til ordformer.

Historisk set udspringer CG fra morfologisk analyse, og "flad syntaks" er en naturlig konsekvens af dette, og også i min parser benyttes en flad repræsentation af syntaktisk struktur (Bick 1997-1). 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 kerne og samler konstituenten til en kohærent helhed med implicitte syntagmegrænser.

Hvor kernen ikke er hovedverbet, bliver det anført ved pilespidsen (fx. N for nominal-hoved, A for adjekt-hoved[23]). Dependensmarkører bliver enten hæftet til de funktionelle tags (fx. @<SUBJ, @ADVL>, @N<PRED), eller står, ved visse bestemmerled, alene (fx. @>N for [bestemmer-] prænominalled).

(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 M P	@>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 eller bøjle den selv hænger i (kernen, 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

[22] Selvom der i systemet indgår et CG-modul der tagger koordinatorer for hvad de koordinerer, blev paratagmer ikke som sådanne transformeret til træstrukturer på evalueringstidspunket. Det kan således ikke udelukkes at fejlprocenten vil være noget højere efter en eventuel introduktion af paratagmer.

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

faktisk godt skære den i stykker og gemme den i en skotøjsæske - den strukturelle information bevares[24].

I eksemplet befinder 'muito' sig langt nede i uroen, men kender sin 'adverbial-adjekt'-snor (@>A) til 'velho'. Dette adjektiv fastgøres så 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.

Den ordbaserede opmærkning indebærer imidlertid en risiko for, at dependensbeskrivelsen begrænses til syntagmeniveauet, og "ignorerer" mange mere komplekse, især ledsætningsrelaterede, dependens- og funktionsrelationer.

Min løsning har været (a) at forsyne *alle* de syntaktiske tags med direktionelle dependensmarkører (jf. ovenfor), og (b) at hæfte 2 tags til de centrale forbinderord (subordinerende konjunktioner, relativer og interrogativer) i finitte og averbale ledsætninger, samt til infinitiver, gerundier og participper i infinitte ledsætninger[25]. 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 disse to tag-typer som adskilte, således at de kan disambigueres uafhængig af hinanden, af distinkte regelmoduler.

(6)	Sabe	[saber] <vq> V PR 3S IND	@FMV		*[Han] ved*
	que	[que] KS	**@#FS-<ACC**	**@SUB**	*at*
	os	[o] <art> DET M P		@>N	*[de]*
	problemas	[problema] N M P		@SUBJ>	*problemerne*
	são	[ser] <vK> V PR 3P IND		@FMV	*er*
	graves	[grave] ADJ M/F P		@<SC	*alvorlige*

[@FMV = finit hovedverbum, @#FS-<ACC = finit ledsætning, der fungerer som direkte (akkusativ-) objekt og styres af et hovedverbum til venstre, @SUB = subordinator, @>N = prænominalbestemmer, @SUBJ> = subjekt for hovedverbum til højre, @<SC = subjektsprædikativ med (kopula-) hovedverbum til venstre, V = verbum, KS = subordinerende konjunktion, DET = determiner, N = substantiv, ADJ = adjektiv, PR = præsens, IND = indikativ, 3S = 3. person singularis, 3P = 3. person pluralis, M = maskulinum, F = femininum, S = singularis, P = pluralis, <art> = artikel, <vq> = kognitiv verbum (med que-valens), <vK> = kopula-verbum]

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** '[det]'
baque [baque] <cP> N M S **@SUBJ>** 'faldet'

[24] 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.

[25] 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 nummererede "links" mellem kerne og dependenter.

VP	foi	[ser] <x+PCP> V PS 3S IND VFIN @**FAUX** 'blev'
	atenuado	[atenuar]<vt><sN>V PCPMS @**IMV**@**#ICL-AUX<**'dæmpet'
PP-PASS	por	[por] <sam-> <+INF> <PCP+> PRP @**<PASS** 'af'
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 [at]'
SUBJ	sua	[seu] <poss 3S/P> DET F S @**>N** 'hans'
	mulher	[mulher] <H> N F S @**SUBJ>** 'kone'
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** 'et'
	emprego	[emprego] <stil> <ac> N M S @**<ACC** 'job'
SUBJ & **FS-N<**	que	[que] <rel> SPEC M/F S/P @**SUBJ>** @**#FS-N<** 'som'
	garante	[garantir]<vt><vq>V PR 3S IND VFIN @**FMV**'garanterer'
ACC	as	[a] <art> DET F P @**>N** 'de'
	despesas	[despesa] <ac> N F P @**<ACC** 'udgifter'
	básicas	[básico] <jn> ADJ F P @**N<** 'basale'
PP-N<	de	[de] <sam-> PRP @**N<** 'af'
P<	a	[a] <-sam> <art> DET F S @**>N** 'den'
	família	[família] <HH> N F S @**P<** 'familie'

finit relativ ledsætning, postnominal modifikator
infinit (infinitiv-) ledsætning, argument af præposition
præpositionssyntagme, passivagent-adjunkt
finit hovedsætning

[@>N =prænominal-modifikator, @SUBJ> =subjekt (til venstre for hovedverbum), @FAUX =finit hjælpeverbum (kernen i verbalkæden), @IMV =infinit hovedverbum, @AUX< =auksiliarargument, @<PASS =passivagent, @P< =styrelse (præpositionsargument), @<ACC =direkte (akkusativ-) objekt (til højre for hovedverbet), @N< =postnominal- modifikator, @FMV =finit hovedverbum]

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

[NP =substantivsyntagme, PP =præpositionssyntagme, FS =finit sætning, ICL =infinit sætning, S =hovedsætning]

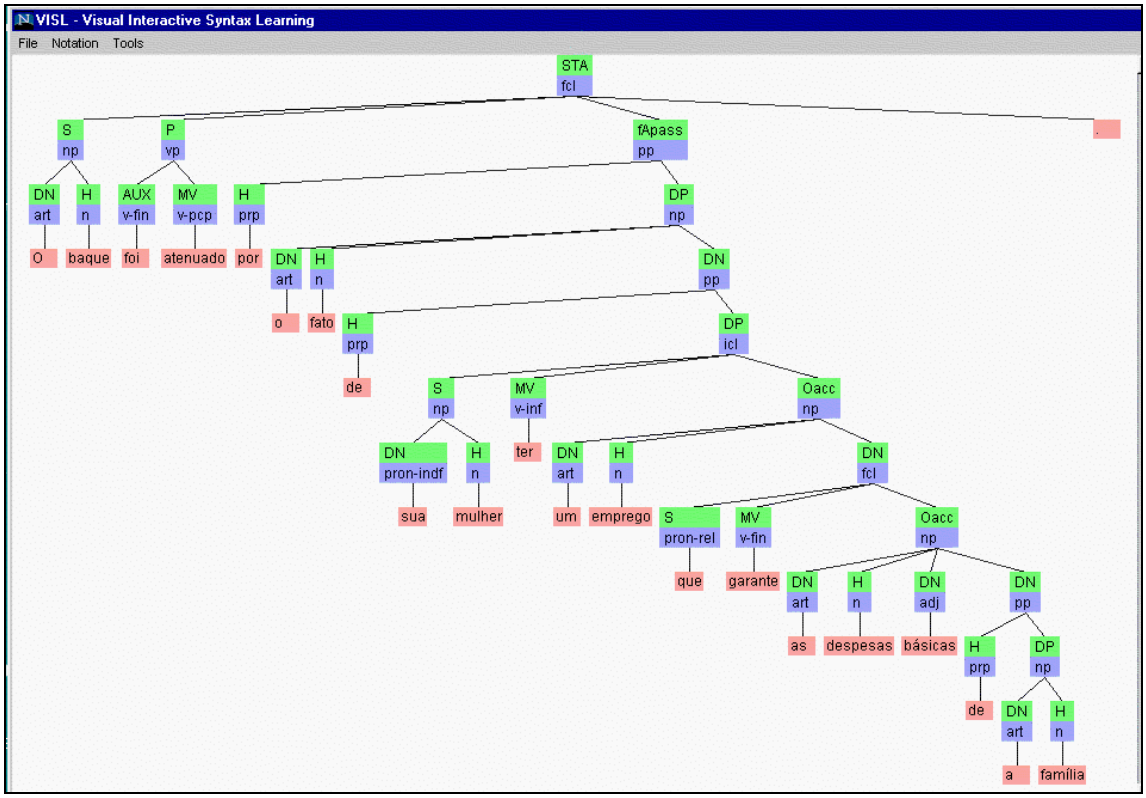
5.2. Transformation af flad dependenssyntaks til træstrukturer

I lyset af den store popularitet og pædagogiske pondus som konstituentgrammatikkerne nyder i moderne lingvistik, er det nærliggende at undersøge om en flad CG-syntaks på denne måde kan opnå en vis ækvivalens og "transformerbarhed" i forhold til en konstituentbaseret trænotation. For at vise, at dette godt kan lade sig gøre, har jeg skrevet et kompilerprogram for substitutionsregler[26] der identificerer syntagme- og ledsætningsgrænserne i en flad

[26] I det engelske VISL-system har jeg som alternativ mulighed realiseret en deklarativ generativ genskrivningsgrammatik.

CG-beskrivelse, markerer dem som *form* (np, pp, icl m.m.) og tildeler dem som*funktion* kernens syntaktiske CG-tag[27]. Herefter kan der genereres input til grafiske programmer som det interaktive Java-program der er implementeret af Martin Carlsen i den didaktiske VISL-brugerflade (http://visl.hum.sdu.dk):

Fig.4: Grafisk træanalyse



Her sigtes primært på syntaks-undervisning, og det konkrete træ-eksempel viser kun ét af flere notationelle valg, hvor de oprindelige CG-termer delvis erstattes med andre, pædagogisk motiverede, symboler. Således indledes alle objekt-symboler med et stort O, og alle gruppe-dependenter med et D. Af strukturelle tilpasninger kan nævnes at den flade beskrivelse af verbalkæder (som dependenshierarki af hjælpeverber og hovedverbum) er opgivet til fordel for prædikatorgrupper (P:vp). Systemet er beskrevet i kompendiet *Portuguese Syntax* (Bick 1999).

En vigtig forskel mellem den flade CG-notation og træ-notationen er, at sidstnævnte - på godt og ondt - *skal* opløse visse flertydigheder[28], som den flade syntaks underspecificerer, fx. i forbindelse med tilhæftningen af postnominaler (især præpositionssyntagmer), koordination og frie prædikativer. Denne underspecificering bliver imidlertid nærmest til et gode, når man betragter den udfra et MT-perspektiv: For det første er mange af tilfældene eksempler på "ægte *syntaktisk* flertydighed", der kun kan tydes af den fuldt kontekstualiserede - menneskelige - lytter/læser. Og for det andet er en række af disse strukturelle

[27] Som nævnt i 3.2., stiger fejlprocenten ved denne transformation til træstrukturer kun med 0.3 procentpoint i forhold til den rene CG-analyse.

[28] enten - i tilfælde af en generativ overbygningsgrammatik (jf. engelsk VISL) - ved at generere flere træer, eller - i tilfælde af en rent transformativ overbygningsgrammatik (jf. portugisisk VISL) - ved at forkaste nogle af mulighederne og kun vise ét af de mulige træer.

ambiguiteter forholdsvis universelle, dvs. sproguafhængige, således at de kan bevares i oversættelsen, der baseres direkte på den "flade" beskrivelse (c). At gøre en sådan flertydighed eksplicit, for et sprogpar der ellers håndterer den éns, forekommer unødvendigt kompliceret.

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

Der findes dog på den anden side også en del tilfælde hvor en trægenereringsgrammatik formår at opløse en sådan (i forhold til CG-beskrivelsen) strukturelt øget ambiguitet vha. sætningsintern information alene. Adjektiviske bestemmere, enten postnominalt eller som frie prædikativer, kan således i portugisisk udvise kongruensrelationer mellem kerne og bestemmerled (e), der ikke kommer til udtryk på dansk:

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

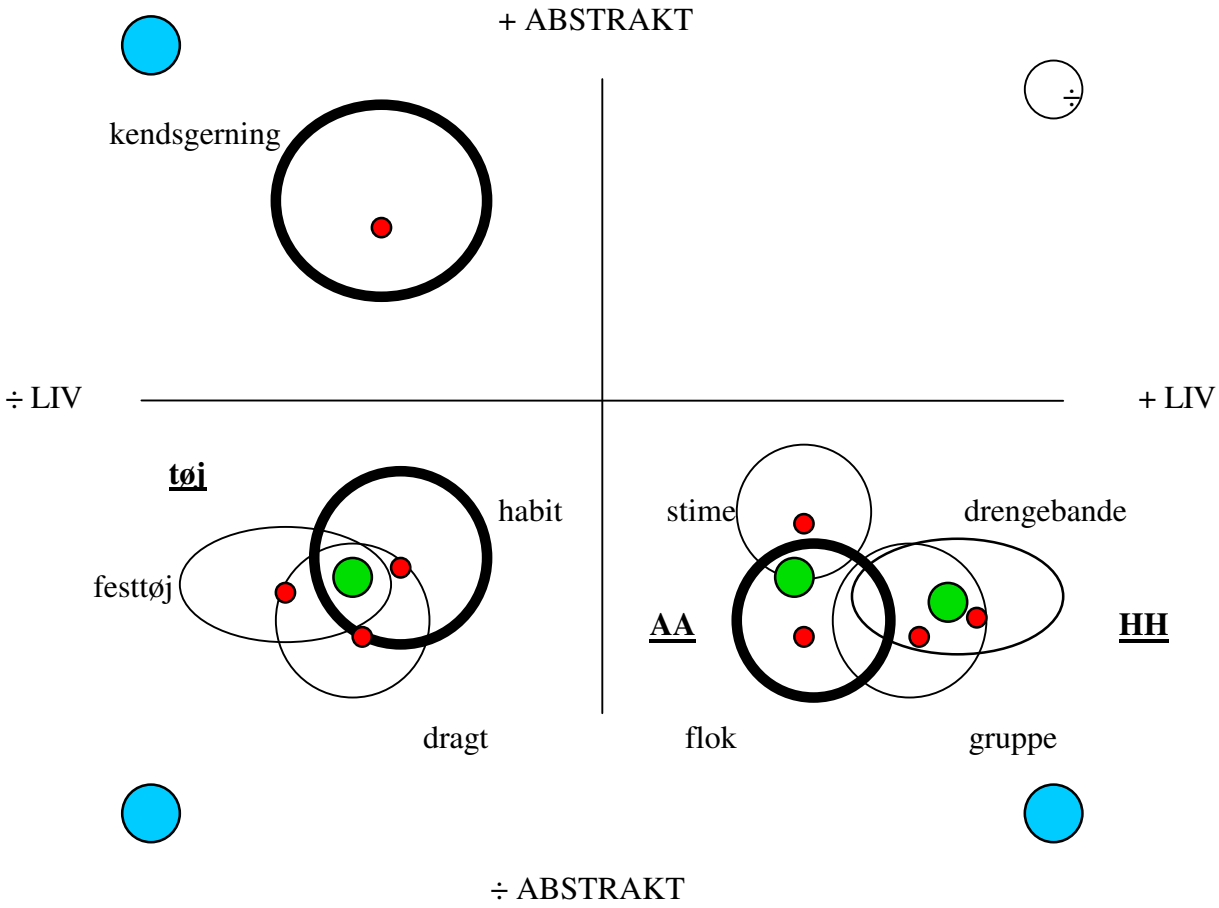
Træstrukturen for (e) - men ikke for (d) - ville altså være entydig og den flade CG-syntaks notationelt overlegen.

6. Det semantiske perspektiv

Forskellige applikationer kræver forskellige grader af modulprogression. Således er en morfologisk analyse tilstrækkelig for en forsker der arbejder med korpusbaserede frekvensanalyser, mens den internetbaserede grammatikformidling i projektet VISL kræver en syntaktisk analyse, og maskinoversættelse en semantisk.

Det er imidlertid min bedste empiriske overbevisning at Constraint Grammar som redskab kan "presses" til stadigt højere analyseniveauer - 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 - mindre almindeligt - "flok" ved hjælp af *atomare semantiske træk* som henholdsvis *abstrakt & ikke levende* ("kendsgerning"), *ikke abstrakt & ikke levende* ("habit") og *ikke abstrakt & levende* ("flok"). Disse træk er tilmed 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 CG-parser kan et hierarki af leksikon- og kontekst-drevne grammatiske regler "forbyde" eller "selektere" disse træk eller prototypiske trækfamilier[29] i den konkrete sætning.

[29] 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.



Diagrammet placerer en række ord i et semantisk felt, i forhold til hinanden 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 underopdeling af ÷ABSTRAKT.

En særlig elegant og "inkrementel" løsning for polysemireduktion af indholdsmæssigt flertydige ord er den semantiske udnyttelse af "lavere parsing-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 (her: aspekt). 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 præpositionsobjekt 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 præpositionsobjekt, a^ =præposition "a", de^ =præposition "de"; morfologi: IMPF =imperfektum, PERF =perfektum]

Endeligt kan de semantisk entydige (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 leksikonet, i form af semantisk beriget valensinformation (såkaldte selektionsrestriktioner). 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 mængde-ord)*
__ <+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 leksikonet, hvoraf nogle er valensinstantieringer (<+que>, <+de+que>, <+de+INF>), nogle semantiske prototyper (<ac><tøj><AA>) og den sidste en oplistning 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=>

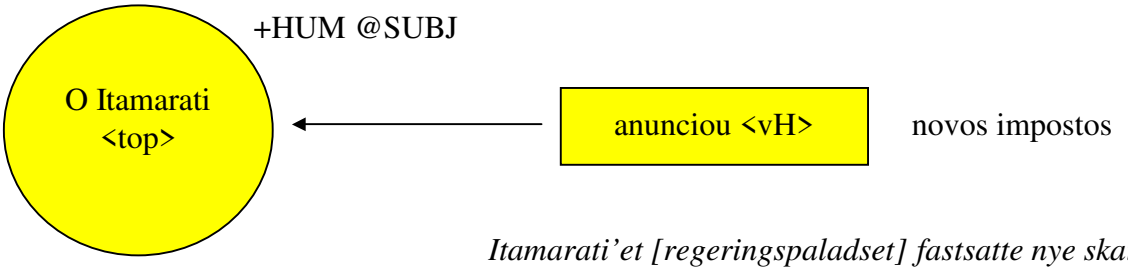
__ <ac><+de+que><+de+INF>	kendsgerning
__ <tøj>	habit, kostume
__ <AA>	flok {fx. geder}
fato=de=banho N M	badedragt

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 entydige kontekstbetingelser.

(12a)

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

De samme CG-regler der afstemmer de semantiske selektionsregler i en sætning, og udnytter dem til polysemiresolution, kan i princippet også bruges til at håndtere produktive metaforer. Verber som 'anunciar' (meddele) eller 'falar' (tale), for eksempel, bærer <vH> tags der "kræver" et subjekt med trækket +HUM. Forudsat det syntaktiske CG-modul har identificeret subjektet korrekt, kan denne projektion af egenskaben +HUM udnyttes til semantisk disambiguering af et polysemt subjekt (fx. 'cara' - 'ansigt'/'fyr'), men den kan også - i tilfælde af et éntydigt -HUM subjekt, tolkes som metaforisk transfer, som i følgende eksempel:



Itamarati'et [regeringspaladset] fastsatte nye skatter

7. En korpusnær grammatik

Når Constraint Grammar-konceptet tillægges attributet 'robust', sker dette som regel ud fra metodologiske overvejelser, dvs. med fokus på CG som en *parsing teknik* (bl.a. fordi den reduktionistiske tilgang garanterer at der altid vil være én læsning der overlever disambigueringen). Jeg vil imidlertid hævde at Constraint Grammar, som det praktiseres af dets nuværende forskersamfund, også udmærker sig ved at være robust som *grammatisk system.*

For det første udarbejdes CG-grammatikker typisk i et korpusbaseret miljø, hvor regelskrivningen og regelkorrektur foregår på baggrund af løbende kvantitativ performanskontrol, der sikrer at grammatikken forbliver tæt på det "naturlige" sprog, og aldrig fredes for syntaktisk variation, produktiv morfologisk derivation etc., idet hverken leksikon eller grammatik begrænses til små eksperimentelle subsegmenter af det undersøgte sprogsystem.

For det andet mener jeg at have dokumenteret for portugisisk, at en CG-inspireret flad dependensgrammatik er *notationelt* robust dels som udgangspunkt for en konvertering til andre grammatiske systemer, dels som en selvstændig og deskriptivt elegant, ordbaseret repræsentation af *syntaktisk* struktur, der inden for samme formalisme kan håndtere flere grammatiske analyseniveauer, og som gør det muligt, at underspecificere syntaktisk flertydighed (fx. koordination og postnominale

præpositionssyntagmer) der først meningsfuldt ville kunne opløses på et semantisk-pragmatisk kontekstualiseret niveau.

For det tredje er den implicitte grammatiske sprogbeskrivelse der ledsager udarbejdelsen af en CG-parser, immanent empirisk på en unik måde der sikrer en særlig form for "grammatisk autenticitet". Fordi der foruden nye regler også løbende introduceres nye tag/leksem-sæt og nye sekundærtags i systemet, får korpusdata og korpusmotiverede distinktioner lov at forme selve grammatikken, og man kunne tale om en korpusdreven analysemodel. Denne proces er grundliggende forskellig fra den statistiske, leksikografiske og stilistiske måde korpora normalt bruges på i humanistisk forskning. Kategorien <vq>[30] (kognitivt verbum), for eksempel, blev introduceret "on the fly", som en hybrid syntaktisk kategori med en semantisk interpretation, ikke *a priori*, men pga. behovet for en valenspotentiale-markør til disambigueringen af 'que'-styrede objekt-ledsætninger (svarende til danske finitte at-sætninger). På samme måde blev kategorien 'ergativt verbum' defineret empirisk *a posteriori*, med udgangspunkt i korpuseksempler hvor verber havde en øget frekvens af efterstillede subjekter (@<SUBJ). Og arbejdet med aviskorpora førte på et tidspunkt til introduktionen af et set for tale-verber (V-SPEAK), som viste sig nødvendig for korrekt at kunne tagge efterstillede, ofte sætningsfinale, subjekter i citat-konstruktioner som i: *"....."*, *diz <V-SPEAK> Fernando Santiago (@<SUBJ), 47, de São Paulo.* I denne proces bliver CG-forskerens oprindelige grammatiske overbevisninger og intuitioner løbende kontrolleret og modificeret af systemets empiriske "behov", og simplicitet og effektivitet af den deskriptive model konkurrerer på lige fod med rent teoretiske overvejelser, når det gælder udseendet af den resulterende (også implicitte) sprogbeskrivelse[31].

En robust korpus-baseret parser med en korpus-dreven grammatik har korpus-annotation som et naturligt applikationsområde, og PALAVRAS har i den senere tid fundet anvendelse i en række større eksterne tagging-opgaver.

PALAVRAS-taggede korpora

Corpus	Size		Genre
ECI-corpus*	ca.	670.000	mixed genre Brazilian text
VEJA 1996	ca.	600.000	news magazine text
NURC	ca.	100.000	urban speech data (Brazil)
Folha de São Paulo	ca.	90.000.000	running newspaper text
Tycho Brahe**	ca.	50.000	historical Portuguese
NILC-corpus*	ca.	39.000.000	jounalistic, didactic (Brazilian)
CORDIAL-SIN**	ca.	30.000	dialectal speech data (Portugal)

[30] En lignende kategori, <Vcog>, bruges også i den engelske CG beskrevet i Karlsson et. al. (1995). For portugisisk blev listen over "kognitive" verber kompileret ud fra korpus-søgninger på verbale ordformer fulgt af 'que' eller et interrogativ.

[31] Disse empiriske og proces-styrede distinktioner *kan* komme i konflikt med grammatikkerens teoretiske ståsted eller applikative intentioner, og har her åbenlyse metodologiske begrænsninger. Den ord- og tag-baserede notation gør det imidlertid muligt i hvid omfang at løse problemet igennem post-processering af CG-output'et, således at CG-systemet som sådant sikres stor metodologisk uafhængighed.

NATURA*	ca. 7.250.000	newspaper text (Portugal)
CETEMPúblico*	ca. 180.000.000	mixed text (Portugal)
all	ca. 320 million words	written language, speech data historical texts

*internet-searchable: * http://www.portugues.mct.pt, ** http://corp.hum.sdu.dk*

Således blev 90 millioner ord (3 års løbende avistekst fra *Folha de São Paulo*) annoteret med det originale PALAVRAS-tagsæt for et brasiliansk universitet, og PALAVRAS-annoterede versioner af flere store brasilianske og portugisiske corpora er blevet tilgængeliggjort på http://www.portugues.mct.pt, i samarbejde med det Oslo-baserede AC/DC-projekt. Tagningen af et 180 millioner ord stort portugisisk avis-korpus er lige påbegyndt. I flere tilfælde[32] har det været muligt at matche "fremmede" morfologiske tag set igennem en simpel ned-filtrering fra PALAVRAS-tagsættet.

Morfosyntaktisk annoterede korpora gør det muligt at løse mere komplekse forskningsopgaver end det er tilfældet med simpel tekstsøgning, og den ordbaserede notation tillader en nem tilpasning til forskellige output- eller søgekonventioner, som fx. når ordklasser farvekodes til padagogiske formål i VISL-brugerfladen:

8. Konklusion

[32] Der er tale om dele af det historiske Tycho-Brahe-korpus og det blandede NILC-korpus, der begge blev tagget for at gennemføre en sammenligning med resultaterne fra probabilistiske parsere.

Parsing-systemet PALAVRAS implementerer en korpus-dreven grammatisk model af portugisisk skriftsprog i et Constraint Grammar-miljø med systematisk modulprogression. Ved automatisk, fuldt disambigueret, ordform-tagging af fri løbende tekst opnås fejlprocenter på under 1% i den morfologiske, og 2-3% i den syntaktiske analyse.

På det morfologiske plan synes portugisisk, et stærkt flekterende sprog med relativ fri ordstilling, at udvise en lignende grad af ambiguitets- og regelkompleksitet som engelsk, et flektionsfattigt sprog med fast ordstilling, et faktum der underbygger Constraint-Grammer-traditionens påstand om formalismens universalitet og sproguafhængighed[33].

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 på det semantiske plan 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.

Som programsystem har parseren vist sig at være robust og effektiv nok til at kunne integreres i applikative kontekster som fx. maskinoversættelse, grammatiske tutoring systemer[34] og grammatiske filtre til korpussøgning. Et særligt perspektiv ligger herudover i den potentielle portabilitet af PALAVRAS' (portugisiske) CG-grammatik. Efter udarbejdelsen af de nødvendige sprogspecifikke monolinguale leksika og morfologiske programmer, for bøjningsanalyse og derivation, har det således vist sig muligt direkte at eksportere store dele af det portugisiske CG-regelsæt til tilsvarende grammatikker for spansk, esperanto og - på det morfologiske plan - dansk[35].

Bibliografi

Arndt, Hans, "Towards a syntactic analysis of Danish computer corpora", in Heltoft, Lars & Haberland, Hartmut (eds.): *Proceedings of the 13th Scandinavian Conference of Linguistics,* Roskilde, 1992

Bick, Eckhard, *Leksikografiske overvejelser i forbindelse med udarbejdelsen af en portugisisk-dansk ordbog* (cand.mag.-speciale), Århus, 1993

Bick, Eckhard, *Portugisisk - Dansk Ordbog,* Mnemo, Århus, 1993, 1995, 1997

Bick, Eckhard, "Automatic Parsing of Portuguese", i Sánchez García, Laura (ed.): *Proceedings of the Second Workshop on Computational Processing of Written Portuguese,* Curitiba, 1996

[33] Sproguafhængighed gælder formalismen og compiler-implementeringen, *ikke* de enkelte regler, der ikke kan overføres fra et sprog til et andet.

[34] Parseren er blevet forsynet med en tilsvarende (prototypisk) brugerflade i forbindelse med VISL-projektet ved Institut for Sprog og Kommunikation, SDU *(Visual Interactive Syntax Learning).*

[35] Som man måske kunne forvente, var portabiliteten mest udpræget for spansk, hvor de fleste ændringer var systematiske, og ofte kunne håndteres ved at erstatte portugisiske ordformer og leksemer med tilsvarende spanske former i sæt-definitioner og ordbaserede regler. I mit aktuelle danske CG-projekt (http://visl.hum.sdu.dk) stammer ca. halvdelen af de nuværende morfologiske regler direkte fra PALAVRAS.

Bick, Eckhard, "Dependensstrukturer i Constraint Grammar Syntaks for Portugisisk"*, i:* Brøndsted, Tom & Lytje, Inger (eds), *Sprog og Multimedier,* pp. 39-57, Aalborg, 1997

Bick, Eckhard, "Automatisk analyse af portugisisk skriftsprog", i: Jensen, Per Anker & Jørgensen, Stig. W. & Hørning, Anette (eds.), *Danske ph.d.-projekter i datalingvistik, formel lingvistik og sprogteknologi,* pp. 22-20, Kolding, 1997

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

Bick, Eckhard, "Structural Lexical Heuristics in the Automatic Analysis of Portuguese", i: Maegaard, Bente (ed.): *The 11th Nordic Conference on Computational Linguistics (Nodalida '98), Proceedings,* pp. 44-56, Copenhaguen, 1998

Bick, Eckhard, "Portuguese Syntax", Århus, 1999 (undervisningskompendium)

Brill, Eric, "Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging", i *Natural Language Processing Using Very Large Corpora,* Kluwer Academic Press, 1997.

Briscoe, Ted & Carroll, John, "Generalised LR Parsing of Natural Language (Corpora) with Unification-Based Grammars", i *Computational Linguistics,* 19(1): 25-60, 1993.

Castilho, Ataliba Teixeira de (ed.), *Português culto falado no Brasil,* Campinas, 1989

Chanod, Jean-Pierre & Tapanainen, Pasi, "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

Church, Kenneth, "A stochastic parts program and noun phrase parser for unrestricted text", i *Proceedings of the Second Conference on Applied Natural Language Processing,* Austin, Texas, 1988. (through Cutting, 1992)

Collins, Michael John, "A New Statistical Parser Based on Bigram Lexical Dependencies", i *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics,* Santa Cruz, USA, 1996.

Cutting, Doug & Kupiec, Julian & Pedersen, Jan & Sibun, Penelope, "A Practical Part-of-Speech Tagger", i *Proceedings of the Third Conference on Applied Natural Language Processing,* ACL, Trento, Italy, 1992

Garside, Roger & Leech, Geoffrey & Sampson, Geoffrey (eds.), *The Computational Analysis of English. A Corpus-Based Approach,* London, 1987

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

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

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

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

Lezius, Wolfgang & Rapp, Reinhard & Wettler, Manfred, "A Morphology-System and Part-of-Speech Tagger for German", ", i: Gibbon, Dafydd (ed.): *Natural Language Processing and Speech Technology*, Berlin, 1996

Lindberg, Nikolaj, "Learning Constraint Grammar-style disambiguation rules using Inductive Logic Programming", i: *Proceedings of COLING/ACL '98,* volume II, pp. 775-779, Montreal, 1998 og http://www.speech.kth.se/~nikolaj/papers/colingac198/ (15.5.1999)

Oostdijk, Nelleke, *Corpus Linguistics and the Automatic Analysis of English,* Amsterdam, 1991.

Padró i Cirera, Lluís, *A Hybrid Environment for Syntactic-Semantic Tagging* (Dissertation ved Universitat Politècnica de Catalunya, Barcelona, 15.12.1997), postscript internet version fra 11.2.1998

Phillips, John D. & Thomsen, Henry S., "A Parser for Generalised Phrase Structure Grammars", i: Haddock, Nicholas & Klein, Ewan & Morrill, Glyn (eds.), *Categorial Grammar, Unification Grammar and Parsing* (Edinburgh Working Papers in Cognitive Science, Vol.1), pp. 115-136 , Edinburgh, Centre for Cognitive Science, University of Edinburgh, 1987 (igennem Karlsson, 1995)

Ratnaparkhi, Adwait, "A Maximum Entropy Part-Of-Speech Tagger" in *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania , 1996

Samuelsson, Christer & Voutilainen, Atro, "Comparing a Linguistic and a Stochastic Tagger", i *Proceedings of the 35th Annual Meeting of the Association for Computional Linguistics and the 8th Conference of the European Chapter of ACL,* Madrid, 1999 (forthcoming), og http://www.ling.helsinki.fi/~ãvoutila/cg/doc/e-ac197/e-ac197.html (10.11.99)

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

Tapanainen, Pasi, *A Dependency Parser for English,* University of Helsinki, Department of Linguistics, Technical Reports, No. TR1, 1997

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

Wauschkuhn, Oliver, "Ein Werkzeug zur partiellen syntaktischen Analyse deutscher Textkorpora", i: Dafydd Gibbon (ed.): *Natural Language Processing and Speech Technology*, Berlin, 1996