MORFOSYNTAKTISK OPMÆRKEDE KORPORA FOR DANSK: KORPUS90/2000 OG ARBORETUM

Af Eckhard Bick (Institut for Sprog og Kommunikation, Syddansk Universitet)

1. Introduktion

En lang række lingvistiske applikationsområder, herunder statistisk baseret sprogbeskrivelse, leksikografi, informationssøgning og maskinoversættelse, efterlyser stadig større tekstkorpora til forskning og programudvikling. Imidlertid er kvaliteten af disse korpora ikke kun afhængig af deres størrelse og kompileringsparametre som sproglig variation og tekstuel-sociologiske kildeoplysninger, men også af graden af lingvistisk bearbejdning af materialet. Man kan her skelne mellem simple strukturelle parametre (fx markering af ord- og periodegrænser), morfologisk *tagging* og syntaktisk *parsing,* samt evt. opmærkning af semantiske og pragmatiske forhold. Inspireret af tilsvarende materiale for andre sprog, indgik VISL-projektet ved Syddansk Universitet (http://visl.sdu.dk) og Det Danske Sprog- og Litteraturselskab (http://www.dsl.dk) i 2001 en aftale om automatisk grammatisk opmærkning af DSL's korpusmateriale, det nuværende Korpus90 og Korpus2000 (Asmussen 2002), i alt ca. 60 millioner ord, der som sætningsrandomiserede citatkorpora begge tillader internetbaseret tilgængeliggørelse uden større ophavsretslige problemer.

2. Korpus-opmærkning

Opmærkningen blev gennemført på det morfosyntaktiske niveau med DanGram-systemet (Bick 2001), en flerniveau-parser baseret på Constraint Grammar-paradigmet (Karlsson 1995, Bick 2000) - en metode der tillader automatisk opmærkning af løbende tekst med stor robusticitet og en forholdsvis lille fejlprocent[1]. Hvert ord i teksten tildeles ud over ordklasse- og fleksionsoplysninger en syntaktisk *tag*, der angiver dels en grammatisk funktion (fx. subjekt @SUBJ, adverbial @ADVL), dels ordets dependensrelation (fx. venstre/højre nominaldependent eller verbalkomplement). Nedenstående tabel viser en statistik over samtlige funktionstags i Korpus2000. Pilespidserne angiver dependensen, @SUBJ> "peger" således på et verbum til højre ("**Julemanden** kommer i dag."), @<SUBJ peger til venstre ("I dag kommer **julemanden**."). En dobbelt pilespids til højre angiver "raising", fx @>>P for styrelsen ("**Hvem** danser hun med?").

[1] Regelkompileringen blev gennemført med Pasi Tapanainens cg2 (Tapanainen 1996) og Martin Carlsens beslægtede GNU-version, vislcg (frit tilgængelig på http://visl.sdu.dk/download/).

tag	kategori	ord		finitte sætninger		infinitte sætninger		averbal sætn.	alle
		n	%	n FS	%	n ICL	%	n AS	%
<ACC	akkusativ-objekt	1.483.771	5,2	251.932	0,9	98.603	0,3		6,4
<ADVL	adverbial	2.433.163	8,5	230.451	0,8	1.608	0,0	247	9,3
<DAT	dativ-objekt	58.555	0,2						0,2
<OA	objektsrel. arg.-adverbial	124.335	0,4			23.120	0,1		0,5
<OC	objektsprædikativ	53.156	0,2			1.845	0,0		0,2
<PIV	middelbart (præp.) objekt	428.072	1,5						1,5
<PRED	frit (subjekts-)prædikativ	47.550	0,2			880	0,0		0,2
<SA	subjektsrel. arg.-adverbial	294.245	1,0	1.062	0,0	198	0,0		1,0
<SC	subjektsprædikativ	761.732	2,7	27.780	0,1	11.818	0,0	301	2,8
<SUBJ	subjekt	844.611	3,0	96.006	0,3	50.502	0,2		3,5
>>A	adverbiel depend., raised	603	0,0						0,0
>>P	styrelse, raised	47052	0,2	287	0,0				0,2
>A	adverbiel dependent	391.585	1,4						1,4
>AUX	auksiliardependent	0	-			55	0,0		0,0
>N	adnominel dependent	4.392.546	15,3						15,3
>P	præpositionsfokus	73.892	0,3						0,3
>S	subordinatorfokus	8561	0,0						0,0
A<	adverbiel dependent	148.463	0,5	2.031	0,0	8.631	0,0	27	0,5
A<<	adverbiel dep., forskudt	0	-			1.142	0,0		0,0
ACC>	akkusativ-objekt	89.765	0,3	101.371	0,4	68	0,0		0,7
ADVL	adverbial (uden sætning)	145.505	0,5			7	0,0		0,5
ADVL>	adverbial	955.348	3,3	104.326	0,4	1.230	0,0	46	3,7
ADVL>>	adverbial, raised	2	0,0						0,0
APP	apposition	61.992	0,2						0,2
AS<	sætn.-stamme i averbal s.	445	0,0	1	0,0	2	0,0		0,0
AUX<	auksiliardependent	0	-			1.115.394	3,9		3,9
CO	koordinator	1.150.098	4,0						4,0
DAT>	dativ-objekt	2.953	0,0						0,0
F-<ACC	formelt akkusativobjekt	2.519	0,0						0,0
F-<SUBJ	formelt subjekt	73.089	0,3						0,3
F-SUBJ>	formelt subjekt	166.853	0,6						0,6
FAUX	finit hjælpeverbum	501.099	1,8						1,8
FMV	finit hovedverbum	1.442.917	5,0						5,0
FOC>	fokus-markør	35.268	0,1						0,1
IAUX	infinit hjælpeverbum	596	0,0						0,0
IMV	infinit hovedverbum	10.426	0,0						0,0
INFM	infinitivmarkør	461.741	1,6						1,6
KOMP<	komparativargument	39.255	0,1	17.624	0,1	7	0,0		0,2
MV<	verbalpartikel	148.111	0,5						0,5
N<	postnominaldependent	1.392.561	4,9	515.538	1,8	36.331	0,1		6,8
N<FUSE	substantivkompositum	23.295	0,1						0,1
N<PRED	prædikativ, np-niveau	216.255	0,8			283	0,0	60	0,8
NPHR	nominal (uden sætning)	193.538	0,7						0,7
OA>	objektsrel. arg.-adverbial	144	0,0						0,0
OC>	objektsprædikativ	92	0,0						0,0
P<	styrelse	3.449.268	12,0	157.836	0,6	292.854	1,0		13,6
PIV>	middelbart (præp.) objekt	1.128	0,0						0,0
PRED>	frit (subjekts-)prædikativ	18.409	0,1			237	0,0		0,1
S<	sætningsdependent	0	-	12.658	0,0				0,0
S-<SUBJ	situativt subjekt	4.801	0,0						0,0
S-SUBJ>	situativt subjekt	15.113	0,1						0,1
SA>	subjektsrel. arg.-adverbial	3.518	0,0						0,0
SC>	subjektsprædikativ	19.188	0,1						0,1

STA	fremsættnede ytring	0	-	8.309	0,0				0,0
SUB	subordinator	705.646	2,5						2,5
SUBJ>	subjekt	2.493.997	8,7	9.663	0,0	18.004	0,1		8,8
SUBJ>>	subjekt, raised	3.126	0,0						0,0
TOP	topic	227	0,0			14	0,0		0,0
VOK	vokativ-led	1.271	0,0						0,0
X	dummy-funktion	1.310	0,0						0,0
i alt: 28.623.149 tokens		25.422.761	88,8	1.536.874	5,4	1.662.833	5,8	681	100

En egentlig evaluering af data'ene i tabellen ligger hinsides denne artikels rammer, men tallene kan hjælpe at sætte det opmærkede korpus i et kvantitativt relief. Eksempeltvis fremgår det tydeligt, at dansk - statistisk set - har en SVO-struktur, men også, at VSO er langt mere almindelig end OVS, idet ca. hver 4. subjekt er efterstillet, mens venstrestilling af objekter er sjælden og i øvrigt begrænset til bestemte kontekster, især citeret tale, relativ- og interrogativpronominer (Bick 2002). For substantivisk @ACC> (i alt under 0,3% af alle syntaktiske tags) var distributionen - i et korpusudsnit på 1.1 millioner ord - følgende:

Subtype	**n**	**Frekvens**	**Definition**
interrogativ	79	29.0 %	at se, **hvilken interesse** kineserne skulle *have*
topic	74	27.2 %	**Denne interesse** *overførte* han på virksomheden **De problemer** *har* jeg slet ikke.
fokus	55	20.2 %	**Blot 6-7 kr.** vil sparekassen *se* som betaling **Sin spillefilmsdebut** *fik* han i 1962 med ...
frontstilling i verbalkæde	43	15.8 %	... få **tyvekosterne** *bragt* hjem ... får man **billeder** at *se* gratis ... at lære **de nødvendige redskaber** at *kende*
raised	12	4.4 %	**Den slags** er vi jo nogle stykker der kan *lide*
faste vendinger	7	2.6 %	Hvad **udvalget af værker** *angår*, har ...
negativt inde i vp	2	0.7%	... at min søn **ingen huller** *havde* ... hun har **ingen kage** *bagt*

Den beskrevne ordbaserede opmærkning tillader en forholdsvis enkel filtrering til alternative notationssystemer, samt til html- eller sgml-opmærkning (jf. konkordansformatet på http://corp.hum.sdu.dk), men det strukturelle informationsindhold i korpusset kan yderligere øges, visualiseres og tilgængeliggøres ved at bruge CG-output som input til en særlig PSG-grammatik, der leverer en egentlig konstituentanalyse med eksplicit specificering af syntagmegrænser. Resultatet er en "skov" af syntaktiske træer (arbejdsnavn for dette korpus: "Arboretum"), der dels tillader manipulation med de grafiske VISL-redskaber, dels søgning/ekstraktion af fx hele substantiv- syntagmer eller bestemte syntagmesekvenser (http://corp.hum.sdu.dk/arboretum.html).

En automatisk opmærkning kan dog i sagens natur aldrig være fejlfri[2], og jo større distinktionsniveauet, desto større behovet for "manuel" korrektur. Et nyligt påbegyndt projekt er derfor, som led i det tværnordiske projekt PaNoLa *(Parsing Nordic Languages)* , med lingvistøjne at revidere dele af det automatisk opmærkede Korpus90/2000, der herefter vil kunne bruges som en slags *gold standard* til evalueringsformål, parserudvikling (herunder også statistiske systemer), dokumentation og undervisning. Den lingvistiske revision følger den automatiske opmærknings to trin, således at der opnås to versioner af samme korpus, en dependensgrammatisk og en konstituentgrammatisk.

3. Leksikografiske muligheder

Den syntaktiske opmærkning tillader bl.a. at søge for sekvenser af nominal subjekt (N @SUBJ>) - hovedverbum (@MV) - nominal akkusativobjekt (N @<ACC). En enkel grep-søgning med efterfølgende filtrering af leksemformerne i konstituentkernerne tillader et tre-ordsformat af typen "hest - æde - hø" med verbet i infinitiv og substantiverne i singularis nominativ (med mindre det fx netop er numerus der ønskes undersøgt). Ved at ordne ekscerpterne efter verbet og sekundært objekt eller subjekt, fås leksisk information mht. selektionsrestriktioner i verbets valensmønster.

DanGram-parseren arbejder med et tag sæt af ca. 200 forskellige semantiske prototyper for substantiver (en oversigt over disse kategorier fås på http://visl.sdu.dk/visl/da/info[3]). Disse tags disambigueres ikke selv, men indgår som kontekstoplysning i disambigueringen af syntaktiske tags, valensinstantiering etc. I forbindelse med leksikografisk korpusarbejde tillader de semantiske tags at løfte ovennævnte selektionsrestriktioner fra det rent leksiske til et mere generelt plan. Bemærk at de semantiske tags i nedenstående statistikker blev optalt isoleret, og at den manglende disambiguering på dette niveau derfor betyder at sjældne komplementer skal ignoreres idet de kunne stamme fra semantisk flertydige ord. Eksemplet "aflyse" viser at dette verbum foretrækker arrangementer, foranstaltninger og aktiviteter som direkte objekt.

21 aflyse <occ> (arrangementer)
19 aflyse <act-c> (tællelige handlinger og aktiviteter)
4 aflyse <ac> (tællelige abstrakta)
4 aflyse <act> (handlinger og aktiviteter)

[2] DanGram leverer p.t. analyser af løbende tekst med ca. 99% korrekte ordklasselæsninger, og ca. 96% korrekte syntaktiske funktioner (dog alt efter finkornetheden i det anvendte tag-sæt).

[3] Prototypeopmærkningen af substantivleksikonnet blev gennemført af Lone Hegelund i 2001 under min supervision. Kategoriinventaret tog udgangspunkt i et lignende system for portugisisk (Bick 2000) og er i store træk kompatibel med de semantisk-ontologiske kerne kategorier i ker i det europæiske SIMPLE projekt.

4 aflyse <sem-l> (musikstykker m.m.)
3 aflyse <event> (hændelser)
3 aflyse <sit> (situationer)

Med endnu et filterprogram opnås en egentlig ordbogsformatering. Bemærk at selektionsrestriktionen "mennesker" (Hprof, H, HH) ved "forhindre" skyldes den syntaktisk korrekte konstruktion "forhindre ngn i at ...".

forflytte <Hprof>_2 (professionelle mennesker)
forfægte <pp>_3 (tankeprodukt)
forfølge <ac>_8 <Hprof>_6 <H>_4 .... (aktiviteter og mennesker)
forføre <H>_3 (mennesker)
forgylde <H>_4 <Hprof>_3 (professionelle mennesker)
forhale <act-c>_3 <act>_3 (handlinger og aktiviteter)
forhandle <ac>_17 <sem-r>_9 <conv>_8 .... (tællelige abstrakta, "readables", aftaler)
forhaste <pp>_3 <sem>_3 (tankeprodukter)
forhindre <act>_35 <Hprof>_23 <ac>_18 <act>_18 <H>_17 <HH>_14 <event-c>_9
forhøje <ac>_13 <mon>_7 <mon-c>_5 ... (abstrakta og pengebeløb)
forkaste <pp>_5 <Hprof>_4 <ac>_3 <conv>_3 .. (tankeprodukter, professionelle, aftaler)
forklare <ac>_39 <act-c>_7 <act>_6 ... (abstrakta og handlinger)
forkorte <per>_4 (perioder)

Med en mindre ændring i filteret opnås en tilsvarende liste for subjekt-selektionsrestriktioner:

advare <Hprof>_44 <HH>_10 <ac>_6 <inst>_6 ... (professionelle, grupper, institutioner)
afblæse <HH>_3 <Hprof>_2 ... (grupper og professionelle)
afbryde <Hprof>_28 <HH>_10 <H>_8 <ac>_6 <Hfam>_4 ... (professionelle og almindelige mennesker)
afdække <act-c>_7 <sem>_6 <Hprof>_5 <ac>_4 (handlinger, intellektuelle frembringelser, professionelle)
affyre <H>_8 <Vair>_7 <inst>_7 <HH>_5 ... (mennesker, fly og grupper)
affærdige <Hprof>_3
afføde <ac>_12 <act-c>_10 <act>_8 ... (abstrakta, handlinger og aktiviteter)
afgive <Hprof>_34 <HH>_24 <inst>_17 ... (professionelle, grupper og institutioner)
afgøre <ac>_25 <HH>_14 <act-c>_11 <H>_6 ... (abstrakta, grupper, handlinger)

Tilsvarende kan en ekstraktion af subjekt - subjektsprædikativ-sekvenser eller prænominale attributter levere kollokationel information for substantiver og adjektiver. Således er en PC med faldende sandsynlighed bærbar (28), ny (14), stationær (9), kraftig (6), billig (5) eller fabriksny (4), mens man omvendt kan sige at hvad der kan være akut, er sandsynligvis et behov (49), et problem (47), en mangel (14) eller prototypen sygdom <sick>: skade (20), delirium (16), psykose (12), sygdom (9), leukæmi (7), hepatitis (6), smerte (5), rygsmerter (4) eller terapi: indlæggelse (12), behandling (11), hjælp (11), operation (8). En af kollokationerne stammer fra metaforisk transfer fra <sick>- til <act>-prototypen: *en akut indlæggelse* (12).

Det er ikke altid kernekollokationerne der er hyppigst. For adjektivet *ambitiøs,* for eksempel, overhales <H>-prototypen (+HUM, *politiker 6, menneske 4, kvinde 3, mor 3*)

således af den metaforiske brug i forbindelse med <pp>-prototypen (kognitive frembringelser, *plan 59, projekt 50, mål 42, målsætning 15).*

Man kan i øvrigt skelne mellem "grammatiske kollokationer" som ovennævnte, fra et opmærket korpus, og rent statistiske kollokationsfrekvenser for naboord, der også kan opnås med et ikke-opmærket korpus, især når kollokationsfrekvensen sættes i forhold til (læs: divideres med) hyppigheden af de involverede ord isoleret set i hele korpuset.

4. Teksttypologi: Passivkonstruktioner

Passivkonstruktioner bruges i vid udstrækning uden passivagent, til at "anonymisere" ytringer ved at gøre dem subjekt/agent-løs, svarende til konstruktioner med "man" som subjekt og "én" som objekt eller "éns" som possessiv. I dansk menes en høj passivfrekvens at være et stilmærke for abstrakte, videnskabelige tekster og det såkaldte kancellisprog, og man kunne forestille sig at type-klassificere opmærkede tekster ud fra deres "passivprocent"[4]. Bruger man det genremæssigt blandede Korpus2000 som standard, er en normal passivprocent 3.1% for alle former, 2.3% for finitte former (inkl. aktive participier) og 5.9% for infinitiver. Det er imidlertid ikke ligegyldigt hvilke ord der bøjes i passiv, og om det er s-passiv eller blive-passiv der bruges, og man burde ved korte tekster med få passiv-tokens (og tilsvarende statistisk usikkerhed) sætte selve passivprocenten i relation til ordenes individuelle passivtendenser:

(a) Børnene flokkedes omkring ismaskinen. *Børnene blev flokket.
(b) Løgene svitses. Løgene bliver svitset.
(c) Aktieudbytte beskattes med 25%. Aktieudbytte bliver beskattet med 25%.
(d) Minimælk fås kun fra Arla. *Minimælk bliver fået.
(e) Der arbejdes på en løsning. Der bliver arbejdet. *Den bliver arbejdet.
(f1) Bøgerne er solgt d. 10. oktober (=er blevet). *Bøgerne er solgte d. 10. oktober.
(f2) Tallene er vist (=vises) med rød skrift. *Tallene er viste med rød skrift.

Således er (a) et eksempel på et verbum der stort set kun forekommer som s-passiv, og ikke *kan* sættes i blive-passiv. Mange ord af denne type vil ligefrem være leksikaliseret i ordbøgerne med passiv-formen som grundform *(synes, slås),* og betydet således lidet for teksttypologien. "Svitse" (b) er typisk for madopskrifter, der som bekendt er rige på imperativer og s-passiver. Det er således interessant at ord som *svitse, purere, aftørre, udbløde, rengøre* ikke kun har en høj s-passivprocent, men også en høj s/blive-procent. Ord som *dømme, formene, føde* derimod, der ikke kun har en høj s-passivprocent, men en endnu

[4] For Korpus90/2000 vil dette kunne gøres, så snart forfatter- , kilde- og teksttypekoderne er tilgængeliggjort.

højere blive-passivprocent (brøk under 50 i tabellen), er mere typisk for en rapporterende teksttype.

De bedste markører for kancellistil er imidlertid måske de ord, der ikke har en leksisk alt for høj s-passivprocent, men en s/blive-procent på over 50, noget der indikerer s-passivisering af intransitive verber (e) eller transitive verber med ikke-agentivt subjekt (d). Bemærk at (e) nok tillader blive-passiv, men kun med formelt subjekt, ikke nominelt subjekt, en distinktion, der ligeledes kan kvantificeres i opmærkede korpora. Eksempler på (d-e) fra nedenstående data-liste er *fås, ønskes, menes, forventes, ventes, anses* (med stigende passivtendens).

Bemærk i øvrigt at dansk kan benytte ubøjelige passiv-participier også efter 'være', og at dette ikke altid er ækvivalent til en konstruktion med 'blive' (f1-2).

I nedenstående tabel indgår kun finitte verber og participier, ikke infinitiver. Bemærk at listen i princippet selvfølgelig dækker hele verballeksikonet, som dog ikke kan vises her. For de mest passiviske verber er forekomster under 10 udelukket, for de moderat passiviske verber er forekomster under 1000 udelukket, og for de mindre passiviske verber gives der kun eksempler med over 10.000 forekomster.

Verbum (n)	Spas/akt	Spas/Bpas
flokke 93	96	98
forefinde 31	96	98
besværliggøre 23	92	96
synliggøre 25	91	95
afvaske 20	90	95
fastgøre 10	88	93
færdiggøre 17	88	93
klargøre 13	88	93
væmme 25	88	97
ælde 33	87	66
afbøje 14	84	75
mistænkeliggøre 16	84	91
rengøre 43	84	80
omgå 132	83	96
tydeliggøre 21	82	95
aftørre 18	81	90
udbløde 11	81	90
beskatte 222	79	56
pristalsregulere 10	78	75
umuliggøre 32	77	87
formene 35	76	**22**
svitse 34	76	81
aflønne 59	75	61
desinficere 36	75	74
henregne 16	75	89
afkøle 82	74	63
aftrappe 20	74	85
purere 14	73	78
....		
udsætte 2632	53	**20**
anvende 2019	50	68
føde 2501	43	**12**
afgøre 1422	39	43
anse 1516	34	73
fremstille 1204	30	40
behandle 2349	29	34
vente 7109	29	**96**
forvente 3974	28	**91**
offentliggøre 1542	27	30
omtale 1132	26	45
...		
dømme 1426	18	**9** (PA)
etablere 1376	18	35
fjerne 2066	18	38
forhandle 1095	18	**55** (IA)
...		
bruge 12892	17	49
sætte 17565	8	22
lægge 13834	7	28
holde 13845	6	50
spille 11082	6	51
give 26077	3	50
tage 29355	3	32
gøre 35462	2	34
mene 24368	2	**73**
skrive 12730	2	15
vise 18573	2	43
fortælle 13126	1	53
ønske 11297	1	**62**
begynde 13383	0	0
blive 62820	0	50
finde 16358	0	0
få 62192	0	**96**
gå 58406	0	13
have 149945	0	0
komme 58846	0	6
ligge 18166	0	0
se 32252	0	0
sidde 11519	0	2
ske 14729	0	5
stå 28094	0	0
synes 13146	0	50
tro 13604	0	2
vide 24255	0	50
være 703308	0	0

5. Foranstillede adverbier i præpositionsstyrede infinitiver og i verbalgruppen

En dansk infinitivmarkør er en næsten sikker sætningsgrænse (for infinitte sætninger). I modsætning til tysk, fx, kan infinitivsætningers objekter ikke stå til venstre for 'at'. Imidlertid findes der konstruktioner med adverbier i denne position. En distributionel statistik over hvilke adverbier der tillades vil dels være af almen typologisk interesse, dels hjælpe en automatisk parser, både med at disambiguere adverbiet og 'at' mht. ordklasse, og med at slå adverbialets dependens fast som højrevendt (mod infinitiven) eller venstrevendt (mod verbalet i en evt. hovedsætning). For at sikre netop dependensen som vendt mod infinitiven, undersøgte jeg først sekvensen PRP ADV+ INFM @ICL-P<, altså tilfælde for infinitiven er præpositionsstyret, og adverbiet dermed "præpositionsisoleret" fra hovedsætningen. Af i alt 6.485 tilfælde havde følgende adverbier en præ-infinitiv-frekvens over 10 i Korpus2000:

ikke	2039	fortsat	84	i dag	26	ligefrem	13
selv	789	så	79	frivilligt	26	ikke blot	13
også	468	hurtigt	58	for alvor	24	frit	13
slet ikke	216	for eksempel	50	pludselig	22	eksempelvis	13
blot	211	i stedet	48	måske	22	effektivt	12
derefter	169	til sidst	47	dermed	22	reelt	11
tidlig	168	samtidig	47	stadig	20	bevidst	11
først	150	aldrig	46	alligevel	20	officielt	10
overhovedet	130	dog	43	aktivt	19	med det samme	10
fx.	117	straks	41	ulovligt	14	dels	10
både*	103	atter	39	således	14	automatisk	10
bl.a.	99	yderligere	34	i går	14		
altid	93	virkeligt	32	fremover	14		

Den mest almindelig adverbiumsklasse på positionen synes at være **fokusadverbier**, der også forekommer prænominelt *(ikke, selv, også, blot, fx., bl.a.),* med tillæg af de mere konjunktionelle *både* og *dels.* Man kan dog diskutere om *selv* her har den samme betydning som prænominalt i "selv Peter måtte indse ...". Den næste store gruppe er **tidsadverbier** *(derefter, tidlig, først, altid, fortsat, hurtig, til sidst, samtidig, aldrig, straks, atter ...).* **Bøjede adverbier** er mere sjældne *(hurtigt, virkeligt, frivilligt, aktivt),* og præpositionssyntagmer eller substantivsyntagmer er gerne fasttømrede udtryk fra især tidsdomænet *(år efter år, ad åre, i går, i givet fald, i hvert fald).* Kombinationer af flere adverbier eller adverbium plus pp forekommer, men sjældent og som regel i et gensidigt dependensforhold *(lige præcis, ikke i tide, først og fremmest).* Sjældne eksempler på flere end 2 adverbier er:

for så først derefter at ...
ved ikke blot passivt at ...
for derefter alligevel straks at ...

At *overhovedet* og til en vis grad *dog,* kan indgå i konstruktionen, er i øvrigt et argument for at infinitvsætninger i nogen henseender kan ligestilles finitte ledsætninger, idet hverken hovedsætninger eller gruppesyntagmer tillader disse ord som konstituenter.

6. Pronominal-ellipse i relativsætninger

Ud over den almindelige ledsætningsordfølge er syntaksen i danske relativsætninger topologisk karakteristisk ved at frontstillingen af relativpronominerne resulterer i OSV- og ASV-ordstilling, samt at relativpronomen (i modsætning til fx tysk og de romanske sprog) i nogle tilfælde helt kan udelades. Jeg har benyttet 938 løbende relativsætninger (svarende til ca. 70.000 ord) fra det opmærkede Korpus2000 til at undersøge distributionen af de enkelte danske relativpronominer, herunder ellipseprocenten, i forhold til forskellige syntaktiske relationer. Leksisk er det interessant at 'der' udfylder halvdelen af alle relativpladser (49%), og at ordets brug er begrænset til subjektpladsen, mens 'som' (27%) dækker - foruden subjektet - alle andre "nominale" syntaktiske funktioner, og 'hvor' (12%) de adverbielle funktioner:

Syntakt. funktion	der		som		zero (udeladt)		samlet	
	antal	%	antal	%	antal	%	antal	%
SUBJ	421	**44,9**	175	**18,7**	15	**1,6**	611	**65,1**
raised	-	**-**	3	**0,3**	-	**-**	3	**0,3**
det-fokus	33	**3,5**	10	**1,1**	-	**-**	43	**4,6**
ACC	-	**-**	34	**3,6**	37	**3,9**	71	**7,6**
raised	-	**-**	7	**0,7**	2	**0,2**	9	**1,0**
det-fokus	-	**-**	-	**-**	6	**0,6**	6	**0,6**
>>P	4	**0,4**	16	**1,7**	12	**1,3**	32	**3,4**
raised	-	**-**	7	**0,7**	1	**0,1**	8	**0,9**
det-fokus	-	**-**	-	**-**	5	**0,5**	5	**0,5**
DAT	-	**-**	1	**0,1**	-	**-**	1	**0,1**
CS	-	**-**	2	**0,2**	-	**-**	2	**0,2**
CO	-	**-**	2	**0,2**	-	**-**	2	**0,2**
	458	**48,8**	257	**27,4**	78	**8,3**	793	**84,5**
	hvor		når		zero (udeladt)			
ADVL-adv	111	**11,8**	10	**1,1**	10	**1,1**	131	**14,0**
	PRP + hvor		PRP + hvilken		88	**9,4**	924	**98,5**
P< (ADVL)	7	**0,7**	1	**0,1**			8	**0,9**
	hvis		at		hvilket			
>N (SUBJ)	1	**0,1**					1	**0,1**
SUB			4	**0,1**			4	**0,4**
S<					1	**0,1**	1	**0,1**
							938	**100,0**

Tabellen viser, at 'der' er generelt mere almindelig som relativsubjekt end 'som', og at bias'en er endnu større i forbindelse med fokuskonstruktioner ("Det er Anne der har lavet

aftensmad"). Mens subjekter i almindelighed kun er dobbelt så hyppige som direkte objekter (jf. kap. 2), er relationen i relativpositionen snarere 8:1. Dette skal imidlertid ses i lyset af at relativpronominer udgør *venstrestillede* objekter, og som sådanne *er* frekvente (Bick 2002), idet "som" udgør 2/3 af alle venstrestillede akkusativobjekter (hvis man ser bort fra citerede sætninger) og er hele 6 gange mere almindelig end venstrestillede *substantiviske* akkusativobjekter.

Pronominalellipsen, undersøgelsens primære sigte, viser sig uden videre at kunne konkurrere med det eksplicitte 'som' for både direkte objekter (@ACC> - 3,9 mod 3,6%) og frontstillede styrelser (@>>P - 1,3 mod 1,7%). Kun ved *raising* foretrækkes det eksplicitte 'som' ("tegningen viser en løsning ***som*** jeg ikke tror *han selv har fundet* ***på***"). Pronominalellipse bruges ikke for relativsubjekter (tilfældene stammer fra sideordnede sætninger), og er ualmindelig for adverbialer ("det år [hvor] han blev født", "den konference / det sted [hvor] vi skal mødes). I nogle tilfælde forekommer valget mellem hhv. brug eller ellipse af 'som' og 'hvor' at være grammatikaliseret til fordel for den ene eller den anden løsning, og det store antal relativsætninger i Korpus2000 (ca. 400.000) vil da også gøre en fremtidig, mere kvalitativ undersøgelse attraktiv.

Bibliografi:

Asmussen, Jørg (DSL). 2002. "Korpus 2000, et oberblik over projektets baggrund, fremgangsmåde og perspektiver". I: *Nys30 - Korpuslingvistik.*København: Akademisk Forlag

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

Bick, Eckhard. 2001. "En Constraint Grammar Parser for Dansk". I: Widell, Peter & Kunøe, Mette (udg.): *8. Møde om Udforskningen af Dansk Sprog.* Århus: Århus Universitet.

Bick, Eckhard. 2003. "PaNoLa, The Danish Connection". I: *Årbog 2002 for Nordisk Sprogteknologisk Forskningsprogram 2000-2004.* Forthcoming.

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

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