

CORPUS EYE – ET BRUGERVENLIGT WEB-INTERFACE FOR GRAMMATISK OPMÆRKNING AF KORPORA

Af Eckhard Bick (Syddansk Universitet)

1. Indledning

I de senere år har teknologiske og datalingvistiske fremskridt gjort det muligt at kompilere og annotere stadigt større tekstsamlinger, til gavn for empirisk, korpus-baseret sprogforskning. For flere af de store sprog foreligger der i dag korpora med over 100 millioner ord, og internettet kan i princippet betragtes som ét stort multilingvalt korpus. For dansk er de største offentligt tilgængelige tekstkorpora Korpus90 og Korpus2000 (kompileret af Dansk Selskab for Sprog og Litteratur), med hver 26 millioner ord, mens der på talesprogsområdet findes fx BySoc (Henriksen 1998) og transskriberede parlamentsdiskussioner fra Folketinget og Europaparlamentet (Europarl). Imidlertid er brugsværdien af et korpus ikke kun afhængig af designparametre som størrelse, genre, tidsperiode m.m., men også af eksistensen og kvaliteten af tilføjet grammatisk meta-information, samt tilgængeligheden og acceptansen blandt lingvistiske forskere, lærere, leksikografer og andre. Jeg har tidligere præsenteret et automatisk korpus-opmærkningssystem for dansk (DanGram, MUDS8), samt et projekt til manuel lingvistisk revision af det opmærkede Korpus90/2000 (MUDS9), og vil denne gang fokusere på det sidste aspekt – vejen fra korpus til bruger.

2. Et integreret, internetbaseret søgeinterface

Her er det afgørende, om der foreligger brugervenlige redskaber til korpussøgning – dvs. redskaber, der (a) ikke kræver køb og installering af specialiseret software, og (b) ikke forudsætter, at brugeren tilegner sig et korpusspecifikt kodesprog. En elegant løsning på (a) er internetbaseret korpusadgang, idet kompatibiliteten med brugerens computersystem her sikres igennem browseren – og hjemmesiderne af både BySoc og Korpus2000 er gode eksempler herpå. Begge systemer har dog visse begrænsninger. For det første er de tilpasset til eet bestemt korpus og tillader ikke kombination og sammenligning med andre korpora i samme

søgning og samme interface. For det andet har brugeren ingen mulighed for at anvende morfologiske eller syntaktiske kategorier i sin søgning.

CorpusEye-projektet (<http://corp.hum.sdu.dk>) på Syddansk Universitet er et forsøg på at designe og programmere et internetbaseret søgeinterface, der dels tilbyder ens redskaber og samme formalisme på tværs af flere korpus-typer og på tværs af flere sprog, dels tillader at udnytte den grammatiske information i opmærkede korpora på en brugervenlig og menubaseret måde.

3. Et stort og voksende korpusudvalg for flere sprog

Samtlige korpora i CorpusEye er blevet forsynet med morfologiske og syntaktiske tags vha. VISL's Constraint Grammar-baserede parsere, for træbankernes vedkommende med et efterfølgende PSG-modul (Bick 2003-1), der i stedet for ordformer benytter syntaktiske funktioner som terminaler i sine genskrivningsregler. På nuværende tidspunkt er der tale om følgende materiale:

- **Dansk:** 5 korpora, ca. 50 millioner ord (bl.a. Korpus90/2000, Skalk, Europarl¹, Folketingsdebatter og Arboretum-træbanken²)
- **Portugisisk:** 4 korpora, ca. 250 millioner ord (bl.a. Público, Folha de São Paulo³ og Floresta-Sintá(c)tica-træbanken)
- **Engelsk:** 4 korpora, ca. 120 millioner ord (herunder BNC, KEMPE⁴ og Europarl)
- **Tysk:** 4 korpora, ca. 50 millioner ord (herunder MAK, BZK⁵ og Europarl)
- **Fransk:** 3 korpora, ca. 35 millioner ord (herunder Le Monde, Europarl og Arboratoire⁶)
- **Spansk:** 3 korpora, ca. 30 millioner ord (herunder El Diario Sur og Europarl)
- **Esperanto:** 5 korpora, ca. 17 millioner ord (bl.a. Monato, Eventoj⁷, klassisk litteratur)

¹ *Europarl* er et stort, frit parallelkorpus med debatudskrifter fra Europaparlamentet, der dækker i alt 11 sprog med 20-30 millioner ord hver, fra perioden 1996-2003. Korpuset er oprindeligt kompileret af Philip Koehn.

² Arboretum er stadig under opbygning, og indeholder nu reviderede analyser af ca. 15.000 sætninger (ca. 300.000 ord) i både Constraint Grammar- og træbank-format.

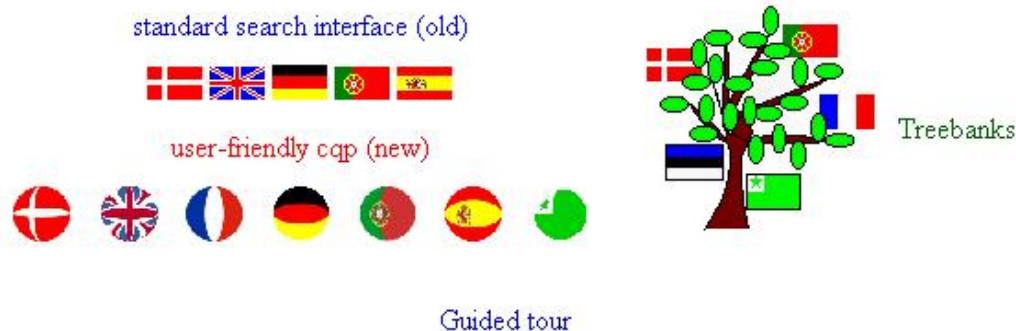
³ *Público* og *Folha de São Paulo* er store dagblade i hhv. Portugal og Brasilien. Korpusserne er kompileret af Linguatca-projektet (v/ Diana Santos), og opmærket med forfatterens PALAVRAS-parsere. Et uddrag fra begge tekstsamlinger underkastes løbende lingvistisk revision på træbank-niveau (*Floresta Sintá(c)tica*).

⁴ *Kempe*, 'Korpus of Early Modern Playtexts in English', er kompileret af Lene B. Petersen and Marcus X. Dahl og opmærket i samarbejde med VISL. *BNC* (British National Corpus) indeholder både skrift- og talesprog.

⁵ Både *MAK* (Mannheimer Korpus) og *BZK* (Bonner Zeitungskorpus), samt det spanske *El Diario Sur* og det franske *Le Monde*-Korpus stammer fra *European Corpus Initiative*, og er siden opmærket med forfatterens parsere, der for fransk og tysk få leveret morfologisk input fra hhv. Achim Stein og Helmut Schmid's DTT-tagger og Lingsofts GERCG.

⁶ *Arboratoire*-træbanken er en del af *Freebank*-initiativet, et samarbejde mellem VISL og ATILF/Loria (Susanne Salmon Alt, Nancy). Teksterne opmærkes med forfatterens FrAG-parsere, efterfulgt af uddragsvis revision.

- **Estisk:** 1 korpus, 3.500 ord (Arboresc-træbanken⁸)



4. Søgformalisme og datastruktur

Den interne søge-database benytter IMS' *Corpus Query Protocol* (Christ 1994, <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>) for CG-korpora og linux-værktøjet *tgrep2* (<http://tedlab.mit.edu/~dr/Tgrep2/>) for træbankerne. Internet-brugen af CQP er inspireret af andre søge-interfaces der tidligere er programmeret af hhv. Paul Meurer for norsk (Oslo Universitet) og Diana Santos for portugisisk (Linguateca). Selvom CorpusEye tillader direkte brug af CQP⁹ samt såkaldte *regular expressions*¹⁰ (joker-søgninger, sets og matematiske operatorer), henvender projektet sig primært til den humanistiske bruger uden formalistiske forkundskaber. Det er derfor muligt at komme i gang med simple tekstsøgninger, der præsenteres i konkordans-format, som her for ordet "høj":



Searched for: [word="høj"]
In corpus: DAN_C2000
Found 2782 results (2782).
1 - 24 [next](#)

```
INFO          heraf , som udbudsmaterialet , er af uhørt høj kvalitet , konklud
INFO lige ved alfarvej eller steder , hvor huslejen er høj .
INFO          Jeg kan blive høj af en god forretning
INFO yderligere en film med Steven Seagal , Kapring i høj fart , som Geoff M
INFO har svært ved at holde sig på ligeud-kursen ved høj fart og i sidevind
INFO Indtrækket er af tilsyneladende høj kvalitet og med ve
INFO mod øst , så mange amerikanske eksperter det som høj gambling .
INFO bedre tale om gradbøjninger af en gennemgående høj moral , skriver Jø
INFO mennesker , men kun for folk med firmabil eller høj indtægt .
```

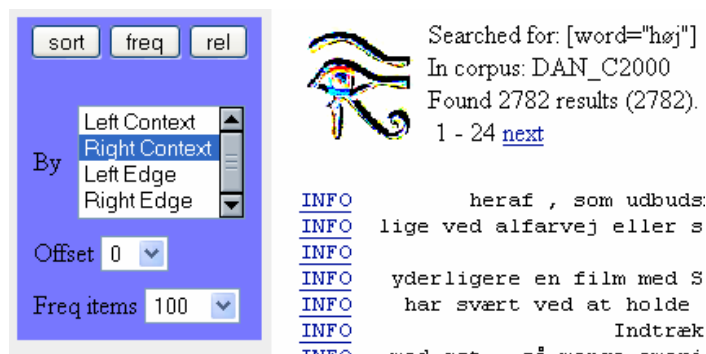
⁷ *Monato* er et internationalt nyhedsmagasin, *Eventoj* er et internetbaseret nyhedsbrev.

⁸ *Arboresc* er blevet til i forbindelse med det NorFa-støttede Nordiske Træbank-Netværk, og bygger på CG-analysen af estisk avismateriale, der underkastes en efterfølgende PSG-analyse (revideret af Heli Uibo).

⁹ Her formuleres søgekriterier for hvert token (ord) for sig, fx. [morph="PR AKT"] [pos="N" & func="<SUBJ"] for to på hinanden følgende ord, det første (et verbum) i præsens aktiv, det andet et nomen (substantiv) med subjekt-funktion og venstreventet dependens (<), som fx i *siger talskvinden*, ... *Her har hjorten ligget*.

¹⁰ *Regular expressions* tillades både i rene tekst-søgestrengene, og i CQP-udtryk. De vigtigste operatorer er '?' (ingen eller en), '*' (ingen eller flere) og '+' (en eller flere), og kan tillægges bogstaver (fx 'korpus+er' = korpuser, korpuser), sæt (fx 'k[æøå]be[rn]?e?' = kæber, køben, kåbe, ...) og joker-tegn (fx * for en vilkårlig streng).

Med et enkelt klik kan brugeren producere statistiske, absolutte eller frekvensnormerede, oversigter over bestemte positioner i søgningen eller dens kontekst. Sorteringspositionen udregnes fra enten "kantord" (*edges*, første og sidste ord i det fundne og markerede sætningsudsnit) eller "kontekstord" (sidste ord *før* og første ord *efter* det markerede udsnit).



The screenshot shows a search interface with a blue sidebar on the left and a main content area on the right. The sidebar contains buttons for 'sort', 'freq', and 'rel'. Below these is a 'By' section with a dropdown menu showing 'Left Context', 'Right Context', 'Left Edge', and 'Right Edge'. There is also an 'Offset' field set to 0 and a 'Freq items' field set to 100. The main content area displays search results for the word 'høj'. It includes a search bar with 'Searched for: [word="høj"]', the corpus 'In corpus: DAN_C2000', and the number of results 'Found 2782 results (2782)'. A link '1 - 24 next' is visible. Below this, there are several lines of text, each preceded by an 'INFO' tag, showing examples of the word 'høj' in context.

5. Frekvensundersøgelser

Allerede de første eksempler giver en "leksikografisk" idé om hvilke ord der på dansk kan kvalificeres med adjektivet "høj" – *kvalitet, fart, moral, indtægt*, og en frekvensanalyse med udgangspunkt i højre kontekst vil give et udtømmende billede af de involverede selektionsrestriktioner. I eksemplet er der søgt på "højt?", for også at fange neutrumsformer. Imidlertid vil dette søgemønster også tillade adverbiet "højt" (*højt besungne, højt oppe, højt på, højt til*). Listens nytteværdi øges betydeligt, hvis der i stedet for absolutte frekvenser (*freq*, midterste eksempel) beregnes relative tal, hvor den lokale frekvens (i konkordansen) normaliseres vha. at dividere med kontekstordets normalfrekvens fra hele korpuset (*rel*, højre eksempel). Endnu bedre bliver resultatet når der – igennem "refine search" søges på adjektiv-lemmaet "høj" efterfulgt af et substantiv (højre eksempel). Fordi dette producerer 2-ords-konkordanser, skal der her sorteres på "right edge" snarere end "right context":

DAN_C2000 (6017)			DAN_C2000 (6017)			DAN_C2000 (7601)		
frequencies:	rel	freq num	frequencies:	rel	freq num	frequencies:	rel	freq num
grad	38551	7.2 [436]	grad	38551	7.2 [436]	grad	223781	17.4 [1327]
,	70	6.7 [406]	besugne	14244	0.3 [19]	lærestalter	34893	1.1 [84]
.	45	5.1 [311]	"	13534	0.1 [7]	kvalitet	6017	1.9 [148]
,	33	4.6 [279]	kvalitet	9603	2.4 [148]	enhed	5575	0.9 [74]
på	89	3.2 [194]	prioritet	8842	0.7 [47]	prioritet	5541	0.6 [47]
.	17	3.1 [192]	niveau	7935	2.3 [141]	handseksamen	4984	0.1 [12]
og	133	5.6 [340]	stressniveau	6905	0 [5]	niveau	4972	1.8 [141]
kvalitet	9603	2.4 [148]	beskæftigelsesniveau	6905	0 [5]	stressniveau	4327	0 [5]
niveau	7935	2.3 [141]	cigarføring	6629	0.1 [12]	beskæftigelsesniveau	4327	0 [5]
oppe	3236	1.6 [97]	kolesteroltal	5892	0.1 [8]	cigarføring	4154	0.1 [12]
op	165	1.4 [90]	HURRA	4971	0 [6]	kolesteroltal	3692	0.1 [8]
til	11	1.2 [74]	aktivitetsniveau	4710	0.2 [18]	priser	3282	1.2 [93]
i	7	1.2 [74]	prioritet	3551	0.4 [30]	prioritet	3250	0.4 [36]
pris	1151	0.9 [57]	oppe	3236	1.6 [97]	luftlag	3250	0.1 [13]
som	8	0.8 [50]	skattet	2946	0.1 [8]	HURRA	3115	0 [6]
for	6	0.8 [49]	respekterede	2694	0.3 [19]	hæle	2951	0.2 [18]
prioritet	8842	0.7 [47]	tempo	2619	0.6 [42]	aktivitetsniveau	2951	0.2 [18]

Undersøgelser som denne kan bruges på en kontrastiv måde, fx i fremmedsprogsundervisning (her: dansk for indvandrere), eller til at øge den sproglige bevidsthed i modersmålsundervisning (emne: semantiske kategorier, selektionsrestriktioner). Hedder opgaven således at undersøge brugs mønstret for hhv. *høj* – *stor* ('high' – 'big?') og *stærk* – *kraftig* ('strong' – ?), vil *CorpusEye* kunne sandsynliggøre, at *høj* bruges for målinger på en skala af diskrete enheder, mens *stor* bruges mere beskrivende end målende, med en selektionsrestriktion for ting og ikke-tællelige abstrakta. Absolutte frekvenser fra Korpus2000 er angivet i parentes:

høj/t ... grad (434), kvalitet (148), niveau (141), pris (57), prioritet (47), tempo (42), fart (40), indhold (36), humør (35), alder (34), kurs (33), hastighed (31), klasse (28), arbejdsløshed (24)

stor/t ... del (1214), betydning (432), succes (297), forskel (267), antal (245), interesse (205), vægt (201), problem (196), flertal (156), indflydelse (154), gruppe (142), rolle (141), glæde (141)

I det andet eksempelpar synes animate substantiver som udgangspunkt at foretrække *stærk*, mens handlinger og hændelser som prototype associeres med ordet *kraftig*. Bemærk den metaforiske "human"-kategorisering af *position*, *ønske* og *økonomi*.

stærk/t ... hold (28), pres (22), mand (21), vilje (15), position (15), ønske (15), vækst (15), kontrast (14), modstander (14), leder (14), økonomi (13)

kraftig/t ... stigning (31), vækst (30), kritik (22), jordskælv (13), afstand (9), advarsel (9), pres (8), opfordring (8), forbedring (8), fald (8), slag (7), mistanke (7)

Ordet *vækst*, der forekommer i begge lister, viser betydningen af brugen af *relative* frekvenser. I en korpus søgning på "ADJ + *vækst*" vil en relativ frekvenssortering således placere *kraftig* på plads 2, mens *stærk* forvises til plads 11.

Køres der, ud fra fx et diakront eller genre-perspektiv, søgninger på mere end et korpus ad gangen, understøtter frekvensanalysen også en direkte sammenligning (i samme

vindue) af hyppigheden på tværs af kilderne. Med sit nuværende korpusinventar tillader *CorpusEye* således allerede at sammenligne skrift- og talesprogsbrug i det engelske BNC eller syntaksen i europæisk og brasiliansk portugisisk. For dansk arbejdes der på en varieret genre-repræsentation, og der kan drives diakrone studier ved at sammenligne Korpus90 og Korpus2000, fx hvad angår det danske fremmedordsleksikon, ortografisk variation eller betydningsdrift i politisk aktive ord som "etnisk" eller "tosproget".

6. "Refine search": Menu-baserede kategorisøgninger

Ønsker man at søge på fx leksemer (bøjningsformsneutraliseret), ordklasse (fx nominativ-substantiver efter "hendes") eller syntaktiske funktioner (fx frontstillede objekter), kan dette gøres igennem kategorimenuer¹¹, der knytter sig til den enkelte søgeposition (det enkelte ord), der så – usynligt for brugeren – oversættes til CQP-udtryk og CG-tags. Menuerne fremkommer som et selvstændigt java-vindue, der har sin egen søgningshukommelse, og desuden "arver" allerede indtastede fuldformssøgninger fra det simple, tekstbaserede indgangsinterface. Overgangen til kategorisøgningsvinduet foregår ved at klikke på "Refine search" knappen, fx når en konkordans indeholder for mange falsk positive hits efter en ren tekstsøgning.

¹¹ Hvilke kategorier der medtages i menuerne, indstilles igennem sprogspecifikke konfigurationsfiler. Menupunktet *subjekt* vil altså kunne indeholde *forskellige* subjektclasser for hhv. dansk og spansk, og enkelte menupunkter, fx *diatese*, vil enten slet ikke, eller ikke på ordniveau, give mening for nogle af sprogene.

Ud over menuerne PoS (ordklasse), morfologi og (syntaktisk) funktion kan man også søge på leksem/lemma ("base") snarere end ordformer ("word"), og det er muligt at negere dele af søgningen ("Neg"). "Extra"-feltet benyttes til mere korpus-specifikke, sekundære tags (fx. ca. 200 semantiske kategorier for danske substantiver). Menupunkter med "more" dækker over tag-sæt og kan udfoldes, med mulighed for differentierede valg, fx. – for dansk – 'subjekt', 'foreløbigt subjekt' og 'situativt subjekt' under punktet *subjekt*. I morfologi-menuen (hvis konkrete udformning i øvrig afhænger af den valgte ordklasse) skal underpunkter vælges for at specificere kategorien, idet et valg som "numerus" eller "genus" som sådan ikke giver mening. "Finiteth" skal således, hvis brugt, specificeres som en eller flere af følgende: *finit verbum*, *infinitiv*, *præsens participium*, *perfektum participium*. Også antallet af søgefelter (ord) er variabel, og søgepositioner kan tilføjes eller fjernes ved at bruge felternes + og ÷ knapper.

Eksemplet, en simpel søgning på OVS-ordstilling, specificerer for 1. felt et substantiv eller proprium med objektfunktion, for 2. felt et finit verbum og for 3. felt et (formmæssigt) uspecificeret subjekt, og resulterer i konkordanser som følgende (fra Arboretum):

```

INFO Samme synspunkt indtager afdelingschef Karin Frøidt
INFO Og deres udstillinger organiserede de i stuerne på gård
INFO gældposter , som bliver solgt fra , desto større mangvremuligheder får ØK-ledelsen til at skri
INFO Øverste terrasse buskryddede jeg også , da den tildel
INFO Klokken havde han opgivet .
INFO Og inden_for den type sex , hvilken afart ville hun gerne have ?
INFO I et forsøg på at kontrollere Thorsens forklaring undersøger politiet også , hvor Tr
INFO Den oplevelse har uddeler Søren Mortensen også .
INFO I stedet_for at blive ved med at diskutere formålet vil jeg opfordre til , at alle forer

```

Bemærk, at sidste eksempel i udsnittet *ikke* er OVS, men skyldes en frontstillet ledsætning med eget objekt ('formålet'). Opmærkningen skelner imidlertid mellem <ACC (venstrevendt direkte objekt) og ACC> (højrevendt objekt), og sidstnævnte *er* derfor en søgbar kategori, i hvert fald i en egentlig CQP-søgestreng¹². Den fulde opmærkning synliggøres som *pop op*-vindue ved at føre musemarkøren hen over et ord i konkordansen, eller – for en hel sætning ad gangen – ved at klikke på *INFO* ud for den enkelte konkordanslinje¹³.

¹² Undtagelsen er de engelske korpora, idet engelsk er det eneste af VISL's 7 korpus-sprog, der benytter sig af en ekstern syntaktisk CG-modul (Lingsofts), der underspecificerer objekters dependensretning, og kun bruger eet tag for direkte objekter, OBJ.

¹³ INFO-vinduet vil desuden – hvor det pågældende korpus indeholder denne information – oplyse om kilde, dato, taler, sætnings-id o.l.

<i>Øverste</i>	<i>terrasse</i>	<i>buskryddede</i>	<i>jeg</i>	<i>også</i>
øverst	terrasse	busk+rydde	jeg	også
		compound mv		
ADJ	N	V	PERS	ADV
nG S DEF NOM	UTR S IDF NOM	IMPF AKT	1S NOM	
@>N	@ACC>	@FMV	@<SUBJ	@<ADVL

id=.badbrhiu.preom=.0.bop=.1.eop=.0.

Det er muligt grafisk at knytte såkaldte operatorer (repetition, negation, optionalitet m.m.) til en given søgeposition:



For eksempel kunne der i nærværende søgning indføjes et nyt felt for "optionelt prænominale" (*-operator, 4. gule checkboks) mellem verbums- og subjektsfeltet, for at tage højde for subjekter, der ud over definite substantiver, pronominer og proprier også dækker over nominalsyntagmer indeholdende prænominalle artikler, adjektiver o.l. Knapperne er i høj grad selvforklarende igennem pop op-vinduer, og indgangssiden byder på en introducerende flash-film, der guider brugeren igennem systemet.

7. Søgning i træbanker



Select a Danish treebank:
 Arboretum - Korpus2000 (ca. 170.000 words)
 Arboretum - Korpus90 (ca. 20.000 words)



Compose a Tgrep2-search for:
 Search Reset

text, whole sentence
 text, whole sentence
 text, section
 tree-section
 whole tree, flat

Selvom ordbaseret grammatisk opmærkning er robust og tillader hurtige søgninger og en effektiv kompleksitetsfiltrering, vil visse strukturelt komplekse søgninger være vanskelige at formulere i det præsenterede interface, simpelthen fordi mængden af optionelt materiale kan være for stor mellem de (kerne-)ord, der bærer den relevante opmærkning. Dette gælder fx hierarkisk betingede ledsætningsfunktioner og – i mindre grad – verbalkomplementeringen, der nemmest exciperes af konstituent- eller dependensgrammatiske træbanker med deres eksplicite syntaktiske dybde. Træbankdelen af *CorpusEye* er ikke menubaseret i samme grad

som CQP-interfacet, men det tilbyder internetadgang til tgrep2-søgninger og et grafisk output med fuld integration i VISL's visualisering- og manipuleringsværktøjer for syntaktiske træer.

/np/ < (DN:pp . DN:fcl)

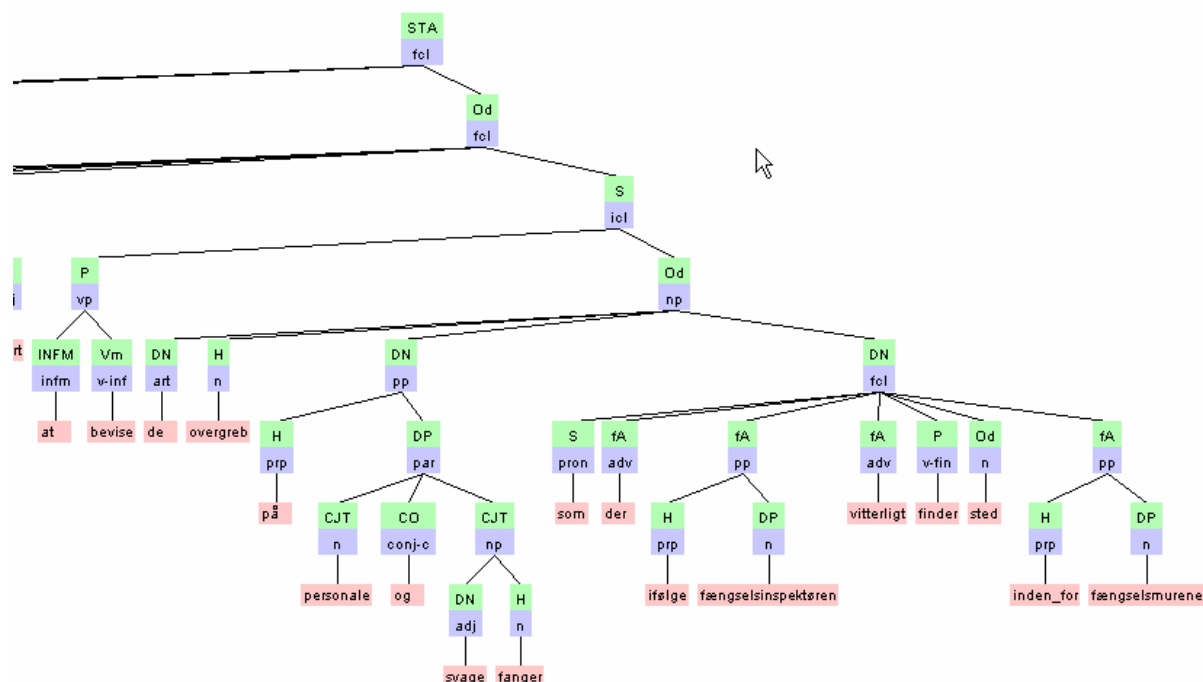
I eksemplet søges der efter nominalsyntaxer (/np/) med to postnominale (DN) søsterdependenter (.), den første et præpositionssyntaxme (pp), den anden en (finit) relativsætning (fcl). Bemærk at VISL-notationen (<http://beta.visl.sdu.dk/visl2/treebanks.html>) konsekvent skelner mellem syntaktisk form og funktion, her adskilt af et kolon (funktion før form). Pga. den større kompleksitet præsenteres resultaterne ikke som konkordans, men enten som hele sætninger (jf. eksempel), eller som det sætningsudsnit, søgningens topknode repræsenterer.

#107 ID=[aotpegcx](#) Han erkender, at det er ekstremt svært at bevise **de overgreb på personale og svage fanger, som der ifølge fængselsinspektøren vitterligt finder sted inden for fængselsmurene**. #A1/4

#189 ID=[aijangan](#) RSO lægger op til årtusindskiftet med en række hovedværker fra vort eget århundrede, som bl.a. er lagt i eksperthænder hos dirigenterne Roshdestvensky og Temirkanov. #A2/4

#241 ID=[aevjevqv](#) Der, hvor han nænsomt og hudnært lader sproget åbne sig for **den glædesmagt i tilværelsen, der giver ham og alle os andre styrke til Ø leve livet på trods af alle genvordighederne**. #A1/14

Både tekst (ovenfor) og trænotation understøttes, sidstnævnte i 3 formater: (a) PENN-træbank-format (tekst beriget med start- og slutparenteser for alle konstituenten); (b) VISL-kildeformat med "vertikaliserede" træer og indrykning af datterknuder; (c) grafisk format:



8. TextPainter: Korpora "on the fly"

Som tidligere antydte, er *CorpusEye*-søgeinterfacet en del af VISL-projektet (*Visual Interactive Syntax Learning*), der også udvikler grammatiske CALL-applikationer (*Computer Aided Language Learning*). Traditionelt fokuserer mange CALL-øvelser på et snævert emne ad

gangen, såsom ordklasser, et bøjningsproblem eller kommatering af relativsætninger, og hvis læreren ikke kan finde en færdigstøbt CALL-øvelse der passer ind i undervisningsstoffet, vil der som regel ikke være mulighed for at ændre i eller tilpasse eksisterende moduler. Problemet er særligt relevant i faget "Almen Sprogforståelse", der på relativ kort tid søger at dække en lang række emner og øge elevernes sproglige bevidsthed som sådan. Her er en af de didaktiske strategier at lade eleven selv "opdage", hvilke karakteristika, distribution og brugsregler der knytter sig til bestemte grammatisk kategori. Et brugervenligt korpusinterface kan hjælpe eleven at finde relevante eksempelsætninger og ændre i søgningerne på en fleksibel og inkrementel måde.

I forbindelse med URKAS- og VISL-SEM-projekterne har forfatteren forsøgt at integrere korpusopmærkning, "text grading" og grammatiske øvelser i et nyt redskab, *TextPainter*¹⁴, der tillader emne/kategori-specifik opmærkning af brugertekster, der løbende underkastes en automatisk grammatisk analyse. *TextPainter* accepterer således cut and paste-tekst på 7 sprog, og fremhæver ord med en ønsket grammatisk kategori eller kategorikombination, fx *subjekter, objekter, verber* eller *prædikativt brugte adjektiver*.

En overordnet øvelse kan bestå i genrebestemmelsen af en tekst: Ved fx at farve verber røde og adjektiver blå vil man kunne skelne mellem en mere handlingspræget action-fortælling og en mere deskriptiv landskabsskildring.

For at opnå en robust analyse og en lav fejlprocent, arbejdes der videst muligt med regel-baserede Constraint Grammar¹⁵-parsere, og al grammatisk information markeres på ord-niveau. Komplekse syntaktiske funktioner repræsenteres således på konstituentens kerneled, i

¹⁴ <http://beta.visl.sdu.dk/visl2/texttyping.htm>

¹⁵ http://beta.visl.sdu.dk/visl2/constraint_grammar.html

en dependensgrammatisk tradition. Ledsætningsfunktion, som fx "relativsætning" (@CL-N<), knyttes således til den pågældende sætnings første verbum:

categories: @CL-N< ... OR ... NONE

Den del af planloven , der **begrænser** nye dagligvarebutikker til højst 3.000 kvm og reelt **forhindrer** et Bilka i Horsens , skal væk og erstattes af noget mere fleksibelt og tidssvarende . ¶ Det mener både kommuner og eksperter , der **betegner** loven som alt=for restriktiv og firkantet og **peger** på , at maksimumsgrænsen gør det umuligt at gennemføre en fornuftig planlægning for detailhandelen

I interaktiv modus skal brugeren selv *finde* alle ord med en bestemt kategori, i nedenstående eksempel kategorien *direkte objekt*. Feedback gives i form af røde og grønne Grammy-bævere, og performansen evalueres med udgangspunkt i en vægtning af falsk positive og falsk negative svar, den såkaldte F-score.

check!

☐ HTS ☐ har ☐ spurgt ☐ 650 ☐ medlemsvirksomheder  ☐ om ☐ deres ☐ indstilling ☐ til
☐ medarbejderes ☐ private ☐ brug ☐ af ☐ internettet ☐ i ☐ arbejdstiden ☐ Halvdelen
☐ af ☐ de ☐ adspurgte ☐ mener , ☐ at ☐ medarbejderne  ☐ bruger ☐ stadig ☐ mere
☐ tid ☐ på ☐ nettet ☐ til ☐ det ☐ private . ☐ En ☐ fjerdedel ☐ synes ☐ allerede
☐ nu , ☐ at ☐ der ☐ bliver ☐ brugt ☐ for ☐ meget ☐ tid  ☐ på ☐ det . ☐ ¶ ☐
Halvdelen ☐ af ☐ arbejdsgiverne ☐ overvåger ☒ internetforbruget  ☐ i ☐ et=eller=andet ☐
omfang . ☐ Kun ☐ ganske ☐ få ☐ overvåger ☐ det  ☐ konstant , ☐ mens ☐ flere ☐
tager ☐ stikprøver 

categories: @ACC ... OR ... NONE

All in all, there were 66 words in the text.
For the category/categories in question, you found 4 out of 5 possible words, and had 2 false positives. This equals a recall of 80 % and a precision of 66.66 %, combining into an **F-score of 72.72 %**. To compare your own with the computer's opinion, check the highlights below!

9. Perspektiv

Selvom udviklingen på ingen måde er afsluttet, er *CorpusEye*-systemet på nuværende tidspunkt fuldt funktionelt, og der afholdes introducerende workshops. For at imødekomme den voksende interesse i skolemiljøet vil der blive udarbejdet mere målrettet dokumentation og eksempler på didaktisk brug af korpus-data i sprogundervisningen. På forskningssiden planlægges der tilbud om opmærkning og tilgængeliggørelse af brugerindsendte tekstsamlinger parallelt med de eksisterende korpora, samt lancering af joint ventures til lingvistisk revision

af benchmark-korpora for forskellige sprog og genrer, med et særligt fokus på træbank-data. I takt med videreudviklingen af VISL's NLP-værktøjer for de enkelte sprog vil nye annotationsniveauer blive tilføjet til eksisterende korpora. For dansk gælder dette p.t. semantiske prototyper for substantiver og en semantisk klassifikation af proprier (Bick 2004), samt evt. kasusroller og visse former for anaforer.

Litteratur

- Bick, Eckhard (2003-1) "A CG & PSG Hybrid Approach to Automatic Corpus Annotation". In: Kiril Simow & Petya Osenova (eds.), *Proceedings of SProLaC2003* (at Corpus Linguistics 2003, Lancaster). pp. 1-12.
- Bick, Eckhard (2003-2) "Morfosyntaktisk opmærkede korpora for dansk". I: Peter Widell & Mette Kunøe (udg.), *9. Møde om Udforskningen af Dansk Sprog*, pp. 43-54. Aarhus Universitet.
- Bick, Eckhard (2004) "A Named Entity Recognizer for Danish." In: Lino et al. (eds.), *Proc. 4th International Conf. on Language Resources and Evaluation, LREC2000* (Lisbon, 2004), pp. 305-308.
- Christ, Oli (1994) "A modular and flexible architecture for an integrated corpus query system". *Proceedings COMPLEX'94*, Budapest.
- Henrichsen, P.J. (1998) "Peeking Into the Danish Living Room – Internet access to a large speech corpus". *Proceedings of 11th NoDaLiDa*. pp.109-119.