Utgreiing om tekstkorpus til språkteknologisk forsking og utvikling

Av professor Janne Bondi Johannessen*, førsteamanuensis Lilja Øvrelid**, senioringeniør Kristin Hagen*, vitenskapelig assistent/rådgiver Kari Kinn*# og rådgiver Per Erik Solberg#

*Tekstlaboratoriet, Institutt for lingvistiske og nordiske studier, Universitetet i Oslo

** Institutt for informatikk, Universitetet i Oslo

Språkbanken, Nasjonalbiblioteket

Formål: å utrede behovet for trenings- og testkorpus som er manuelt annotert av fagpersoner.

Denne utgreiinga bygger på brev av 1. juli fra Nasjonalbiblioteket til Tekstlaboratoriet ved professor Janne Bondi Johannessen. Utgreiinga er i hovedsak gjort av de tre førstnevnte forfatterne, mens Kinn og Solberg kom inn i prosessen mot slutten av arbeidet. Nedenfor følger vi hovedsaklig de punktene som er nevnt i bestillingsbrevet.

Vi vil først og fremst vektlegge syntaktisk annoterte korpus i denne utgreiinga, og komme inn på morfologisk oppmerkede korpus bare der det sies eksplisitt. Målet med utgreiinga er å diskutere bakgrunn og behov for syntaktiske korpus, og dessuten komme med anbefalinger om valg av annotering, størrelse, metode o.a. for å oppnå korpus av høy standard.

Bak i dokumentet finnes en liste over url-er og referanser til prosjekter og litteratur som er nevnt i utgreiinga.

Bakgrunn

Noen begreper

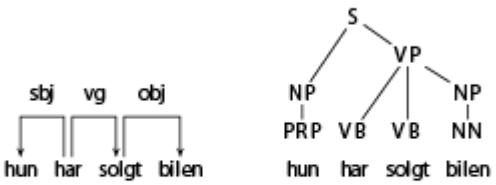
Innledningsvis vil vi slå fast at denne utgreiinga i hovedsak omhandler korpus som er annotert med syntaktisk struktur, også kalt trebanker. Slike korpus inneholder gjerne også ulike typer morfologisk og semantisk informasjon. Et korpus som er manuelt annotert og kvalitetssikret, sies å ha gullstandard. Et slikt korpus vil vi kalle gullkorpus. Dersom korpuset er manuelt sjekket av to personer, vil vi her si at det har platinastandard, mens et korpus som ikke er manuelt gjennomgått i sin helhet, vil ha sølvstandard. Når vi nedenfor snakker om korpus, mener vi, med mindre annet er sagt, et *tekstkorpus til språkteknologisk forsking og utvikling*. Hvilke typer informasjon et norsk gullkorpus bør inneholde, vil vi komme tilbake til. Vi antar at gullkorpuset skal være fritt tilgjengelig for alle – både forskning og industri.

Noen korpus er annotert med hensyn til en spesifikk grammatisk teori, for eksempel BulTreeBank (Simov et al. 2002), som er annotert i *Head-driven Phrase Structure Grammar* (se for eksempel

Pollard & Sag 1994) eller INESS (se hjemmeside), som skal annoteres i *Lexical Functional Grammar* (se for eksempel Bresnan 2001). Mange korpus er mer teoriuavhengige. For syntaktisk annotering er det to hovedtyper: ordene i korpusene er organisert enten som *frasestrukturtrær* eller *dependenstrær*. Frasestrukturtrærne representerer en hierarkisk organisering av frasene, såkalte *konstituente*r, se for eksempel Penn Treebank (Marcus et al. 1993). En dependensrepresentasjon kjennetegnes av relasjoner mellom leksikale enheter, altså ord. Her finnes ikke noe begrep om fraser, selv om disse i stor grad kan utledes basert på en dependensrepresentasjon, se for eksempel The Prague Dependency Treebank (se hjemmesiden).

Illustrasjon 1 viser to analyser av samme setning med et dependenstre til venstre og et frasestrukturtre til høyre.

Illustrasjon 1: to setningstrær, et dependenstre til venstre og et frasestrukturtre til høyre.



I dependenstreet markerer pilene dependentene slik at man kan lese at i setningen ”Hun har solgt bilen” er ”har” hode med ”hun” og ”solgt” som dependenter. Dependensrelasjonene er videre markert med syntaktisk funksjon: ”hun” er subjekt (sbj), ”bilen” er objekt (obj), osv. I frasestrukturtreet er det gjort en første inndeling i NP og VP, altså to fraser som indirekte representerer subjekt og predikat.

Kort historisk bakgrunn

Det språkteknologiske forskningsområdet har de siste tiårene vært preget av det som ofte omtales som en empirisk revolusjon. Økende tilgjengelighet av annoterte korpusdata har gjort det mulig å anvende statistiske metoder og såkalte maskinlæringsalgoritmer for å utlede modeller som utfører ulike former for lingvistisk annotering automatisk. Korpusdata annotert med syntaktisk informasjon har stått helt sentralt i utviklingen og må nok ta noe av æren for at parsing er i ferd med å bli en moden teknologi med imponerende resultater for en rekke språk.

Da Penn Treebank for engelsk ble tilgjengelig på begynnelsen av 90-tallet, førte det til en eksplosjon av parsere trent på dette engelske datasettet. De såkalte *CoNLL shared tasks* har også har også preget utviklingen innenfor parsingsfeltet. Dette er forskningskonkurranser som organiseres årlig under det fulle navnet Computational Natural Language Learning, og som i årene 2006-2009 helt eller delvis var viet til emnet data-drevet dependensparsing (Nivre et al. 2007; Hajič et al. 2009). Via CoNLL har dependenstrebanker for en rekke språk blitt tilgjengeliggjort for forskersamfunnet: svensk, dansk, tysk, nederlandsk, tsjekkisk, spansk, arabisk, kinesisk o.a. Konkurransene har

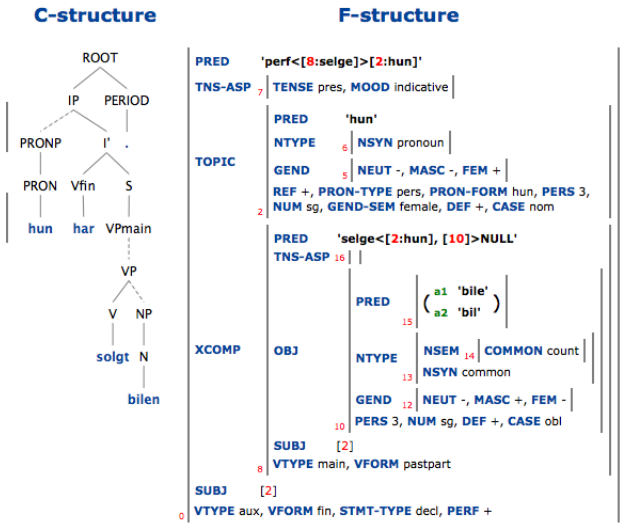
muliggjort evaluering av forskjellige parsere på en rekke forskjellige språk og har stimulert til videre forskning innenfor dependensparsing. Dessverre har ikke norsk vært en del av denne utviklingen siden det per i dag ikke finnes noen egnet trebank for norsk.

Norske og nordiske syntaktiske korpus

I de fleste land det er naturlig å sammenlikne seg med, er det utviklet syntaktiske korpus til forskning og utvikling. I Norden kan vi nevne Swedish Treebank, Danish Dependency Treebank, Turku Dependency Treebank (finsk) og Icelandic Parsed Historical Corpus (The Icelandic Treebank), se lenker til hjemmesidene. Disse har enten helt eller delvis gullstandard.

I Norge har vi så langt få offentlig tilgjengelige syntaktiske korpus, men INESS-prosjektet ved Universitet i Bergen (med forprosjekt TREPIL) er i ferd med å utvikle en stor trebank for bokmål. INESS - Norwegian Infrastructure for the Exploration of Syntax and Semantics (2010 – 2015) skal bli tilgjengelig for forskning, men vil ikke kunne brukes til kommersiell utvikling uten videre, siden det brukes en proprietær løsning (XLE-systemet, se under *lenker*). INESS er en LFG-trebank med en svært detaljert annotasjon både morfologisk, syntaktisk og semantisk, se illustrasjon 2 og vedlegg 3.

Illustrasjon 2: setning fra INESS med frasestruktur og funksjonell struktur



Ellers finnes det bare små tilløp til syntaktiske korpus for norsk. Gjennom samarbeidet The Nordic Treebank Network (2003 – 2005) ble The Sofie Treebank påbegynt med setninger fra de to første kapitlene av Jostein Gaarders roman *Sofies verden*, tilgjengeliggjort ved Tekstlaboratoriet, UiO. Setningene er analysert for 10 språk, deriblant bokmål. Tekstlaboratoriet har dessuten i samarbeid med VISL-prosjektet ved Syddansk universitet i Odense laget en samling av flere hundre analyserte setninger for bokmål og nynorsk. Dette er stort sett enkeltsetninger hentet fra barne- og ungdomslitteratur. Setningene brukes til grammatikkspill og setningsanalyseaktiviteter for barne-, ungdoms- og videregående skole.

For nordsamisk finnes et fritt tilgjengelig korpus på i underkant av 50 000 ord fra Senter for samisk språkteknologi ved Universitetet i Tromsø. Teksten er morfologisk annotert, dependensanalysert og manuelt sjekka, og er hentet fra Johan Muris roman *Muitalus sámiid birra*. Senter for samisk språkteknologi har også små gullkorpus på ca 1300 ord per språk for nordsamisk, lulesamisk, og sørsamisk.

For et lite språk som norsk vil det være naturlig å dra lærdom av hva som er gjort i andre tilsvarende korpusprosjekt. Vi tenker først og fremst på store prosjekt i nabolandene Sverige og Danmark, men selvfølgelig også på det pågående INESS-prosjektet nevnt ovenfor. I tillegg finnes PROIEL-prosjektet ved Universitetet i Oslo og Menotec-prosjektet ved universitetene i Oslo og Bergen. Disse prosjektene arbeider riktignok med data fra tidligere språkhistoriske perioder men spørsmålene de har tatt stilling til med hensyn til analyse og metode, likner mye på de valgene vi må gjøre for norsk.

INESS-prosjektet har vi alt nevnt. Se også vedlegg 3. Nedenfor vil vi kort beskrive Swedish Treebank, Danish Dependency Treebank og PROIEL/Menotec. I kapitlet om faglige krav til et norsk gullkorpus nedenfor og i vedlegg 2 er taggsettene fra disse tre korpusene sammenliknet.

Swedish Treebank består av ca 1,3 millioner ord fra Talbanken (ca 300 000 ord) og SUC-korpuset (se hjemmesider). Talbanken er et korpus med skriftlige og muntlige tekster fra 1970-tallet. Korpuset er manuelt annotert med hensyn til morfologi og syntaks, senere er annoteringen konvertert til dependensstrukturer. SUC – Stockholm Umeå Corpus – er et balansert korpus med svenske skriftlige tekster fra 1990-tallet. Dette korpuset er annotert manuelt med hensyn til morfologi og automatisk med hensyn til dependenssyntaks.

Materialet fra Talbanken er mye brukt til parserutviklingen i det språkteknologiske universitetsmiljøet i Sverige, for eksempel til utviklingen av dependensparseren MaltParser, som har gjort det svært godt i ulike CoNLL-konkurranser (se hjemmesiden til Malt-Parseren for referanser og CoNLL-resultater).

Den svenske trebanken har et taggsett som er forholdsvis detaljert, men som likevel med noen forenklinger og tilpassinger ligger nært opp til output-et fra Oslo-Bergen-taggeren, jf diskusjonen senere i dokumentet og vedlegg 2.

Danish Dependency Treebank er en del av Copenhagen Dependency Treebank som består av to deler, en dansk-engelsk på 95 000 ord og en ren dansk. Den siste er på 100 000 håndtaggede ord med morfologi og syntaks. Setningene er hentet fra PAROLE-korpuset, et korpus med flere sjangre representert. Taggsettet er stort, men det er utarbeidet en svært detaljert manual for annotasjonen. Danish Dependency Treebank var med i CoNLL 2006.

Innenfor rammen av PROIEL-prosjektet utvikles et parallellkorpus med tekster fra gammelgresk, latin, gotisk, armensk og kirkeslavisk. De fleste av tekstene er hentet fra Det nye testamentet. Analysene skal brukes til en lingvistisk studie av de gamle indoeuropeiske språkene, og språkteknologi er ikke nevnt blant bruksområdene. Egnet annotasjons- og søkeverktøy er utviklet av prosjektet. Menotec-prosjektet bruker hovedsakelig de samme verktøyene som PROIEL, og har et liknende siktemål.

Praktisk anvendelse for forskning og industri

Gullkorpus med syntaktisk analyse har en rekke anvendelser innenfor språkteknolgiske applikasjoner. Med et slikt korpus vil det være mulig å trene en eller flere ordklassetaggere og parsere for et språk, samt evaluere disse mot den manuelle annoteringen. Det siste punktet peker på en svært viktig egenskap ved et gullkorpus, nemlig det faktum at man kan være sikker på at det gir den korrekte analysen i henhold til de retningslinjene som er gitt. Dette muliggjør kvantitativ evaluering av forskjellige metoder og algoritmer. Et gullkorpus kan dermed brukes til å trene en data-drevet parser direkte, det vil si trene en parser som ikke er basert på en håndkodet grammatikk. En data-drevet parser er billig å utvikle når et gullkorpus foreligger. Gullkorpus er også viktige for utviklingen av mer kostbare grammatikk-baserte parsere, og de aller fleste store grammatikk-drevne parsere inkluderer statistiske modeller trent på gullkorpus.

En god parser kan brukes til annotering av store tale- eller skriftspråkskorpus, som i sin tur kan brukes i for eksempel lingvistisk forskning. I tillegg er parsere nødvendige moduler innenfor språkteknogiske anvendelsesområder: maskinoversettelse (Ding & Palmer 2005), grammatikkontroll (se ett eksempel i Deksne & Skadiņš 2011), automatisk tekstsammendrag og setningskompresjon (Martins & Smith 2009), semantisk søk (Poon & Domingos 2007), ontologilæring (Snow et al. 2006), spørsmål-svar-systemer (Wang et al. 2007) og sentimentanalyse (Wilson et al. 2009).

Utvikling av gullkorpus er kostbart siden hver eneste setning må kvalitetssikres av en eller flere kompetente personer. For forskningsmiljøer og industri i et lite språksamfunn som norsk, vil alternativet som regel være ikke å utvikle teknologi som er avhengige av parsere, siden dette blir for kostbart.

Kommersielle firmaer røper sjelden hvilke moduler som er brukt i deres produkter, men det er naturlig å anta at det er brukt parsere i for eksempel den engelske grammatikkontrollen i Word eller til sammendragproduksjonen for engelsk i søkemotoren Bing.

Faglige krav til et norsk korpus

I avsnittene nedenfor vil vi diskutere hvilke faglige krav som bør stilles til et norsk gullkorpus og komme med anbefalinger. Først vil vi slå fast at dersom et slikt korpus skal ha noen vesentlig verdi for forskning og utvikling, må det ha en viss størrelse. Jo større jo bedre. Dette hensynet får konsekvenser for diskusjonen nedenfor og for de valgene vi til slutt vil anbefale med hensyn til taggsett, granularitet, arbeidsmetode for oppbyggingen og så videre.

Videre vil vi gå ut fra at et norsk gullkorpus bør være både morfologisk og syntaktisk annotert. Begge deler er etterspurt både fra forskning og industri, og det finnes ikke slike korpus så langt for norsk. Alle trebanker det er naturlig å sammenlikne seg med, har dessuten både morfologisk og syntaktisk annotasjon, jf Penn Treebank, Danish Dependency Treebank, Prague Dependency Treebank, TIGER, Swedish Trebank og mange flere.

Frasestruktur eller dependens

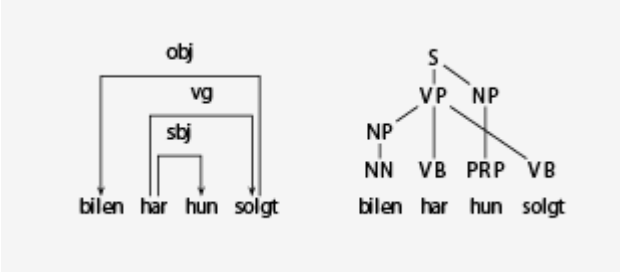
Et grunnleggende spørsmål er hvilken syntaktisk-semantisk annotasjon som skal velges for en norsk trebank. Siden INESS-prosjektet ved Universitetet i Bergen allerede er i gang med å lage en trebank basert på LFG-analyser, synes vi det mest fornuftige ville være å velge en annen innfallsvinkel. Tidsaspektet er også viktig her. Skal det lages et stort gullkorpus, må det velges en metode som ikke er så ressurskrevende som LFG eller HPSG. Da gjenstår valget mellom frasestruktur og dependens. Nedenfor vil vi argumentere for dependens ut i fra følgende kriterier: (i) transparens, (ii) kompleksitet, (iii) ordstilling.

Med transparens mener vi i hvilken grad strukturene er leselige eller anvendbare for videre bruk i språkteknologiske applikasjoner. En fordel med dependensstrukturer i denne sammenhengen er at de representerer predikat-argument-strukturen i en setning (dvs. hva som er subjekt, predikat, objekt, osv.) direkte. I en dependensanalyse vil syntaktiske funksjoner eller relasjoner være primitive, mens i en frasestrukturanalyse vil man måtte avlede slik informasjon fra den syntaktiske strukturen som et ekstra ledd i prosesseringen. Dependensstrukturer har, som nevnt tidligere, blitt mye brukt i språkteknologiske anvendelser, og det er nettopp transparensen som gjør denne typen analyser så populære. Eksempelapplikasjonene og referansene som er angitt i seksjonen om praktisk anvendelse ovenfor, gjelder stort sett bruk av parsere som tildeler en dependensanalyse.

Når det gjelder kompleksitet, kan vi skille mellom praktisk kompleksitet, hensyn som påvirker maskinell prosessering, og teoretisk kompleksitet, som vil påvirke de algoritmiske hensynene (tids-minnebruk) når disse strukturene skal prosesseres. En dependensstruktur er gitt ved visse formelle kriterier, som f.eks. at en dependent har ett og kun ett hode. Dette gjør at parsing kan reduseres til å merke hvert ord i en tekst med a) dets hode, og b) dets dependensrelasjon. Dette går tydelig frem i CoNLL-formatet (se eksempel 4 og 5 nedenfor), der vi finner ett ord per linje. Når det gjelder teoretisk kompleksitet, har det blitt vist at visse egenskaper ved dependensstrukturer gjør at det i visse tilfeller er mulig å forbedre algoritmisk kompleksitet sammenliknet med frasestrukturparsing (Eisner 1996).

Dependensgrammatikk er blitt særlig mye brukt av lingvister med andre språk enn engelsk som morsmål. En grunn til dette er sannsynligvis at dependensrepresentasjoner teoretisk sett er uavhengig av leddstilling og enkelt kan representere variasjon i leddstilling. Da norsk inneholder en del variasjon i leddstilling er det naturlig å evaluere disse formalismenes uttrykkskraft i forhold til denne typen variasjon. I illustrasjon 1 ovenfor så vi en dependensstruktur og en frasestrukturanalyse for setningen "Hun har solgt bilen". I illustrasjon 3 ser vi en enkel setning med topikalisering.

Illustrasjon 3: Dependens- og frasestrukturanalyse av ”bilen har hun solgt”



Vi antar her en enklest mulig (ateoretisk) frasestruktur, grunnet både annoteringskompleksitet og videre språkteknologisk anvendelse.[1] Vi ser at den topikaliserte varianten i dependensanalysen får en struktur der de syntaktiske relasjonene tydelig fremgår, til tross for at objektsleddet er foranstilt. Frasestrukturanalysen, derimot, blir ikke velformet, og det er vanskelig å tenke seg en frasestrukturanalyse for denne setningen uten at det må innføres flere nivåer av syntaktisk struktur. Selv da vil ikke den frasestrukturelle analysen vise de syntaktiske relasjonene som holder mellom det topikaliserte leddet og det leksikalske verbet. En frasestrukturanalyse gir ikke uten videre opplysning om hvor i setningen spørreleddet hører til, som i: "Når tror du at dere kommer?".

I tillegg til kriteriene beskrevet over kan vi også nevne at en norsk trebank i dependensformat vil bidra til at vi vil kunne ta del i det internasjonale forskningsfellesskapet som CoNLL-konkurransene har resultert i, i form av språkteknologiske verktøy for trening av taggere og parsere, program for evaluering av disse verktøyene, samt muliggjøre sammenligninger på tvers av språk.

Konklusjon: Vi anbefaler en dependensanalyse for et nytt norsk gullkorpus.

Formater og verktøy

Det finnes mange dependensformater å velge mellom. Heldigvis inneholder de ulike formatene ofte de samme kategoriene, slik at det vil være en enkel programmeringsoppgave å konvertere korpusmateriale fra et format til et annet[2]. Mange korpus er også nedlastbare i flere formater, for eksempel Swedish Trebank som finnes i MALT XML, Tiger XML og CoNLL-format .

I MALT XML har et ord fem attributter: id, ordform, ordklasse, syntaktisk hode og dependensrelasjon til hode, se eksempel 1 fra Swedish Treebank:

Eksempel 1:

```
<sentence id="24">
          <word id="1" form="Dessutom" postag="ab" head="2" deprel="ADV"/>
          <word id="2" form="höjs" postag="vb.prs.sfo" head="0" deprel=""/>
          <word id="3" form="åldergränsen" postag="nn.utr.sin.def.nom" head="2" deprel="SUB"/>
          <word id="4" form="till" postag="pp" head="2" deprel="ADV"/>
          <word id="5" form="18" postag="rg.nom" head="6" deprel="DET"/>
          <word id="6" form="år" postag="nn.neu.plu.ind.nom" head="4" deprel="PR"/>
          <word id="7" form="." postag="mad" head="2" deprel="IP"/>
</sentence>
```

MALT XML er innformatet for Malt-Parseren, som er utviklet ved universitetene i Växjö og Uppsala (se hjemmeside).

TIGER XML er et utvekslingsformat som er bygget opp på en annen måte. I eksempel 2 er samme setning som ovenfor i TIGER XML.

[1] De aller fleste trebanker per i dag som inneholder frasestrukturanalyse, opererer med flate strukturer, se f.eks. Penn Treebank og den tyske NEGRA-trebanken.
[2] Mange konverteringsprogrammer finnes allerede, se for eksempel en oversikt her: http://beta.visl.sdu.dk/treebanks.html

Eksempel 2:

```
<s id="s24">
   <graph root="s24_500">
      <terminals>
         <t id="s24_1" word="Dessutom" pos="ab"></t>
         <t id="s24_2" word="höjs" pos="vb.prs.sfo"></t>
         <t id="s24_3" word="åldergränsen" pos="nn.utr.sin.def.nom"></t>
         <t id="s24_4" word="till" pos="pp"></t>
         <t id="s24_5" word="18" pos="rg.nom"></t>
         <t id="s24_6" word="år" pos="nn.neu.plu.ind.nom"></t>
         <t id="s24_7" word="." pos="mad"></t>
      </terminals>
      <nonterminals>
         <nt id="s24_506" cat="nn.neu.plu.ind.nom">
            <edge idref="s24_6" label="--"></edge>
            <edge idref="s24_5" label="DET"></edge>
         </nt>
         <nt id="s24_504" cat="pp">
            <edge idref="s24_4" label="--"></edge>
            <edge idref="s24_506" label="PR"></edge>
         </nt>
         <nt id="s24_502" cat="vb.prs.sfo">
            <edge idref="s24_2" label="--"></edge>
            <edge idref="s24_1" label="ADV"></edge>
            <edge idref="s24_3" label="SUB"></edge>
            <edge idref="s24_504" label="ADV"></edge>
            <edge idref="s24_7" label="IP"></edge>
         </nt>
         <nt id="s24_500" cat="ab">
            <edge idref="s24_502" label="--"></edge>
         </nt>
      </nonterminals>
   </graph>
</s>
```

For setninger i TIGER XML er det laget et søkeprogram – TIGERSearch. Dette programmet blir dessverre ikke vedlikeholdt for øyeblikket, og lar seg ikke installere på Windows7-maskiner. TIGER XML er brukt av flere prosjekter, blant annet den tyske TIGER-trebanken (se hjemmeside).

Constraint Grammar-taggere (se Karlsson et al. 1995 og VISL-nettsiden) som Oslo-Bergen-taggeren (Johannessen et al. under utgivelse) bruker gjerne et format som VISL-formatet i eksempel 3. Eksemplet er hentet fra VISL-siden og viser en dansk setning som også har en dependensanalyse:

Eksempel 3:

```
Når [når] KS @SUB #1->4
Sofies [Sofie] PROP GEN @>N #2->3
mor [mor] N UTR S IDF NOM @SUBJ> #3->4
var [være] V IMPF AKT @FS-ADVL> #4->9
sur [sur] ADJ UTR S IDF NOM @4
over [over] PRP @A< #6->5
et=eller=andet [en=eller=anden] DET NEU S NOM @P< #7->6
$, #8->0
skete [ske] V IMPF AKT @FS-STA #9->0
det [den] PERS NEU 3S NOM @F-9
at [at] KS @SUB #11->13
hun [hun] PERS UTR 3S NOM @SUBJ> #12->13
kaldte [kalde] V IMPF AKT @FS-9
deres [de] PERS 3P GEN @>N #14->15
hus [hus] N NEU S IDF NOM @13
for [for] PRP @13
et [en] ART NEU S IDF @>N #17->19
værre [dårlig] ADJ COM nG nN nD NOM @>N #18->19
menageri [menageri] N NEU S IDF NOM @P< #19->16
$. #20->0
```

CoNLL-formatet er brukt til de tidligere nevnte shared task-oppgavene ved konferansene *Conference on Computational Natural Language Learning*, og mange gullkorpus har derfor en nedlastbar versjon i dette formatet (jf. for eksempel Swedish Treebank, Danish Dependency Treebank, Prague Dependency Treebank).

CoNLL består av ti faste felter atskilt med tab: ID, FORM, LEMMA, CPOSTAG (morfologisk tagg), POSTAG (mer detaljert morfologisk tagg), FEATS (morfologiske og syntaktiske trekk atskilt med |) HEAD, DEPREL, PHEAD, PDEPREL. Bare ID, FORM, CPOSTAG, POSTAG, HEAD og DEPREL må fylles ut.

Eksempel 4 er hentet fra Danish Dependency Trebank, eksempel 5 fra Swedish Treebank. Som eksemplene viser, kan taggene ha ulikt format, navn og kompleksitet. Det essensielle er at de obligatoriske feltene er fylt ut.

Eksempel 4:

1	Træ	_	N	NC	gender=neuter|number=sing|case=unmarked|def=indef				2	subj
	_	_								
2	bliver	_	V	VA	mood=indic|tense=present|voice=active			0	ROOT	_	_
3	til	_	SP	SP	_	2	pred	_	_		
4	papir	_	N	NC	gender=neuter|number=sing|case=unmarked|def=indef				3	nobj
	_	_								
5	.	_	X	XP	_	2	pnct	_	_		

Eksempel 5:

1	De	_	PO	PO	_	2	SS	_	_
2	kan	_	QV	QV	_	0	ROOT	_	_
3	alltså	_	AB	AB	_	2	+A	_	_
4	inte	_	AB	AB	_	2	NA	_	_
5	få	_	FV	FV	_	2	VG	_	_
6	full	_	AJ	AJ	_	7	AT	_	_
7	Atp	_	PN	PN	_	5	OO	_	_
8	.	_	IP	IP	_	2	IP	_	_

Ved Prague Dependency Treebank-prosjektet er det laget verktøy for både søk og redigering (Netgraph og TrEd). Prosjektet bruker sitt eget format (PML - Prague markup language), men har laget konverteringsverktøy slik at setninger i CoNLL-format kan søkes i og redigeres via Netgraph og TrEd. Verktøyene videreutvikles og vedlikeholdes fremdeles, og det er gode brukermanualer tilgjengelig via prosjektsidene på nettet. Swedish Treebank har brukt TrEd ved den siste redigeringen av trebanken.

Formatene ovenfor gir bare plass til forholdsvis enkel morfologisk og syntaktisk informasjon. Skulle man ønske å bygge ut et gullkorpus med mer informasjon – for eksempel semantisk – gir PML-formatet til Prague Dependency Treebank anledning til det.

PROIEL-prosjektet har, som vi nevnte, også et annoteringsverktøy. Dette har vi sett på, men det har ikke den samme brukersupporten som PML, og er også mindre fleksibelt.

Konklusjon: Vi anbefaler PML-verktøyet og CoNLL-formatet som grunnformat.

Tekstutvalg

Tekstene til mange kjente trebanker består av avistekster (for eksempel tyske TüBa-D/Z og TIGER eller Prague Dependency Treebank), og mange parsere har utelukkende vært trent på tekster fra en avis, for eksempel Wall Street Journal (en del av Penn Treebank). Andre korpus har tekster også fra andre kilder, for eksempel sakprosatekster og skjønnlitteratur (Danish Dependency Treebank og Swedish Treebank). Noen har også med tekster fra muntlige kilder (Penn Treebank og Swedish Treebank).

Dersom det norske gullkorpuset utelukkende skulle brukes av lingvister, ville det uten tvil være en fordel at det var så bredt sammensatt som mulig med hensyn til teksttyper. Når formålet derimot er språkteknologisk forsking og utvikling, mener vi det er en fordel at hoveddelen består av avistekst og sakprosatekst. I litterære tekster er språkbruken ofte poetisk, og brudd på skriveregler og normer er vanlig. Dette gjør det vanskelig å trene gode parsere. Dette er en erfaring noen av forfatterne av dette dokumentet har etter å ha arbeidet med tagging av korpus som inneholder mye skjønnlitteratur.

Tekster fra sosiale medier som Twitter og Facebook er populære analyseobjekter for både forsking og industri akkurat nå. Dette er tekster med muntlig preg i forhold til avis- og saktekster.

Siden avistekster er det som oftest brukes i andre syntaktiske korpus, og avis inneholder mye forskjellig type tekst, anbefaler vi en god del av dette. I tillegg synes vi det er riktig med en del tørr sakprosa av typen rapporter, som NOU-utredninger. Dette er vår anbefaling:

60 % avistekst

20 % rapporttekst

10 % tekst med muntlig preg (sms, sosiale medier)

10 % skjønnlitteratur

Tekstene må naturligvis merkes på en slik måte at brukere kan velge å plukke ut tekster de ikke ønsker å bruke.

Innenfor hver sjanger må det også foretas valg. Avistekster inneholder for eksempel overskrifter, bylines, bildetekster, fotballtabeller, vinanmeldelser osv. som ikke oppfører seg slik vanlig norske

setninger gjør. Vi vil likevel anbefale at korpuset inneholder slike elementer slik at parsere trent på materialet blir i stand til å behandle virkelig tekst fra for eksempel aviser.

Konklusjon: Vi anbefaler et tekstutvalg med mye avistekst (60 %) og noe fra andre sjangre (se ovenfor). Vi anbefaler at trebanken også inneholder setninger som overskrifter og bildetekster.

Metadata

Et norsk gullkorpus må inneholde relevante metadata, for eksempel om tekstene. Språkbanken bør avgjøre hvilke format.

Størrelse

Moderne trebanker er forholdsvis store. Noen eksempler: Prague Dependency Treebank: 1,5 millioner ord med syntaktisk annotasjon, Swedish Trebank: ca 1,3 millioner ord, Penn Treebank: 4,5 millioner ord. Andre er mindre: Danish Dependency Treebank: 100 000 ord.

I faglitteraturen kan man ikke finne noen absolutt anbefaling når det gjelder hvor stort et gullkorpus bør være, bare at det bør være over en viss størrelse, slik de fleste moderne gullkorpus er. Basert på diskusjonen ovenfor vil vi anbefale at det lages et gullkorpus for bokmål og et gullkorpus for nynorsk på 1 million ord hver.

Å lage gullkorpus er tidkrevende – selv om man tar gode preprosesseringsverktøy i bruk. Det er derfor ikke sikkert at det er mulig å analysere 1 million ord innenfor den tidsrammen som er gitt. Vi anbefaler platinastandard (gjennomgått av to annotører) på forslagsvis 200 000 ord, og gullstandard (gjennomgått av én annotør) på resten. Imidlertid bør man vurdere om man skal tåle sølvstandard når prosjektet har vart en stund. Det er viktig å ha en viss størrelse på korpusene. Sølvstandard kan oppnås ved at man trener en statistisk parser på de gullkorpussetningene som er produsert, og så kan disse parserne brukes til å analysere resten av materialet – med sølvstandard. Konklusjon: Det bør lages et korpus på 1 million ord for hver målform. Minst 200 000 ord bør ha platinastandard.

Arbeidsprosess

For å få flest mulig analyser på kortest mulig tid er det, som tidligere nevnt, en stor fordel å kunne preprosessere setningene slik at de faglige annotørene ikke behøver å analysere setningene fra bunnen av, men kun behøver å rette eller endre på analyseforslag. Under arbeidet med Penn Treebank beregnet man at manuell morfologisk tagging tok dobbelt så lang tid som retting av automatisk tagget tekst. Feilprosenten var også omkring 50 prosent høyere ved manuell tagging sammenliknet med retting av automatisk tagget tekst (Marcus et al. 1993).

Til preprosessering av norsk vil det være mest effektivt å benytte seg av gode analyseverktøy som finnes fra før. Oslo-Bergen-taggeren gir en morfologisk analyse av høy kvalitet med en accuracy på 95.7 % for OBT+stat (Johannessen et al., under utgivelse). Den gir også en syntaktisk analyse med

underspesifiserte dependensrelasjoner (ikke evaluert). Oslo-Bergen-taggeren gir et output-format som enkelt lar seg konvertere til formater som CoNLL, som i sin tur kan brukes i for eksempel PML-verktøyet utviklet av Prague Dependency Treebank for søking og redigering. Når man har laget et gullkorpus av en viss størrelse, kan man trene en parser på materialet og se om det er raskest å erstatte OBT+stat med den nyutviklede parseren.

Oslo-Bergen-taggeren bruker Norsk ordbank, en fullformsliste med et taggsett harmonisert med Norsk referansegrammatikk (Faarlund et al. 1997). Leksikonet er ikke beriket med semantiske opplysninger.

Siden korpuset skal holde gullkorpusstandard, bør alle setninger i prinsippet, dvs. så langt man rekker, gjennomgås to ganger av to ulike fagannotører, slik at man oppnår det vi har kalt platinastandard. Slik slipper man korrektur, men setningene man er uenige om, må selvsagt behandles en tredje gang – enten det er snakk om slurvefeil eller mer prinsipielle uenigheter.

Konklusjon: Vi anbefaler at Oslo-Bergen-taggeren med tilhørende verktøy benyttes for morfologisk analyse og preprosessering. TrEd og Netgraph anbefales for redigering og søk i trær.

Morfologisk og syntaktisk taggsett

Siden vi anbefaler å bruke Oslo-Bergen-taggeren til preprosessering, tror vi det vil være mest hensiktsmessig å bruke det morfologiske taggsettet fra denne taggeren (se vedlegg 1). På denne måten vil mye av den morfologiske annotasjonen være klar allerede før den manuelle annoteringsjobben begynner. Det er dessuten lagt ned store ressurser i utviklingen av OB-taggerens taggsett, og taggsettet følger de morfologiske analysene i Norsk referansegrammatikk (Faarlund et al. 1997). Det virker derfor lite hensiktsmessig å utvikle et nytt taggsett fra bunnen av. Enkelte små tilpasninger vil sannsynligvis være nødvendig, for eksempel av hensyn til sammenhengen mellom den morfologiske annotasjonen og den syntaktiske dependensanalysen.

OB-taggeren har i tillegg til det morfologiske taggsettet et sett av syntaktiske tagger (se vedlegg 2 i kolonnen lengst til høyre). Vi anbefaler å bruke disse taggene som utgangspunkt for gullkorpusenes syntaktiske taggsett, av de samme grunnene som er nevnt for det morfologiske taggsettet ovenfor. Vi vil likevel gjøre en del justeringer for å kunne gi en lingvistisk adekvat dependensanalyse med et noe høyere presisjonsnivå enn det OB-taggerens taggsett gir grunnlag for. Her kan vi dra nytte av arbeidet som er gjort med å utvikle syntaktiske tagger i korpus for andre språk, og vi mener at det er særlig mye å hente i taggsettet som brukes i Talbanken 05 (se også vedlegg 2, andre kolonne). Dette taggsettet har i hovedtrekk mye til felles med OB-taggerens taggsett, men inneholder flere interessante syntaktiske distinksjoner, for eksempel mellom ulike typer subjekter. Samtidig er det mindre enn taggsettet som brukes i Copenhagen Dependency Treebank (se vedlegg 2, tredje kolonne), som er svært omfattende og inneholder mange distinksjoner som vi ikke vil anbefale å gjøre i gullkorpusene, både av hensyn til prosjektets ressurser og korpusenes formål. Dette gjelder for eksempel detaljerte semantiske og pragmatiske distinksjoner mellom undertyper av samme syntaktiske funksjon.

Taggsettet til PROIEL/Menotec (se vedlegg 2, første kolonne) er ikke utviklet med tanke på automatisk analyse, og vi har derfor valgt å se bort fra det.

Konklusjon: På grunnlag av taggene i OB-taggeren og Talbanken 05 samt egne lingvistiske vurderinger, vil vi foreslå følgende syntaktiske taggsett for gullkorpusene:

Tabell 1:

FINV	Finitt verb (hjelpeverb der dette finnes). ”Hun *har* solgt bilen.”
INFV	Infinitt verb (hovedverb). ”Hun har *solgt* bilen.”
SUBJ	Subjekt. ”*Hun* har solgt bilen.”
FSUBJ	Formelt subjekt. ”*Det* kommer en bil på veien.”
PSUBJ	Potensielt subjekt (dvs. egentlig subjekt/logisk subjekt). ”Det kommer *en bil* på veien.”
SPRED	Subjektspredikativ. ”Du er *snill*.”
OPRED	Objektspredikativ. ”Hun kalte ham *snill*.”
FSPRED	Fritt subjektspredikativ. ”De gikk *slitne* hjem.”
FOPRED	Fritt objektspredikativ. ”De kjøpte huset *usett*.”
DOBJ	Direkte objekt. ”Hun har solgt *bilen*.”
IOBJ	Indirekte objekt. ”Hun ga *ham* bilen.”
FOBJ	Formelt objekt. ”De sa *det* rett ut at hun ikke kunne synge.”
POBJ	Potensielt objekt. ”De sa det rett ut *at hun ikke kunne synge*.”
PUTFYLL	Utfylling til preposisjon. ”Katten satt på *taket*.”
VPART	Verbpartikler. ”Har du husket å slå *av* lyset?”
ADV	Adverbial. ”De spiste middag *på kjøkkenet*.”
DET	Bestemmende adledd (j.f. Faarlund et al. (1997: §3.3)). ”*den* vakre byen”
ATR	Beskrivende adledd (j.f. Faarlund et al. (1997: §3.3)). ”den *vakre* byen”
APP	Apposisjon. ”Harald, *kongen av Norge*, inviterte til middag.”
TITTEL	Tittel. ”*Kong* Harald inviterte til middag.”

KONJ	Konjunksjon. ”Hun kjøpte kaffe *og* te.”
SBU	Subjunksjon. ”Hun sa *at* hun likte kaffe og te.”
SBUREL	Subjunksjon i relativsetning. ”Hun likte bare kaffe *som* var nykvernet.”
INTERJ	Interjeksjon. ”*Hurra*!”
PAR	Hodet i parentetiske innskudd. ”Har man jobbet i 40 år, har man tjent opp full pensjon (halve antall år *gir* rett til halv pensjon).”
FRAG	Hodet i konstituenter som ikke har syntaktisk forbindelse til noe verb, f. eks. i overskrifter: ”Ny *seier* til Ålesund.”
IP	Interpunksjon, tegnsetting som avslutter helsetning. ”Har du kjøpt kaffe*?*”
IK	Interpunksjon inne i setningen (komma). ”Hun hadde kjøpt kaffe*,* te og melk.”

Noen sentrale analysevalg

Nedenfor gjør vi rede for noen sentrale analysevalg vi anbefaler for dependensanalysene i gullkorpusene. Anbefalingene er basert på sammenlikninger av analysene i Menotec og PROIEL, Talbanken 05 og Copenhagen Dependency Treebank (se vedlegg 2), i tillegg til lingvistiske hensyn og hensyn til at korpusene skal kunne trene opp parsere på en best mulig måte. Vi har valgt å bruke Talbanken 05 som modell der vi ikke har sterke motargumenter mot dette, både fordi analysene i mange tilfeller er gode for vårt formål, og fordi det er praktisk å kunne bruke Talbanken som rettesnor i tvilstilfeller.

Øverste hode i treet: Vi følger Talbanken 05, som har det finitte verbet som øverste hode i treet, uavhengig av om det er et leksikalsk verb eller et hjelpeverb.

Nominalfraser: Vi lar substantivet være hode i nominalfraser, slik de gjør i Talbanken 05. Det er vanlig praksis i dependensgrammatikk å la leksikalske ord være hoder og funksjonsord dependenter, og denne praksisen ønsker vi å følge så langt det er mulig. Et alternativ kunne være å la determinativet være hode, noe som praktiseres i Copenhagen Dependency Treebank (se vedlegg 2). Denne løsningen ville imidlertid medført en inkonsekvent analyse og dermed vanskeligheter for parserne som skal trenes opp på korpuset: Ettersom mange nominalfraser ikke inneholder noe determinativ, ville substantivet uansett måtte være hode i mange tilfeller.

Koordinering: For parsere er det problematisk å ha konjunksjonen som hode, fordi det da ikke er synlig hvilken funksjon konjunktene har. Vi lar derfor det første koordinerte leddet være hode, slik

som i analysen beskrevet i Nivre et al. (2006). Denne analysen er fulgt i tidlige utgaver av Talbanken 05, men i den siste utgaven er konjunksjonen ofte valgt som hode (se vedlegg 2). Vi tar ikke på det nåværende tidspunkt stilling til hvordan konjunksjonen og andre koordinerte ledd bør analyseres.

Topikalisering av ledd fra leddsetning: For å unngå kryssende dependenser velger man i noen korpus å la topikaliserte elementer som egentlig hører hjemme i en underordnet setning, være dependent på den øverste noden (se vedlegg 2). Vi ønsker at analysene i gullkorpusene skal være lingvistisk adekvate. Derfor anbefaler vi at slike topikaliserte elementer blir dependent til det aktuelle hodet i den underordnede setningen.

Leddsetninger: Det finitte verbet bør være hode i leddsetninger. Et nærliggende alternativ er å la subjunksjonen være hode, men dette ville ført til en inkonsekvent analyse, ettersom mange leddsetninger mangler subjunksjon, og det finitte verbet i slike tilfeller uansett må overta hodefunksjonen. Også i relativsetninger vil det finitte verbet være hode, men det kan diskuteres hvorvidt *som* skal regnes som en subjunksjon eller ikke. I Talbanken har de valgt å gjøre *som* i relativsetninger til relativpronomen og gi det funksjonen til det relativiserte elementet. OB-taggeren tagger *som* i slike setninger som *SBU-rel,* en analyse som er i tråd med Norsk referansegrammatikk, hvor *som* regnes som subjunksjon. Hvis *som* annoteres som en subjunksjon, vil imidlertid alle relativsetninger mangle en node for det relativiserte elementet. Vi anbefaler eventuelt at *som* blir tagget som subjunksjon i morfologien, mens det gis funksjonen til det relativiserte elementet i den syntaktiske analysen. Leddsetninger som fungerer som argumenter (subjekt eller objekt), får funksjonene SUBJ og OBJ, og ikke noen egen funksjon som i Menotec/PROIEL.

Tomme noder: Vi anbefaler å ikke operere med tomme noder, da dette er problematisk for parsere.

Skilletegn: Vi følger samme praksis som i Talbanken (se vedlegg 2): Punktum, utropstegn og spørsmålstegn er dependenter på setningens øverste hode. Komma er dependent på konjunksjonen eller på hodet til den setningen det markerer slutten på.

Komparasjon: *Enn* og *som* bør være dependenter på det komparative adjektivet/adverbet.

Kontrollsetninger og liknende: Vi anbefaler å følge Talbanken og la infinitivsmerket være hode (se vedlegg 2). Dette er konsistent med de andre analysevalgene vi har tatt, da et infinitt leksikalsk verb heller ikke er hode i konstruksjoner med hjelpeverb.

Formelle subjekter: I konstruksjoner med formelle subjekter, slik som *Det sto en mann i hagen*, foreslår vi at det formelle subjektet, *det*, får funksjonen FSUBJ, mens det potensielle (logiske) subjektet, *en mann*, får funksjonen PSUBJ. Dette er i utgangspunktet samme praksis som de har valgt i Talbanken (se vedlegg 2). I Talbanken bruker de funksjonen FSUBJ kun i de tilfellene hvor det finnes et potensielt subjekt i tillegg til det formelle subjektet *det,* altså ikke i setninger som *Det regner* (Teleman 1974:46). Vi anbefaler å vurdere å bruke FSUBJ også i disse tilfellene for å få en mer adekvat analyse.

Dokumentasjon

En god dokumentasjon vil øke brukervennligheten til korpuset. Vi anbefaler at det skrives dokumentasjon fortløpende under hele prosjektet.

Tilråding og konklusjon

Nedenfor vil vi kort gjengi hovedkonklusjonene fra tilrådingene i forrige kapittel.

- Syntaktisk-semantisk annotasjon: dependens
- Grunnformat: CoNLL-formatet – som kan formateres til andre format som TIGER XML, MALT XML. Dersom arbeidsprosessen legges opp slik vi anbefaler, vil trebanken også foreligge i PML-format og CG-format
- Tekstutvalg: 60 % avistekst, 20 % rapporttekst, 10 % tekst med muntlig preg (sms, sosiale medier, transkripsjoner), 10 % skjønnlitteratur
- Metadata: formatet som Språkbanken velger å bruke for sine ressurser
- Størrelse: Omkring 1 million ord for bokmål og 1 million ord for nynorsk. Platinastandard på anslagsvis 200 000 ord for henholdsvis bokmål og nynorsk (alt annotert av to uavhengige annotører), og gullstandard på resten (annotert av én annotør). Vurdere sølvstandard på det man ikke rekker (korpus analysert med parser trent på platina- og gullkorpus)
- Arbeidsprosess: Bruke Oslo-Bergen-taggeren med tilhørende verktøy for morfologisk analyse og preprosessering. TrEd og Netgraph for redigering og søk i trær.
- Morfologisk og syntaktisk taggsett: bruke det morfologiske taggsettet til Oslo-Bergen-taggeren. For syntaks brukes en tilpasset og redigert utgave av taggsettene til Oslo-Bergen-taggeren og Talbanken 05 (en del av The Swedish Treebank), se tabell 1.
- Sentrale analysevalg: hovedsakelig følge Talbanken 05
- Dokumentasjon: bør utarbeides fortløpende

Referanser

Bresnan, Joan. 2001. *Lexical Functional Syntax*. Blackwell.

Deksne, Daiga & Skadiņš, Raivis. 2011. CFG based grammar checker for Latvian. I Bolette Sandford Pedersen, Gunta Nešpore & Inguna Skadiņa (red.). *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011.* NEALT Proceedings Series, Vol. 11 (2011), 275-278.

Ding Y., & Palmer M. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. *Proceedings of ACL*, 2005, 541-548.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, 340–345.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antonia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue & Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies

in Multiple Languages. *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*. Boulder, Colorado, USA. June 4-5. 3-22.

Johannessen, Janne Bondi, Hagen Kristin, Lynum, André & Nøklestad, Anders. Under utgivelse. OBT+stat: A combined rule-based and statistical tagger. Kommer i Andersen, Gisle (red.). *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian*. Amsterdam/New York: John Benjamins.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, & Arto Anttila, (red). 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing, No 4. Mouton de Gruyter, Berlin/New York. ISBN 3-11-014179-5.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. *Proceedings of the 11th Conference on Natural Language Learning*, 915 – 932. Prague, Czech Republic.

Faarlund, Jan Terje, Lie, Svein & Vannebo, Kjell Ivar. 1997. *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Marcus, Mitchell P, Santorini, Beatrice & Marcinkiewicz, Mary Ann. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, vol. 19.

Martins, André F. T. & Smith, Noah A. 2009. Summarization with a Joint Model for Sentence Extraction and Compression. ILP '09. *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*.

Pollard, Carl & Sag, Ivan A. 1994. *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Poon H, Domingos P. 2009. Unsupervised semantic parsing. *Procedings of EMNLP*.

Simov, Kiril, Popova, Gergana & Osenova, Petya. 2002. HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In Andrew Wilson, Paul Rayson, & Tony McEnery (red.). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World.* Lincom-Europa, Munich, 135-142.

Snow R., Jurafsky D., Ng A.Y. 2006. Semantic taxonomy induction from heterogenous evidence. *Proceedings of COLING/ACL*.

Wang M., Smith N.A., Mitamura T. 2007. What is the Jeopardy model? A quasi-synchronous grammar for QA. *Proceedings of EMNLP*.

Wilson T., Wiebe J., Hoffmann P. 2009. Recognizing Contextual Polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3), 399-433.

Nettadresser

BulTreeBank: http://www.bultreebank.org/

CoNLL: http://www.cnts.ua.ac.be/conll/

Copenhagen Dependency Treebank/Danish Treebank: http://code.google.com/p/copenhagen-dependency-treebank/

Icelandic Parsed Historical Corpus: http://linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_%28IcePaHC%29

INESS: http://iness.uib.no/iness/

MaltParser: http://maltparser.org/index.html

Menotec: www.menota.org/meldingar/2010-01-06.page

Norsk ordbank: http://www.hf.uio.no/iln/om/organisasjon/edd/forsking/norsk-ordbank/

NEGRA corpus: http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html

Nordic Treebank Network: http://w3.msi.vxu.se/~nivre/research/nt.html

Oslo-Bergen-taggeren: http://tekstlab.uio.no/obt-ny/

Penn Treebank: http://www.cis.upenn.edu/~treebank/

PROIEL-prosjektet: http://www.hf.uio.no/ifikk/english/research/projects/proiel/

The Prague Dependency Treebank: http://ufal.mff.cuni.cz/pdt2.0/

Sofie Treebank: http://www.hf.uio.no/iln/om/organisasjon/tekstlab/prosjekter/arkiv/sofie.html

Stockholm-Umeå Corpus – SUC: http://spraakbanken.gu.se/eng/resources/suc

Swedish Treebank: http://spraakbanken.gu.se/eng/stb

TIGER: http://www.ims.uni-stuttgart.de/projekte/TIGER/

Talbanken05: http://w3.msi.vxu.se/~nivre/research/Talbanken05.html

Johan Turis Muitalus sámiid birra: http://giellatekno.uit.no/turi/turi.nno.html

Turku Dependency Treebank: http://bionlp.utu.fi/fintreebank.html

TüBa-D/Z: http://www.sfs.uni-tuebingen.de/en/tuebadz.shtml

XLE-systemet: http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html

VISL grammatikkanalyser og -spill: http://beta.visl.sdu.dk/

VEDLEGG 1: Morfologiske tagger i Oslo-Bergen-taggeren

ordklasse/tegn	kjønn	tall	type	best	tid	person	kasus	gradbøy
adj	m/f nøyt fem	ent fl	<adv> <ordenstall> <perf-part> <pres-part> fork	ub be				pos kom sup
adv								
det	mask nøyt fem	ent fl	dem dem <adj> <adj> forst <adj> kvant kvant poss poss res poss høflig sp forst	ub be				
inf-merke								
interj								
konj			<adv> clb					
prep								
pron	fem mask mask fem nøyt	ent fl	hum res hum sp pers pers hum pers høflig poss hum sp refl sp res			1 2 3	nom akk	
sbu			<spørreartikkel>					
subst	mask fem nøyt	ent fl	appell prop fork	ub be			gen	
ukjent								
verb					pres inf pass inf pres pret perf-part imp pass			

Tilleggstagger og skilletegn

- Ord kan bli markert som **sammensetninger** (*samset*) eller **unormert** (*unorm*) eller **genitiv** (*gen*)
- Taggeren markerer **setningsgrenser** med *clb* (ved f.eks. komma og konjunksjoner). **Helsetningsgrense** e.l. markeres med <<< .
- **Hermetegn** og **punktum** blir satt sammen med ordet til venstre eller høyre dersom tegnet antas å være en del av ordet, f.eks. når punktumet er en del av en forkortelse: *A. subst prop fork* eller når kun ett ord står i hermetegn:
  <"exit"> (Når flere ord er rammet inn av hermetegn, skilles hermetegn ut som eget tegn <">.)
- Taggeren forsøker også å skille ut **overskrifter**, ved å gi dem en helsetningsgrensetagg |
- Legg merke til at alle store bokstaver blir gjort om til små, men at originalteksten beholdes mellom taggene <word></word>.

Ellipse	<...> , $... clb <ellipse>
Hermetegn	<"> , $ <anf>
Kolon	<:> , $: clb <kolon>
Komma	<,> , $, clb <komma>
Overskrift	<|> , $| clb <overskrift> <<<
Parentes begynner	<(> , $(<parentes-beg>
Parentes slutter	<)> , $) <parentes-slutt>
Punktum	<.> , $. clb <<< <punkt>
Semikolon	<;> , $; clb <semi>
Spørsmålstegn	<?> , $? clb <spm>
Strek	"<->" , "$-" <strek>
Utrop	"<!>" , "$!" clb <<< <utrop>

VEDLEGG 2:

Sammenlikning av syntaktiske taggsett og noen sentrale analysevalg i Menotec/PROIEL, Talbanken 05 og Copenhagen Dependency Treebank

Taggsett (syntaktiske funksjoner)

Nedenfor er de grunnleggende syntaktiske funksjonene i hvert av korpusene listet opp. Opplysningene er hentet fra prosjektenes manualer, i tillegg har vi i noen tilfeller sett på faktiske annotasjoner i hvert korpus. Vi har i tillegg tatt med de syntaktiske funksjonene som Oslo-Bergen-taggeren opererer med.

Menotec/PROIEL	**Talbanken 05**	**Copenhagen Dependency Treebank**	**Oslo-Bergen-taggeren**
Rotfunksjoner: PRED PARPRED VOC Hjelpefunksjoner: AUX Argumenter: SUB OBJ OBL XOBJ COMP NARG AG Adjunkter: ADV XADV ATR APOS	(j.f. Teleman (1974:37-38)) ES: egentligt subjekt FS: Formellt subjekt SS: övrigt subjekt FV: finit predikatsverb IV: infinit verb SP: bunden subj. pred.-fylln. FP: fri subj. pred.-fylln. OP: obj. pred.-fylln. EO: egentligt objekt FO: formellt objekt IO: indirekt objekt OO: övrigt objekt VO: infinitivfrasen i obj. m. inf. VS: infinitivfrasen i subj. m. inf. RA: rumsadverbial TA: tidsadverbial KA: komparasjonsadverbial OA: objektsadverbial NA: neg.-adverbial	(j.f. CDT-manual: 10-34) SYNCOMP: syntactic complement @space: valency-bound location/direction adverbial @time: valency-bound time adverbial avobj: adverbial object dobj: direct object fobj: filler object gobj: genitive object iobj: indirect object nobj: nominal object numa: additive numeral complement numm: multiplicative numeral complement part: verbal particle pobj: prepositional object possd: possessed complement possr: possessor complement pred: predicative predo: object predicative	@<adv @<det @<p-utfyll @<sbu @<sbu-rel @<subst @adj> @adv @adv> @app @det> @fv @i-obj @interj @iv @kon @laus-np @o-pred @obj @s-pred @subj @subst> @tittel

	VA: varslande adverbial +A: konjunktionellt adverbial MA: talarattitydsadverbial CA: framhävande adverbial AG: agent XA: *så att säga* m. synonymer AA: övrigt adverbial AT: ffrställt adj.-attribut DT: ffrställt bestämmarattribut XT: *så kallad* m. synonymer AN: apposisjon PT: predikativt attribut EF: relativsatsen i emf. omskr. DB: dubbelt satsled ET: övrigt efterställt attribut +F: koordinasjonsfras på satsnivå XF: fundamentfras IM: infinitivmärke PL: verbpartikel PR: preposition ++: samordnande konjunktion UK: underordnande konjunktion XX: obestämbar satsdel Tagger som ikke står i Teleman (1974): VG PA CC (ikke i Talbanken05, men nevnt i Nivre et a.) CJ UA	preds: subject predicative qobj: quotational object robj: reflexive object subj: subject expl: expletive subject vobj: verbal object SYNADJ: syntactic adjunct ADVERB: adverbial app: apposition appa: parenthetic apposition (comma) xpl: explication appr: restrictive apposition (no comma) arg: genitive aributive conj: conjunct relation coord: coordinator relation correl: correlative coordinator relation fpred: free predicative fpredo: free direct-object predicative fpreds: free subject predicative gapd: gapping dependent RuleGap: gapping dependent name: part of name namef: first name namel: last name title: person title pnct: punctuation rel: relative clause relelab: elaborating relative clause relpa: parenthetic relative clause relr: restrictive relative clause voc: vocative	

	IP IK ...	xtop: external topic with resuming pronoun ADVERB: adverbial ATTRIBUTION: inter-sentential elementary discourse unit RuleAr: aribution BACKGROUND: inter-sentential elementary discourse unit CAUSE: inter-sentential elementary discourse unit COMMENT: inter-sentential elementary discourse unit COMPARISON: inter-sentential elementary discourse unit CONDITION: inter-sentential elementary discourse unit CONTRAST: inter-sentential elementary discourse unit ELABORATION: inter-sentential elementary discourse unit ENABLEMENT: inter-sentential elementary discourse unit EVALUATION: inter-sentential elementary discourse unit EXPLANATION: inter-sentential elementary discourse unit MANNER: inter-sentential elementary discourse unit MEANS: inter-sentential elementary discourse unit SUMMARY: inter-	

		sentential elementary discourse unit TEMPORAL: inter-sentential elementary discourse unit agent: agent adverbial aribution: intra-sentential elementary discourse unit baground: intra-sentential elementary discourse unit cause: intra-sentential elementary discourse unit goal: goal adverbial comment: intra-sentential elementary discourse unit comparison: intra-sentential elementary discourse unit conc: concession adverbial concom: cond: condition adverbial condition: intra-sentential elementary discourse unit cons: consequence adverbial contrast: intra-sentential elementary discourse unit elaboration: intra-sentential elementary discourse unit enablement: intra-sentential elementary discourse unit evaluation: intra-sentential elementary discourse unit event: Adverbial	

		expressing an event exem: example adverbial explanation: intra-sentential elementary discourse unit joint: intra-sentential elementary discourse unit man: manner adverbial accom: companionship adverbial inst: instrument adverbial manner: intra-sentential elementary discourse unit means: intra-sentential elementary discourse unit neg: negation adverbial other: other adverbial prg: pragmatic adverbial discmark: sentence-initial discourse marker epi: epistemic adverbial eval: evaluation adverbial focal: focalizer adverbial scene: pragmatic condition and structural adverbial add: additive adverbial contr: contrast adverbial elab: elaboration adverbial quant: degree adverbial resem: comparison adverbial source: source aribution adverbial space: space adverbial	

		dir: direction adverbial loc: location adverbial summary: intra-sentential elementary	

Noen sentrale analysevalg

	Menotec/PROIEL	**Talbanken05**	**Copenhagen Dependency Treebank**
Øverste hode i treet (forhold mellom hovedverb og hjelpeverb)	Hovedverb er øverste hode, har funksjonen PRED. Alle hjelpeverb henges på PRED som AUX.	Finitt verb (hjelpeverb) er hode. Leksikalsk verb er dependent med funksjonen VG (ved f.eks. perfektum) eller SP (ved passiv m. *bli*)	Finitt verb (hjelpeverb) er hode og tar hovedverb (leksikalsk verb) som dependent med funksjonen vobj (verbal object).
Nominalfraser	Nomen er hode, determinativer og adjektiver henges på hodet som ATR.	Nomen er hode. Determinativer og genitivattributter henges på hodet som DT, adjektiver som AT.	Determinativ er hode, nomen er dependent med funksjonen nobj (nominal object). Adjektiver henger som attr på hodet.
Preposisjonsfraser	Preposisjonen er hode, dependenter får funksjonen OBL (COMP ved setningsformede dependenter).	Preposisjonen er hode, dependenter får funksjonen PA, uavhengig om det er et nomen eller en setning.	Preposisjonen er hode, utfyllingen får funksjonen nobj (nominal object).
Koordinering	Konjunksjon er hode. Alle koordinerte ledd har samme type dependens til konjunksjonen som konjunksjonen har til sin modernode. Elementer som deles av alle koordinerte konstituenter, er koblet til konjunksjonen. Koordinerte finitte	Nivre et al. (2006): Første koordinerte ledd er hode for hele det koordinerte uttrykket og bærer funksjonen som hele det koordinerte uttrykket har i setningen. Det andre koordinerte leddet er dependent på det første med funksjonen CC. Konjunksjonen er	Den andre av to konjunkter er dependent på den første med funksjonen conj. Konjunksjoner er dependenter på den andre konjunkten med funksjonen coord.

	setninger analyseres på en egen måte. Hvis konjunksjon mangler, settes tom konjunksjon inn.	dependent på det andre koordinerte leddet med funksjonen ++. I tidligere utgaver av Talbanken har de valgt analysen som er beskrevet i Nivre et al. (2006). I siste versjon av Talbanken er imidlertid konjunksjonen svært ofte hode, inkludert i eksempelsetningen fra Nivre et al. (2006).	
Topikalisering av ledd fra leddsetning, inkl. spørreord. (Eksempel: "*Når* tror du at dere kommer?")	Topikalisert ledd blir værende i leddsetningen.	Vi har funnet to eksempler på denne konstruksjonen i Talbanken og to i Stockholm Umeå Corpus (SUC) (som sammen med Talbanken utgjør Swedish treebank). I eksemplene fra Talbanken blir de topikaliserte leddene værende i leddsetningen, men i eksemplene fra SUC er de topikaliserte leddene dependenter på setningens øverste hode.	?
Leddsetninger inkl. relativsetninger	Leddsetninger som fungerer som subjekt eller objekt, analyseres som COMP (altså ikke SUB eller OBJ). Ellers kan leddsetninger være f. eks. ADV, ATR (restriktive relativsetninger) eller APOS (ikke-restriktive relativsetninger).	Leddsetninger som fungerer som subjekt, får en SS-dependens, mens objektsleddsetninger får OO. Det er de samme funksjonene som nominale argumenter får. Finitt verb er hode (ikke	Det kommer ikke helt tydelig fram om leddsetninger som er subjekt/objekt får denne funksjonen eller en egen, som i Menotec/PROIEL. Subjunksjonen er hode.

	Subjunksjon er hode. Hvis subjunksjon mangler, rykker hovedverbet opp og overtar funksjonen.	subjunksjonen, altså.) Relativsetninger: Det finitte verbet er hode. *Som* tas for å være et relativpronomen, dependent på relativsetningens leksikalske verb.	Relativsetninger: Finitt verb i relativsetningen regnes som hode, er dependent til relativisert ledd i oversetningen med funksjonen rel. Hvis det finnes et relativpronomen, får det en såkalt *ref*-pil fra hodet i det relativiserte leddet i oversetningen. Hvis det ikke finnes noe relativpronomen, må hodet i det relativiserte leddet fungere som en såkalt sekundær dependent til et ord i relativsetningen, ofte verbet. (Relativsetninger blir altså ikke innledet av subjunksjon etter denne analysen.)
Tomme noder?	Ja, verb og konjunksjoner. Tomme verb brukes i setninger hvor verbet mangler, f. eks. ”Han spiste is og hun kaker”	?	?
Skilletegn	Sorteres ut, er ikke en del av dependenstreet.	Punktum er dependent på det øverste hodet med funksjonen IP. Komma får funksjonen IK. Når komma brukes i forbindelse med koordinering av flere ledd, er kommaet dependent på konjunksjonen. Når komma brukes til å vise	Skilletegn får funksjonen pnct. Punktum er dependent til hovedsetningens verb. Komma som skiller ut bisetninger, er dependent på bisetningens verb. Ved koordinering av flere konstituenter er

		slutten på en relativsetning eller en foranstilt leddsetning, er det dependent på (det leksikalske) verbet til den setningen det markerer slutten på.	kommaer, i likhet med konjunksjonen og de andre koordinerte elementene, dependent på første koordinerte hode.
Komparasjon	Menotec: *En* 'enn' og *sem* 'som' regnes som subjunksjoner, og sammenlikningsledd blir dermed alltid leddsetninger, eventuelt med tomt verb. PROIEL: Komparasjonsordet (*quam*, 'enn' etc.) er datter av det komparative adjektivet/adverbet via en OBL-dependens. Det andre elementet i sammenlikningen er datter av komparasjonsordet. Dependensen er den samme som det første elementet har i sammenlikningen. I en setning som *Mari har større bil enn John*, vil *John* være knyttet til *enn* med en SUBJ-dependens, fordi *Kari* er subjekt. I *Bilen traff Mari hardere enn John* vil *John* ha en OBJ-dependens.	Vi har sjekket *än*. *Än* regnes som en subjunksjon. Den er gjerne dependent på (det leksikalske) verbet, og altså ikke på det komparative adverbet eller adjektivet (men det ser også ut til å finnes andre alternativer). *Än* får funksjonen KA. Det sammenliknede ordet knyttes til *än* med en UA-dependens. UA brukes ellers for å knytte det finitte verbet i en leddsetning til leddsetningens subjunksjon.	Vi har sjekket *end*. *End* regnes som en subjunksjon. Den er dependent på det komparative adjektive/adverbet med funksjonen pobj. Funksjonen til datteren til end varierer (er datteren et nomen, får den ofte nobj, adverber får obl, verb vobj etc.)
Kontrollsetninger o.l.	XOBJ med en sekundærdependens (XSUBJ) til subjektet. Predikativ annoteres også som XOBJ med	I setninger av typen *man prövar att...* er infinitivsmerket hode for infinitivskonstruksjonen,	I setninger av typen *man prøver at...* er infinitivsmerket hode for infinitivs-konstruksjonen,

	sekundærdependens.	knyttet til det finitte verbet med en OO-dependens, dvs. den brukt ved nominale objekter. Infinitiven er knyttet til infinitivsmerket med en IF-dependens. Ved modale hjelpeverb er infinitiven dependent på hjelpeverbet. Funksjonen er VG.	knyttet til det finitte verbet med en dobj-dependens (direct object). Infinitiven er knyttet til infinitivsmerket med en vobj-dependens. Ved modale hjelpeverb er infinitiven dependent på hjelpeverbet. Funksjonen er vobj.
Funksjonen til formelle subjekter og potensielle subjekter ("logiske subjekter"/"egentlige subjekter") (dvs. hvilken funksjon får *det* og *en mann* i *Det sto en mann i hagen*)	Ikke relevant for PROIEL, heller ikke for Menotec. Men i det gammelengelske korpuset på ISWOC (samme applikasjon som PROIEL og Menotec) får formelle subjekter funksjonen EXPL. Når det potensielle subjektet er en leddsetning, blir det COMP, hvis det er en NP, blir det SUB.	Formelt subjekt får FS, potensielt subjekt får ES.	Formelt subjekt får funksjonen expl, potensielt subjekt får forskjellige funksjoner, alt etter konstruksjonen.

Kilder:

Nivre, Joakim, Jens Nilsson, og Johan Hall. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. *Proceedings of LREC*. Genova: 1392–1395.

Teleman, Ulf. 1974. *Manual för grammatisk beskrivning av talad og skriven svenska*. Sverige: Studentlitteratur.

Menotecs retningslinjer: http://dl.dropbox.com/u/8086804/Menotec/Menotec-retningslinjer-2011-09-02.pdf

PROIELS retningslinjer: http://folk.uio.no/daghaug/syntactic_guidelines.pdf

VEDLEGG 3:

INESS – C-STRUKTUR

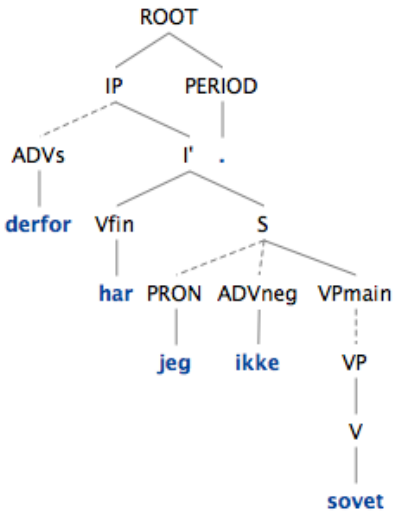
Se: http://iness.uib.no/iness/main-page

Og

http://prosjekt.digital.uni.no/projects/inesspublic/wiki/NorGram_Phrase_Structure_Categories

Først: enkelt sitat fra Iness med tre

Vfin is the head of the **I'** phrase. The reason why the category name **I** is not used, is that finite verbs also occur as daughters of **VPfin**



2. Inventory of phrase structure categories

1conj10P	CPtmprel	IPimprs	NUMlitcoord	PPpart	QuantPmin	
ADVP	DA'	IntADVcoord	NUMmilliardP	PPpost	QuantPnoref	
ADVPdeg	DAP	IntNOMcoord	NUMmillionP	PPppobj	QuantPprpstn	
ADVPdegloc	DAPgen	MeasP	Ncoord	PPprd	ROOT	
ADVPdegnum	DP	MeasPgen	ORDP	PPpred	ROOTconj	
ADVPint	DPgen	NOM'	ORDPgen	PPsel1	ROOTconj2	
ADVPloc	DPint	NOMcoord	P'clk	PPsel2	S	
ASVPloctop	DPintgen	NOMcoordgen	PAREN	PPsom	Scop	
ADVPs	DescP	NP	PARpp	PPtil	Simprs	VPpart
ALLQP	FRAGMENTSTOP	NPcoordgen	POSS'	PPvbobj	Ssub	VRBcoord
ALLQPgen	HALF'	NPgen	POSS'gen	PRONP	Ssub2	VRBfincoord
AP	HOUR'	NPint	POSSP	PROP'	Ssubcoord	WH-ELL
APcoord	HOURP	NPsit	POSSPgen	PROPP	TTLP	WhP
APint	I'	NPssn	POSSPint	PROPPgen	TofD	WhPcoord
AppP	I'aux	NPtemp	POSSPintgen	ParP	VP	WhPfree
CMPNDcoord	I'coord	NUM1000P	POSSPrel	QP	VP2	WhPrel
CONJPcomp	I'cop	NUM100P	PP	QPgen	VP2fin	YEARP
CPadv	I'imprs	NUM10P	PPADVcoord	QPint	VPasp	
CPinf	I'inq	NUM11P	PPapred	QPintgen	VPaspfin	
CPloc	INIT'	NUM1P	PPav	QTENOM	VPattr	
CPnom	IP	NUM21P	PPavint	QuantP	VPcoord	
CPnullc	IPcond	NUMBP	PPclk	QuantPadv	VPfin	
CPpol	IPcoord	NUMP	PPcoord	QuantPgen	VPfincoord	
CPpur	IPcoordgap	NUMdigP	PPint	QuantPint	VPmain	
CPrel	IPgap	NUMdigcoord	PPnum	QuantPintgen	VPmain2	

Noen sentrale analysevalg:

Øverste hode i treet (forhold mellom hovedverb og hjelpeverb)	Root. Har datter IP, som har DP i Spec og Vfin som hode. Kompl til Vfin kan være S (dvs. infinitt verbalfrase). Se fig. 1. Altså: Vfin/I er hode.
Nominalfraser	Det funksjonelle elementet er hode: D, Quant (Num). NP er komplement. Se fig 2 og 3.
Preposisjonsfraser	P er hode, NP komplement. Se fig. 1,2,3.
Koordinering	Conj er hode, konjunktene er likeverdige. Men NB frasen er også nevnt med en kategori: APcoord. Se fig. 4. Setningskoordinering gjøres på samme måte, se fig. 5 Gapping: IPCoord. Med structure sharing. " **IPcoordgap** is a coordination of IPs like **IPcoord** , except that the second conjunct is 'gapped', i.e., it lacks a finite verb, which is understood to be the same as the finite verb of the first conjunct. This is presently restricted to predicative sentences with copula *være*: *Petter er sint og Kari fortvilet.* "
Topikalisering av ledd fra	C-str følger overflaterekkefølgen. F-strukturen ordner

leddsetning, inkl. spørreord. (Eksempel: ”*Når* tror du at dere kommer?”)	med de semantiske relasjonene. Fig. 6a+b
Leddsetninger inkl. relativsetninger	Subjunksjonen er hode. Fig. 7. Relativsubjunksjonen er hode i relativsetning: CRel i CPrel., fig.8a,b. Rel.setning med null subjunksjon: CPnullc: Vet ikke hva som er hode her, fig. 8b.
Tomme noder?	Har ikke sett noen.
Skilletegn	Er med. fig.9.
Komparasjon	enn: CONJcomp fig. 10a,b.Forstår ikke helt setn.komparasjon.
Kontrollsetninger o.l.	**Se fig.11.**
Funksjonen til ekspletive subjekter og “logiske subjekter” (dvs. hvilken funksjon får *de*t og *en mann* i *Det sto en mann i hagen*)	