Syntaktisk informert frasesamanstilling

Kevin Brubeck Unhammer

07/09, 2010

Innhald

Kapittel 1

Krav til frasesamanstilling

1.1 Innleiing

I denne delen prøver eg å finne fram til kva som er den best moglege frasesamanstillinga. Eg argumenterer for at «best» her må tolkast i forhold til eit formål, her å finne samsvar mellom kasusmarkering og semantisk rolletildeling. Som utgangspunkt har eg visse krav for ordsamanstilling gitt i Thunes (2003), saman med krava for frasesamanstilling i Dyvik et al. (2009). Eg viser kvifor ein, for våre formål, må revidere kravet til Thunes om likskap i argumentstruktur. Eg gir nokre døme for å grunngje krava i Dyvik et al. (2009), i tillegg til å utdjupe dei for å gjere dei enklare å implementere i kapittel **??**. Dette involverer au å omformulere krava for c-struktursamanstilling slik at dei ikkje refererer til ordlenkjer, berre f-strukturlenkjer. Sidan eit av måla med Xpar-prosjektet er å finne ut kor mykje frasesamanstillingsinformasjon me kan få ut av parallellismen i f-strukturane (eller, sett frå den andre sida, kor uavhengig ein kan gjere seg av den bottom-up-informasjonen ei ordlenkje gir), blir det eit avleidd mål å formulere frasesamanstillingskrava med referanse til f-strukturane der det går an.

1.2 Formål med frasesamanstilling

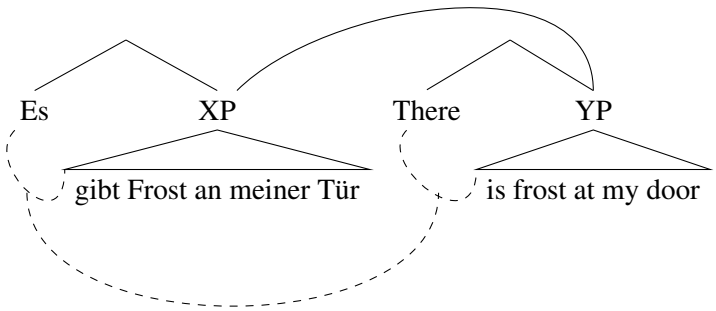
Ei frasesamanstilling er ein slag annotasjon av eit korpus. På same måte som oppbygginga av eit korpus avheng av formålet til korpuset, kan ein ikkje definere den ideelle annotasjonen av eit korpus utan å ta høgd for kva ein skal nytte annotasjonen til.

Me kan illustrere dette med eit enkelt, praktisk døme: ved automatisk ordklassetagging må ein gjerne avvege mellom dekning (å finne flest moglege analysar for flest mogleg ord) og presisjon (å berre ende opp med korrekte analysar). Viss formålet er å annotere ein leksikografisk ressurs, vil det vere viktigare med høg dekning på bekostning av presisjon, sidan leksikografen gjerne leiter etter nye/kreative bruksområde av ord. Skal taggaren nyttast til maskinomsetjing i staden, kan ein ikkje nytte meir enn éin analyse til slutt, så her er presisjon viktigast.

Sjølvsagt kan ein her seie at den *ideelle* annotasjonen vil vere å berre ha korrekte analysar, men sjølv ved ideelle krav er formålet viktig: er ein ute etter å finne N-gram som ofte blir omsett med kvarande, men som *ikkje* er syntaktiske konstituentar, er det klart at retningslinjene nedanfor ikkje er så nyttige.

Sidan utviklinga av automatisk frasesamanstilling hovudsakleg har skjedd innanfor frasebasert statistisk maskinomsetjing (PBSMT), kjem me ikkje utanom ei samanlikning her. I PBSMT er formålet med ei fraselenkje å betre maskinomsetjing på eitt eller anna mål, t.d. BLEU-skåren. BLEU-skåren samanliknar ferdig omsett tekst (ein gullstandard) med det automatisk omsette, ved å sjekke kor mykje N-gram-overlapp det er mellom tekstene. Ei fraselenkje mellom N-grammet *es gibt* og *there is* (dvs. eit auka sannsyn for å nytte slike par i omsetjinga) kan gi ein høgare endeleg skåre i BLEU. Som vist i Koehn et al. (2003) fekk dei ein lågare BLEU-skåre når dei fjerna lenkjer mellom nodar som, i følgje ein robust statistisk PCFG-parsar, ikkje var syntaktiske frasar (konstituentar). Dvs. at i figur 1.1 vil lenkja vist ved den prikkete linja bli fjerna frå mengda over moglege lenkjer om ein berre held seg til syntaktiske konstituentar, og $p(es\ gibt,\ there\ is)$ vil ikkje bli tilsvarande auka i den statistiske omsetjingsmodellen. Sidan PBSMT, som skildra i Koehn et al. (2003), er agnostisk til syntaktiske høve i omsetjingssteget[1] er det for dei ingen grunn til å berre halde seg til samanstilling mellom syntaktiske konstituentar; dei har i utgangspunktet meir nytte av kollokasjonsinformasjon.

todo: referere til den faktiske parsaren? det var Bikel kanskje?



Figur 1.1: N-gram-samanstilling versus syntaktiske frasar

Men sett no at me ikkje har som formål å nytte frasesamanstillinga til reint N-grambasert omsetjing. Kva for *lingvistiske* krav kan me stille til å kalle to frasar samanstilte? Me må i alle fall tillate ein del skilnad. I alle større parallelltekster vil parallellstilte setningar ha visse syntaktiske og semantiske[2] omsetjingsskifte, t.d.

[1] Både omsetjingsmodellen og språkmodellane er reint N-grambaserte her, og har difor ikkje nytte av syntaktisk informasjon (i motsetning til syntaktisk informert generering slik Riezler & Maxwell (2006) implementerer).

[2] Sidan eg går ut frå at data er setningssamanstilt, kjem eg ikkje inn på diskurs-/pragmatiske verknader, med mindre dette fører til forskjellar innanfor setningane (sjå t.d. del 1.5 om lenkjer mellom koreferente substantiv og pronomen).

leksikalisering av syntaktiske konstruksjonar eller omvendt, endring av ordklasse, presisering/depresisering, endringar i leksikalske trekk (t.d. telleleg/utelleleg), osb. (Munday, 2001, s. 56–62), slik at den einaste fullstendige, «perfekte» samanstillinga vil vere identitetsfunksjonen. Kor mykje mangel på samsvar me godtek blir då avgjort av formålet med samanstillinga.

Eitt av formåla med samanstillinga i denne oppgåva er å kunne oppdage korleis ulike språk realiserer semantiske roller syntaktisk; då spesielt i forhold til hypotesane gitt i XPar (2008, s. 7), t.d. at «case marking might be useful to further determine a given argument's semantic role». Skal me finne det siste, må me altså kunne lenkje frasar med ulik kasusmarkering, men ha krav om lik tildeling av semantiske roller; samtidig skal me sjå at me ikkje kan ha krav om lik syntaktisk funksjon. I tillegg vil me sjølvsagt ikkje lenkje på tvers av konstituentgrenser, sidan det er fullstendige konstituentar[3] som fyller dei semantiske rollene.

Eit anna mogleg formål er å nytte desse frasesamanstillingane til maskinomsetjing. Riezler & Maxwell (2006) nyttar ein stokastisk frasesamanstilling til å oppdage transfer-reglar for bruk i LFG-basert generering i maskinomsetjing. Dette er reglar som omsett fragment av ein f-struktur på kjeldespråket til f-strukturfragment på målspråket. (Eit krav på utforminga av moglege transfer-reglar hindrar at ein får reglar som lenkjar ikkje-konstituentar, eg kjem tilbake til dette nedanfor.) Samanstillinga utvikla her burde au kunne nyttast til å finne slike transfer-reglar, men dette er ikkje noko eg har lagt vekt på.

Nedanfor gir eg eit forslag til krav for frasesamanstilling, med desse formåla i tankane. Om alle krava er moglege å implementere, er eit separat problem.

1.3 Frasesamanstilling i ein LFG-trebank

Samanstilte frasar bør ha nok semantisk likskap til å kunne opptre som omsetjingar i liknande omgivnader (Dyvik et al., 2009, s. 74). Thunes (2003) gir nokre prinsipp – som er passande å ha som utgangspunkt – for å fastslå det som kan kallast *omsetjingsmessig korrespondanse* (her for ordsamanstilling). Dette er prinsipp som skal gjelde for eit litt forskjellig formål, men som au «ligger nær opp til det vi intuitivt mener er riktig» (Thunes, 2003, s. 2). Prinsippa blir nytta til å lage ein gullstandard for ordsamanstilling[4], hovudsakleg for dei opne klassene, og er definert ved å vise til kva for rolle eit argumentord speler, eller kva for rolletildeling eit predikat eller modifiserande ord gir. Så for å t.d. samanstille to verb må dei ha like mange semantiske argument (men argumenta treng ikkje alle realiserast syntaktisk) og dei

[3] LFG tillèt som nemnt diskontinuerlege konstituentar, men dette er ikkje det same som ikkje-konstituentar av typen «es gibt» / «there is».

[4] (Thunes, 2003, s. 2): «Våre prinsipper er satt opp for å tjene et bestemt formål, nemlig å samle inn data som metoden i Semantic Mirrors skal anvendes på», ein metode for å automatisk finne WordNet-liknande relasjonar frå parallelltekst. I denne metoden vil det vere naturleg med høge krav til presisjon, men kanskje lågare krav til dekning: speilmetoden skal finne leksikalske semantiske forhold som held på *typenivå*, medan for trebanken er det viktigare korleis me kan annotere eit *token* av t.d. eit verb i ein viss VP i ei gitt korpussetning.

må *tildele same roller*; medan argumenta må *spele same rolle*, og både argument og adjunkt må vere *koreferente*. Lenkja ord må vere del av frasar som speler same rolle i «det som er felles i interpretasjonene av [dei to setningane]» (Thunes, 2003, s. 3).

Viss me tek utgangspunkt i det siste, vil det vere naturleg å i tillegg lenkje desse frasane som speler same rolle i «det som er felles i interpretasjonene».

Krava for ordsamanstillinga må au vere fylt for at desse frasane kan samanstillast. Ei ordsamanstilling er altså naudsynt for ein frasesamanstilling, og omvendt. Dette er berre problematisk om me føreset at det eine er derivert av det andre; men dette har me ingen *a priori* grunn til å gjere. Krava eg her utviklar bør i staden sjåast på som *skrankar* på moglege samanstillingar i modellen (jamfør **??** om modellteoretiske grammatikkar), heller enn derivasjonelle forhold. Samtidig er det som nemnt eit mål å finne ut kor uavhengig me kan gjere oss av ordlenkjingsinformasjonen (dette er au nyttig for implementasjonen), utan at det treng å gi krava ei *retning*.

Ei frasesamanstilling er ei skildring av forhold mellom *fragment* av setningar, dette er endå ein grunn til at det er naturleg å skildre dei ønskelege forholda som skrankar på moglege samanstillingar. Me kan setje skrankar på f-struktur-, konstituent- og ordsamanstilling samtidig, utan å måtte ha krav om at den eine samanstillinga er fullstendig (eller delvis) avleiia av den andre, før me veit om eit slikt avleiingsforhold er empirisk fundert. Me kan i tillegg ha ufullstendige samanstillingar i dei tilfella der det er ufullstendig samsvar mellom setningane (der ei fullstendig samanstilling ville brutt visse krav).

Sidan metoden er mynta på bruk i ein LFG-parsa trebank, og delvis vil nytte denne annotasjonen som datagrunnlag, er det naturleg å nytte same konsept som blir nytta i LFG[5] (f-struktur, c-struktur, endosentrisitetsprinsipp, $\overline{\text{X}}$-tre, osb.) au i desse krava til den «beste» frasesamanstillinga; i den grad LFG gir ein generaliserbar skildring av syntaks, bør desse krava vere generaliserbare til andre teoriar, men ein del forhold som er avleidd av LFG-prinsipp må sjølvsagt modifiserast om krava skal generaliserast til andre teoriar.

Utan skrankar i det heile vil alt kunne lenkjast til alt (noko som er like unyttig som å ikkje lenkje noko); i del 1.4 ser eg på kva for typar element i dei lingvistiske analysane (ord, grammatiske trekk, konstituentar, ...) det er fornuftig å tillate lenkjer mellom. I avsnitta nedanfor spesifiserer eg kva som må til for at me skal lenkje element av desse typane.

[5] I tillegg finst andre positive biverknader av ein LFG-basert frasesamanstilling for bruk i denne samanhengen, som at ein kan studere kor parallelle dei parallelle grammatikkane i ParGram-prosjektet (Butt et al., 2002) faktisk er, på ulike nivå (leksikon og argumentstruktur, c-struktur, f-struktur).

1.4 Kva kan lenkjast?

Viss to uttrykk er samanstilt på setningsnivå (slik at me dimed kan gå ut frå at dei er omsetjingar av kvarandre), og begge har ein LFG-analyse, så har me iallfall tre ulike nivå kor me kan finne ekvivalensforhold under setningsnivå:

1. mellom ord i setningane,
2. mellom f-strukturar,
3. mellom c-strukturnodar.

På begge språk har me alle nivå – det er ingen grunn til å lenkje på tvers av nivå sidan forhold mellom desse nivåa er implisitt i LFG-analysen.

Alle ord i setninga er *kandidatar* for samanstilling med ord i omsetjinga, men det kan godt hende at eit ord *ikkje* har ei lenkje, og me kan heller ikkje utelukke at det finst mange-til-mange-lenkjer som ikkje kan «delast opp». Dette gjeld au nodane i c-strukturen.

Me utelukker lenkjing av ikkje-konstituentar som *there is* på c-strukturnivå sidan ei lenkje mellom to c-strukturnodar impliserer at heile frasen under er lenkja. Det finst ingen c-strukturnodar som dominerer berre *there*, *is* og ingen andre ord (heller ikkje *es*, *gibt*), så dette er ikkje lenkjekandidatar. *There is* og *Es gibt* i figur 1.1 kan då ikkje samanstillast åleine, men berre som del av ei ytre frasesamanstilling[6].

Når det gjeld f-strukturane er det ganske mange element me teoretisk sett kunne ha lenkja, t.d. enkelttrekk som kasus eller dei uordna mengdene med adjunkt, men det som er mest *nyttig* og *meiningsfullt* er nok å berre lenkje der det er ei nær kopling til orda i setninga. Sidan alle PRED-element i ein f-struktur unikt står for predikerande ord, kan me – gitt to samanstilte setningar – la *kandidatane for samanstilling på f-strukturnivå* inkludere alle desse PRED-elementa i f-strukturane til setningane[7]. PRED-element representerer semantiske bidrag som oftast er påkrevde på begge språk i omsetjingar, medan andre f-strukturtrekk gjerne er valfrie på det eine av språka; det er ikkje alle språk som har t.d. obligatorisk kasusmarkering, og ein vil kanskje nytte trebanken til å oppdage nettopp slik variasjon. PRED-elementa er i tillegg gjerne enklare å knyte direkte opp mot den konkrete, observerte tekststrengen (eventuelt testast mot korpora, eller talarintuisjonar), medan t.d. eit trekk som aspekt kanskje er umogleg å skilje frå tempus i affikset (det vil vere vanskelegare å teste om ei lenkje mellom aspekt-trekk er empirisk motivert utan å dra inn ein heil del teori).

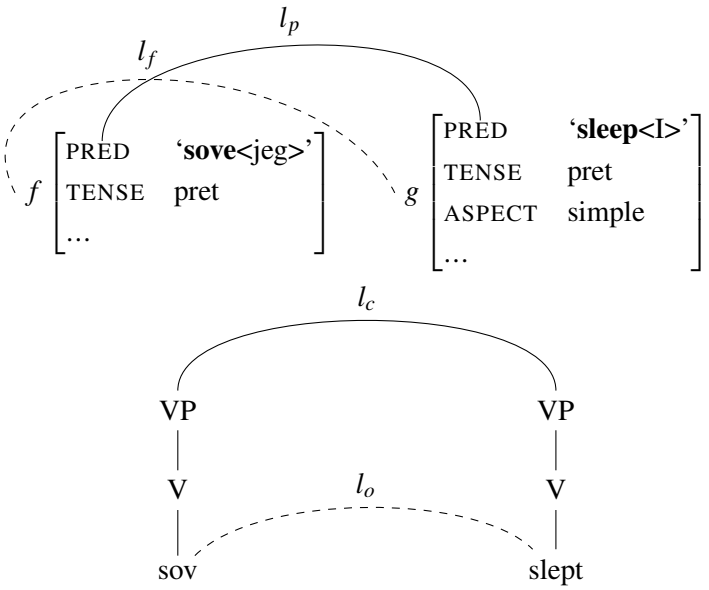
Samtidig er det au eit omsetjingsforhold mellom trekka i same f-struktur som dei lenkja PRED-elementa, og me ville kanskje ikkje ha omsett dei to PRED-elementa i andre f-strukturkontekstar. Difor bør me au sjå på ei PRED-lenkje som

[6] Slike forhold kan me sjølvsagt finne igjen etter lenkjinga, men då vil me au kunne generalisere til andre ordformer. Eg kjem tilbake til dette i kapittel **??**.

[7] I del 1.7.3 kjem eg tilbake til spørsmålet om me vil inkludere visse f-strukturar utan PRED-element i kandidatane for samanstilling.

ei lenkje mellom *f-strukturane til desse PRED-elementa*[8]. Med dette i tankane, kombinert med c-struktur-f-strukturavbildinga ϕ (sjå del **??**), får me følgjande samanheng, illustrert i figur 1.2:

(1) Ei lenkje mellom to PRED-element p og q, kor p er medlem av f-strukturen f, og q er medlem av f-strukturen g, tilseier at:
   a. me tolkar f-strukturane f og g som lenkja,
   b. orda i setningane som projiserer PRED-elementa tek del i ei lenkje (kor andre ord kan vere involvert), og at
   c. nodar innanfor $\phi^{-1}(f)$ og $\phi^{-1}(g)$, dei funksjonelle domena til f-strukturane f og g, kan lenkjast



TODO: teikne inn f-domene

Figur 1.2: Ei PRED-lenkje l_p kan tolkast som ei f-strukturlenkje l_f, og impliserer ei c-strukturlenkje l_c mellom toppnodane i dei funksjonelle domena. Orda som projiserer PRED-elementa er med i ei lenkje l_o (som kan inkludere fleire ord).

Punkt (1-a) og (1-c) over seier at viss PRED-elementa projisert av t.d. to verb i verbfrasar er lenkja, kan VP-ane som heilskap lenkjast, i tilfellet i figur 1.2 kan iallfall dei øvste nodane i VP-ane lenkjast, i tillegg til f-strukturane frå ytre PRED til verba. Det er dette at heile VP-ane (kanskje inkludert objekt) er lenkja som gjer det til ei fraselenkje og ikkje berre ei ordlenkje. Punkt (1-a) er forsvart over, medan punkt (1-c) kjem som ein konsekvens av at det er det funksjonelle domenet som spesifiserer informasjonen i f-strukturane, nodane her bør difor lenkjast berre viss

[8] Eventuelt kunne me ha definert lenkjingskandidatane på f-strukturnivå som alle PRED-haldande f-strukturar, resultatet blir det same.

f-strukturane er lenkja. Men som punkt (1-c) indikerer finst det au situasjonar der nodar innanfor domena skal stå ulenkja.

Alle nodar i c-strukturen (alle syntaktiske *frasar/konstituentar* i setninga) som kan koplast til PRED-haldande f-strukturar, vil vere kandidatar for samanstilling på c-strukturnivå (dette inkluderer diskontinuerlege konstituentar), men ikkje alle vil bli lenkja. I del 1.7 ser eg på kva som må til for å lenkje nodar i det funksjonelle domenet. I tillegg finst det nodar over ord som ikkje projiserer PRED-element, desse kjem eg tilbake til i del 1.7.3.

I følgje punkt (1-b) vil fraselenkja leie til at sjølve verba i to lenkja VP-ar au er lenkja, som tilseier at *ei PRED-lenkje impliserer ei ordlenkje*. I visse tilfelle er dette heilt uproblematisk, t.d. viss *I slept down by the river* skal lenkjast med *Eg sov nede med elva* vil me uansett lenkje *slept* og *sov*; dette kan gjelde transitive verb au:

(2) a. The locusts have no king, just noise and hard language
↔
b. Grashoppene har ingen konge, berre støy og krasse ord

have/har tek del i VP-samanstillinga *have no king.../har ingen konge...*, her au skal det vere uproblematisk å lenkje enkeltorda *have* og *har*.

Men som nemnd treng ikkje ordsamanstillinga vere ein-til-ein, det punkt (1-b) seier er at desse orda iallfall er ein del av ein samanstilling med kvarandre (i døme (2) altså VP-samanstillinga). Kanskje er dette ei mange-til-mange-lenkje som ikkje *kan* reduserast til ein-til-ein-lenkjer; eller kanskje er det som i (2) mogleg å skilje ut delsamanstillingar, som *have/har*. Eg kjem tilbake til dette TODO: når? seinare.

Sidan PRED-lenkjing impliserer ordlenkjing, må me sjekke om krava på ordnivå (del 1.5) er oppfylte for å lenkje to PRED-element. TODO: litt brå avslutning

1.5 Krav på ordnivå

Ord som skal lenkjast må i Thunes (2003) vere del av frasar som speler same rolle i det som er felles i interpretasjonane, her kan me omskrive det til at dei må vere del av *frasar som er lenkja på c-strukturnivå*; forholda i (1) gir då koplinga til krav på andre nivå (t.d. vil krav om tildeling av like mange roller vere meir passande å spesifisere på f-strukturnivå).

Det er visse ting me ikkje kan spesifisere ut frå rein c- og f-strukturinformasjon. Den norske setninga *eg vil ete* kan fint samanstillast med *I want to eat*, med ei lenkje mellom *ete* og *eat*. Men kva står i vegen for å lenkje *ete* til hovudverbet i *I want to drink*? Forskjellen på f-strukturnivå er berre at PRED-verdien er ulik (**eat** mot **drink**). Me må altså ha eit krav om at tydinga til lenkja ord (og deira predikat) er «lik nok» til at me kan sjå på dei som omsetjingar[9]. Dyvik et al. (2009, s. 74) krev at orda generelt, utan kontekst, må vere semantisk plausible omsetjingar, dvs.

[9]Eigentleg burde slike setningar ikkje vere lenkja på setningsnivå ein gong, men som me skal sjå i del 1.6.1 treng me kravet om lik tyding sjølv innanfor setninga.

at målordet er eit medlem av mengda av *linguistically predictable translations* av kjeldeordet. Målordet har då *LPT-korrespondanse* med kjeldeordet. Nedanfor reknar eg LPT-kravet som eit krav på ordnivå, og eg føreset at LPT-informasjonen er ein type bottom-up-informasjon, som viser om to ord generelt (i ulike kontekstar) blir nytta som omsetjingar av kvarandre. Denne informasjonen kan reint praktisk komme frå automatisk ordsamanstilling, eller ei god tospråkleg ordbok, det bør ikkje spele nokon rolle for resten av krava[10].

TODO: Er det mogleg å presisere LPT-kravet meir? Skal det berre vere eit rangerings-krav??

Ein type presisering/depresisering (del 1.2) me ofte ser i omsetjingar er at eit pronomen på kjeldespråket blir nytta der målspråket har eit koreferent substantiv, eller omvendt. Dyvik et al. (2009) opnar for at desse au har LPT-korrespondanse (som nemnt i Thunes (2003) må lenkja ord uansett vere koreferente).

Men kva då med lenkjing av pronomen til verb bøygd for person og tal i pro-drop-språk?

(3) a. iqePa (georgisk)
$\leftrightarrow$
b. han bjeffa

Viss setningane i døme (3) er lenkja, der iqePa har eit pro-argument koreferent med *han* som subjekt, bør dei to subjekta iallfall kunne lenkjast på f-strukturnivå; dei har same referent og speler same rolle i argumentstrukturen til verba (som me går ut frå er lenkja). På ordnivå, derimot, kan me ikkje lenkje *han* til *iqePa* åleine – her må me ha ei mange-til-ein-lenkje mellom $\{\text{han}, \text{bjeffa}\}$ og $\{\text{iqePa}\}$. Generelt må me ha slike lenkjer der eitt ord projiserer fleire PRED-element[11].

1.5.1 Ordklasse

Ulike språk leksikaliserer same konsept på ulike måtar. Cheung et al. (2002, s. 3) nemnar vanskane med å ha eit krav om lik ordklasse i utviklinga av ein kinesisk-engelsk termbank, kor t.d. det engelske ordet *fulfilment* meir naturleg blir omsett til eit verb på kinesisk. På same måte vil eit georgisk verbalsubstantiv (*masdar*) gjerne bli omsett til eit verb i infinitiv på norsk. Slike skifte mellom ordklasser er svært vanlege i omsetjing[12].

Me kan opne for ordklasseoverskridande lenkjer der det er samsvar på andre nivå, me bør iallfall krevje ein likskap i argumentstruktur; så om LPT-kravet og krava på c- og f-strukturnivå er fylt, bør det ikkje vere noko i vegen for å lenkje ord (eventuelt mengder av ord) av ulik ordklasse.

[10] Ein kan au tenkje seg at ei djup semantisk dekomponering av kvart ord sto som grunnlag for LPT-informasjon – men då vil LPT-korrespondanse mellom to ord implisere at orda er synonyme, heller enn generelt plausible omsetjingar.

[11] Me ville au fått ei mange-til-ein-lenkje om me tillot *komplekse predikat* i analysane, t.d. slik Butt (1998) foreslår ved å la kombinasjonen av to ord endre argumentstrukturen til eitt PRED-element.

[12] Munday (Catford (1965), i 2001, s. 61) gir ein gjennomgang av slike *klasseskifte*, og andre typar omsetjingsskifte.

1.6 Krav på f-strukturnivå

problematiser: må me ha /PRED-lenkje/ frå arg til arg/adj? (ikkje krevd no...men skal me rangere ved det?)

kryssande f-lenkjer? mange-mange-f-lenkjer?

På f-strukturnivå har me direkte tilgang til informasjon om argumentstrukturen til eit predikat, og mengda av adjunkt som modifiserer predikatet. Når Thunes (2003, s. 3) skriv at to lenkja ord *a* og *b* må opptre i frasar som har «tilstrekkelig like argumentstrukturer til at uttrykkene i *a*s omgivelser står i de samme semantiske relasjonene til hverandre og til *a* som de korresponderende uttrykkene i *b*s omgivelser gjør til hverandre og til *b*» er det difor passande å prøve å gjere dette til eit krav på f-strukturnivå.

Den enklaste lenkjingssituasjonen, f-strukturmessig, er der rotpredikata kan lenkjast, og første argument av predikatet på kjeldespråket kan lenkjast til første argument på målspråket, andre argument til andre argument, osb., og lenkjinga kan fortsetje slik rekursivt inn i f-strukturane. I ein slik situasjon er det fullstendig samsvar mellom kor mange argument det finst på kvar side, og fullstendig samsvar i det tematiske rollehierarkiet (dvs. kva for posisjon kvar rolle har i argumentstrukturen), i heile strukturen.

Som me skal sjå er det ikkje vanskeleg å komme over situasjonar der dette ikkje held, og me blir nøydt til å tillate lenkjer mellom argument og adjunkt, og lenkjer som går på tvers av følgja i argumentstrukturane. I tillegg kan me ikkje klare oss utan LPT-informasjon for å avgjere *når* me har å gjere med slike meir komplekse situasjonar.

1.6.1 Krav om lik argumentstruktur

Thunes (2003) gir som nemnd eit krav om at *predikat må ha tilsvarande semantiske argument* for å lenkjast.

Om det alltid er slik at to predikat har like mange argument, som kjem i same rekkjefølgje i argumentstrukturen, vil det gjere den praktiske oppgåva med å lenkje predikata, og argument med argument, mykje enklare. Men kan me stille så sterke krav?

Sett at ei setning på språk 1 har ei *at*-setning som adjunkt, medan denne setninga på språk 2 er eit argument, og at desse setningane ville vore lenkja om dei opptredde åleine. Om dei uttrykkjer same proposisjon og *speler same rolle i verbsituasjonen*, synest det naturleg å lenkje desse.

Slike omsetjingsrelasjonar gir data for verbsituasjonen, på eit meir generelt grunnlag enn det me kan få frå einspråklege analysar åleine. Om me har gode semantiske grunnar for å kalle ein deltakar i ein verbsituasjon eit argument på eitt språk, vil dei same grunnane gjelde for omsetjingsmessig korresponderande verb på andre språk. Ein kan då nytte unionen over alle argument til korresponderande verb til å karakterisere kva ein meiner med *deltakarane i verbsituasjonen*. Syntaktiske forhold i språket kan sjølvsagt gi grunnar til å *ikkje* kalle dette eit argument.

For å gjere dette konkret kan me sjå på setning 7 i test-suiten til XPar-prosjektet:

(4) abramsi brouns daenajleva sigaretze, rom cvimda
Abrams.NOM Brown.DAT vedde.3SG sigarett.om, at regne.3SG.IMP
'Abrams veddet en sigarett med Brown på at det regnet'

I følgje LFG-parsen til desse setningane har hovudpredikata svært ulik argumentstruktur[13]. Det norske *vedde* har fire argument, medan *da-najleveba* har to (*Abrams* og *Browne*), kor at-setninga på norsk og *rom cvimda* uttrykkjer same proposisjon og speler same rolle i verbsituasjonen. Den engelske LFG-parsen av den tilsvarande setninga (mine omsetjingar) gir tre argument, *with* blir her adjunkt, medan den tyske grammatikken, som au har tre argument, gjer *at*-setninga til adjunkt. I (5) nedanfor har eg representert dei omsetjingsmessig korresponderande frasane i f-strukturane med dei norske omsetjingane for å illustrere dette:

(5) a. Adams veddet en sigarett med Browne (norsk bokmål)
på at det regnet.

$$\begin{bmatrix} \text{PRED} & \text{'}\mathbf{vedde}\text{<Abrams, sigarett, Browne, regne>'} \\ \text{ADJUNCT} & \{\} \end{bmatrix}$$

b. abramsi brouns daenajleva sigaretze, rom cvimda. (georgisk)

$$\begin{bmatrix} \text{PRED} & \text{'}\mathbf{da\text{-}najleveba}\text{<Abrams, Browne, regne>'} \\ \text{ADJUNCT} & \{\text{sigarett}\} \end{bmatrix}$$

c. Abrams hat mit Browne um eine Zigarette gewettet, (tysk)
daß es regnet.

$$\begin{bmatrix} \text{PRED} & \text{'}\mathbf{wetten}\text{<Abrams, regne>'} \\ \text{ADJUNCT} & \{\text{Browne, sigarett}\} \end{bmatrix}$$

d. Abrams bet a cigarette with Brown that it was raining. (engelsk)

$$\begin{bmatrix} \text{PRED} & \text{'}\mathbf{bet}\text{<Abrams, sigarett, regne>'} \\ \text{ADJUNCT} & \{\text{Browne}\} \end{bmatrix}$$

Om ein skal ha grammatikkane som datagrunnlag er det altså eit reellt problem kva ein skal gjere med mangel på samsvar i argumentstruktur. Om det alltid var fullstendig samsvar i argumentstruktur, ville det vore trivielt å lenkje argument: viss to korresponderande verb hadde tre argument, ville me lenkja det første med det første, det andre med det andre og det tredje med det tredje. Men om me har analysar som dei over, ser det ut til at me er avhengig av LPT-kravet frå del 1.5 for å avgjere kva for adjunkt og argument som samsvarer.

[13] Analysane er henta 18. mai, 2009, frå http://decentius.aksis.uib.no/logon/xle.xml, som implementerer LFG-grammatikkane frå ParGram-prosjektet (Butt et al., 2002).

LPT-kravet blir forresten endå viktigare når det gjeld lenkjing av adjunkt til adjunkt. Adjunkt plukker ut si eiga rolle (argument får rolla tildelt frå verbet) og f-strukturane ordnar ikkje adjunkt etter nokon rekkjefølgje, dei er representert som uordna mengder, medan følgja mellom argument iallfall potensielt kan nyttast til å indikere semantisk likskap.

Ein kan argumentere for at grammatikkane her *burde* hatt like (eller likare) analysar, dette ville letta lenkjingsarbeidet, men sidan stoda no er slik, må krava ta høgd for lenkjer mellom argument og adjunkt. Om seinare utgåver av grammatikkane gir likare analysar, vil det iallfall ikkje gi verre lenkjingsresultat.

Og ei enkel korpusundersøking tyder på at det er relativt sjeldan at ein får slike situasjonar som (5) illustrerer. I Unhammer (2009) analyserte eg setningane frå den manuelt frasesamanstilte trebanken SMULTRON (Samuelsson & Volk, 2006) med LFG-grammatikkane for engelsk og tysk i ParGram-prosjektet (Butt et al., 2002), for å undersøkje følgjande hypotese:

> participants in a verbal situation are expressed as arguments (rather than adjuncts) in the source language of a translation if and only if they are expressed as arguments (rather than adjuncts) in the target language.

Mellom anna fann eg at 2 av 15 korresponderande verbtoken hadde LFG-analysar kor argument korresponderte med adjunkt[14]. Her utgjorde altså dei grammatiske analysane (ein del av) data, og undersøkinga seier nok meir om analysane enn om språklege forhold. På et så tynt datagrunnlag kan me vel berre konstatere at me må kunne handtere argument-adjunkt-lenkjer, men argument-argument-lenkjer bør prioriterast viss alt anna er likt.

1.6.2 Ulik følgje i argumentstruktur

I tillegg til at argument kan lenkjast til adjunkt, kan koreferente argument ha ulik følgje i argumentstrukturen. Det er klart at me vil lenkje objektet til *gefallen* (eller bokmål: *behage*) med subjektet til *like*, og omvendt. Men rekkjefølgje i argumentstrukturane i ParGram-prosjektet er ofte basert på syntaktisk funksjon heller enn rolle, slik at eit verb som har tema som subjekt og opplevar som objekt vil ha tema før opplevar i argumentstrukturen, medan ei omsetjing av dette verbet kan ha opplevar før tema:

(6) a. der Tonfall gefällt mir nicht

$$\left[\text{PRED} \quad \text{`}\mathbf{gefallen}\text{<Tonfall, ich}_i\text{>'} \ldots\right]$$

$$\leftrightarrow$$

[14] 25 om ein inkluderer analysar kor minst eitt av argumenta ikkje hadde korrekt analyse (t.d. eit PRO der grammatikken burde funne eit substantiv).

b. jeg liker ikke tonen
$\left[\text{PRED} \quad \text{‘}\textbf{like}\text{<jeg}_i\text{, tonen>’ ...}\right]$

Argumentstrukturane i (6) har omvendt intern følgje. Igjen må me ha LPT-informasjon for å avgjere kva for lenkjing som er korrekt. Men i visse tilfelle vil ikkje ein gong LPT-informasjon vere nok:

(7) a. sie_j gefallen ihnen_i
$\left[\text{PRED} \quad \text{‘}\textbf{gefallen}\text{<de}_j\text{, de}_i\text{>’}\right]$

$\leftrightarrow$

b. de_i liker dem_j
$\left[\text{PRED} \quad \text{‘}\textbf{like}\text{<de}_i\text{, de}_j\text{>’}\right]$

Det finst ingen f-strukturinformasjon eller LPT-informasjon me kunne nytta til å sikre den korrekte lenkjinga *sie/dem* og *ihnen/de*; og viss me rangerer lik argumentstruktur over ulik, vil me her få feil resultat. Det me *kan* gjere (utanom å endre grammatikkane slik at argumentstruktur korresponderer med eit universelt tematisk rollehierarki) er å sjå på mange lenkjingar av same verbpar, og på den måten oppdage moglege feil. For enkelttilfelle, derimot, vil krava i denne oppgåva ikkje vere nok til å gi korrekt lenkjing.

1.6.3 Krav om argumentlenkjer

Sjølv om me ikkje krev lik følgje i argumentlenkjer, og tillèt argument-adjunkt-lenkjer, er det eit minstekrav for å lenkje to PRED-element at alle argumenta til det eine PRED-elementet kan korrespondere med argument eller adjunkt av det andre PRED-elementet. Dette følgjer av formålet med å finne ut korleis ulike språk realiserer ulike semantiske roller syntaktisk; om eit verbargument ikkje kan lenkjast til noko i omsetjinga (ikkje ein gong eit pro-element), er det usannsynleg at verba uttrykker same situasjon, og tildeler same roller. På same måte må sjølvsagt lenkja predikat ha LPT-korrespondanse. Dyvik et al. (2009, s. 75) gir følgjande krav på f-strukturnivå:

(8) Krav for lenkjing av to PRED-element p og q:
- a. ordformene til p og q har LPT-korrespondanse
- b. alle argument av p har LPT-korrespondanse med eit argument eller adjunkt av q
- c. alle argument av q har LPT-korrespondanse med eit argument eller adjunkt av p
- d. LPT-korrespondansane er ein-til-ein
- e. ingen adjunkt til p er lenkja til f-strukturar utanfor q, og omvendt

Det (8-d) seier er at me ikkje lenkjer t.d. to instansar av «hest» på det eine språket til éin instans av «horse» på det andre. Krav (8-e) kjem eg tilbake til nedanfor.

Det går an å gjere (8) strengare, og krevje at argumenta – i tillegg til å ha LPT-korrespondanse – sjølv er PRED-lenkja. Dette har eg ikkje gjort i implementasjonen min, men det er mogleg å ha det som eit rangeringskriterium, noko eg kjem tilbake til i del 1.8. Ved å *ikkje* krevje at lenkjinga går heilt til botn i f-strukturen blir det mogleg å seie at *setningane* er syntaktisk like, og at kanskje visse overordna frasar er syntaktisk like, men visse *delfrasar* kan likevel vere ulike og dimed ikkje vere lenkja.

Kva med f-strukturomgivnadene til p og q, skal me krevje at dei er like? I (8-e) har me eit krav om at adjunkt til p ikkje er lenkja til f-strukturar utanfor q, og omvendt. Men viss a_p er eit adjunkt til p, kan det lenkjast til ein *dotternode* av argument eller adjunkt til q? La a_q vere eit argument eller adjunkt til q, viss a_q er eit argument må det ved (8) ha LPT-korrespondanse med argument/adjunkt i p, men det treng ikkje vere lenkja – viss det er ulenkja gjeld ikkje krav (8) for a_q, så (8) hindrar ikkje ei lenkje mellom a_p og døtre av a_q.

I tillegg vil ikkje (8) hindre at t.d. den ytste f-strukturen i kjeldespråket er lenkja til eit XCOMP-argument på målspråket; men i dette tilfellet bør kanskje ikkje *setningane* vere lenkja i utgangspunktet.

Sjølv om det er logisk mogleg å gjere slike lenkjingar, er det vanskeleg å finne ikkje-vilkårlege avgrensingar for når ein skal kunne lenkje f-strukturar som står i ulike omgivnader; i implementasjonen min har eg difor følgt eit strengare krav enn (8-e):

(9) PRED-elementa p og q kan berre lenkjast om dei er ytste f-strukturar i lenkja setningar, eller er argument/adjunkt til lenkja f-strukturar.

Dette er ei tentativ formulering. Til no har eg ikkje sett døme kor (9) ikkje bør gjelde, men om det finst slike døme bør sjølvsagt kravet modifiserast.

Krav (8) og (9) bør i enkle situasjonar vere tilstrekkelege for lenkjing på f-strukturnivå, men det finst au meir komplekse korrespondansar mellom PRED-element. Desse ser eg på del 1.6.5.

1.6.4 SKRIV Adposisjonsobjekt

I følgjande setningspar har me eit objekt «sigarett» som svarer til PP-en «sigaretze» («sigareti» + «ze»), eit adjunkt:

```
Abrams veddet en sigarett med Browne på at det regnet.
abramsi brouns daenajleva sigaretze, rom cvimda.

F_s [ PRED sigarett ]

F_t [ PRED ze<1> 1[ PRED sigareti ] ]
```

F_s og F_t er døtre av dei ytre predikata i kvar setning, krav (iii) seier at det må vere LPT-korrespondanse mellom desse for at me skal kunne lenkje «veddet» og «daenajleva». Her synest det feil å føye saman «sigareti» og «ze», `(F_s . (F_t 1))`, sidan «sigarett» ikkje inneheld informasjonen gitt av «ze».

Det finst då to løysingar. Me kan slakke på LMT-kravet ved å la `L'(F_t) = {sigaretze, ze}` (evt. `{sigaret, ze}`), då kan me lenkje `(F_s . F_t)`, medan 1 er ulenkja.

Eller me kan lenkje `(F_s . 1)`, kor me har skikkeleg LMT-korrespondanse, men då må me slakke på (iii) og (iv), og altså ha lov til å «hoppe over» ein f-struktur for å lenkje «veddet» og «daenajleva». F_t er då ulenkja. Det er løysinga valt i Dyvik et al. (2009, s. 75, fotnote 3), og den løysinga eg følgjer vidare i oppgåva.

1.6.5 SKRIV Kausativar og inkorporering

Til no har me føresett at eit PRED-element anten er ulenkja, eller er lenkja til eitt og berre eitt anna PRED-element. Men i visse tilfelle kan det vere ønskeleg å lenkje til fleire PRED-element.

Denne delen treng fleire ekte døme, og forsvar for kvifor me vil ha ein-mange-lenkjer for våre *formål* (georgisk ser ut til å vere borte frå xle-web?)

I ein norsk *la*-konstruksjon, t.d. den me har i «å la noko fryse» (i tydinga å forårsake at noko frys til) har me semantiske bidrag frå både *la* og hovudverbet *fryse*, og begge har PRED-element (sjølv om bidraget frå *la* nok er meir «grammatisk»). Men slike perifrastiske kan gjerne omsetjast til leksikaliserte kausativar som berre har eitt PRED-element, men likevel med tydinga «å la fryse». Påfunnet i (10) illustrerer denne situasjonen:

(10) a. ho lar-fryse huset

$$\left[\text{PRED} \quad \text{‘}\textbf{la-fryse}\text{<ho, hus>’}\right]$$

$$\leftrightarrow$$

b. ho lar huset fryse

$$\begin{bmatrix}\text{PRED} & \text{‘}\textbf{la}\text{<ho, hus, XCOMP>’} \\ \text{XCOMP} & \left[\text{PRED} \quad \text{‘}\textbf{fryse}\text{<hus>’}\right]\end{bmatrix}$$

Her er altså den kausative tydinga leksikalisert, og verbet har berre eitt PRED-element (på same måte som det norske verbet *kjøle* berre har eitt PRED-element, ikkje *la* + *bli kald*).[15]

Den same situasjonen får me der eit argument eller adjunkt er inkorporert i verbet på det eine språket, men uttrykt som eit separat predikat på det andre språket, t.d. samisk *fierpmástallat* som på norsk blir *å fiske med garn* – to predikat på norsk tilsvarer eitt på samisk.

[15] Det går sjølvsagt an å analysere sjølv leksikaliserte kausativar som om dei har fleire PRED-element, men det bør i såfall skje på uavhengig grunnlag, ikkje for å gjere lenkjinga enklare.

I (10) har *la-fryse* to argument, som ved krav (8) begge må finne korresponderande argument eller adjunkt for å lenkje *la-fryse*. Då går det ikkje an å lenkje *la-fryse* til berre *fryse*, som har eitt argument; me får eit XCOMP til overs som manglar lenkje. Me kan heller ikkje lenkje berre *la* til *la-fryse*, sidan det då får ein XCOMP til overs.

Men, ved å ha ei ein-mange-lenkje, frå *la-fryse* til både *la* og *fryse*, kan me oppfylle krav (8). Då treng ikkje XCOMP-argumentet lenkjast til eit argument av *la-fryse*, det er allereie lenkja til PRED-elementet; det som står igjen er unionen av argumenta til *la* og *fryse*, desse må alle ha LPT-korrespondanse med argument eller adjunkt av *la-fryse*, og omvendt må alle argument av *la-fryse* ha LPT-korrespondanse med argument eller adjunkt av *la* eller *fryse* (utanom XCOMP-argumentet til *la*, som allereie har ei lenkje). Ein kan tolke dette som om *la* og *fryse* var samanføyd til eitt predikat som krevde to argument (her: *ho* og *huset*).

Den einaste formelle forskjellen mellom dette og substantivinkorporering blir då at substantivet ikkje krev eigne argument. Det er au mogleg å tenkje seg ein kausativ med eit inkorporert objekt, omsett til *la* + *hovudverb* + *objekt*, altså ei lenkje frå eitt PRED til tre PRED. Igjen vil me då sjå på dei resterande ulenkja argumenta på kvar side; kvar av desse må lenkjast med eit unikt argument eller adjunkt.

Men det bør kanskje vere grenser for kor langt slik samanføying kan gå, om ikkje anna fordi problemet fort blir komputasjonelt vanskeleg. Å opne for ein-mange-lenkjer mellom PRED-element (eller til og med mange-mange-lenkjer) gir ei mykje større mengd moglege løysingar på lenkjingsproblemet; i alle situasjonar der me krev LPT-korrespondanse mellom eit argument a_p av p og eit adjunkt a_q av q for å lenkje p og q, vil me no au ha ei mogleg løysing der a_q er ulenkja, medan a_p er samanføyd med p og difor ikkje treng LPT-korrespondanse med argument/adjunkt av q. Så kan det au hende at a_p sjølv kan samanføyast med eit av sine argument/adjunkt. Skal me sjå etter slike løysingar samtidig som me ser etter løysingar med ein-ein-lenkjer, vil me måtte leite gjennom mange ufruktbare stiar. Ein måte å unngå dette på er å nedprioritere samanføying, og berre prøve dette der det ikkje finst andre alternativ.

Men det er ikkje berre av omsyn til implementasjonen ein bør nedprioritere desse. Ei ein-mange-lenkje tyder på ein type omsetjingsskifte, og det er ønskeleg å først sjå etter samanstillingar som føreset syntaktisk likskap, før ein ser etter omsetjingsskifte. Den viktigaste informasjonen me har å gå på er at setningane er omsetjingar og difor har ein viss likskap – Ockhams barberkniv gir oss då grunn til å velje ei løysing som føreset lik syntaks over ei løysing som føreset ulik syntaks. Viss det er mogleg å opprette ei samanstilling på bakgrunn av lik syntaks, vil me prioritere denne.

I implementasjonen blir difor alle ein-til-ein-lenkjer prøvd først. TODO:implementere: Sidan kan ein prøve å føye saman eit ulenkja PRED-element p med eit ulenkja PRED-element a_p kor a_p er argument eller adjunkt av p, og der p og a_p vil kunne lenkjast med eit ulenkja PRED-element q ved føringane gitt over, og alle dei andre lenkjingskrava er dekkja. Me får då eit modifisert krav (8):

(11) Krav for samanføyd lenkjing frå PRED-elementa p og a_p, kor a_p er eit argument eller adjunkt av p, til PRED-elementet q:

a. ordformene til p og a_p har saman LPT-korrespondanse med ordformen til q
b. la A vere unionen av argument til p og argument til a_p, utanom a_p sjølv; alle element av A har LPT-korrespondanse med eit argument eller adjunkt av q
c. la D vere unionen av argument eller adjunkt til p og argument eller adjunkt til a_p, utanom a_p sjølv; alle argument av q har LPT-korrespondanse med eit element av D
d. LPT-korrespondansane er ein-til-ein
e. ingen adjunkt til p eller a_p er lenkja til f-strukturar utanfor q, og ingen adjunkt til q er lenkja til f-strukturar utanfor p

Det er trivielt å utvide dette kravet til å fungere for mange-mange-lenkjer au.

1.7 Krav på c-strukturnivå

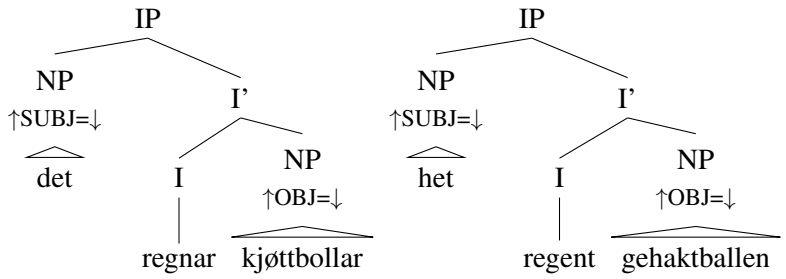
Ein f-struktur er projisert av ei mengd c-strukturnodar, det vil seie at det er desse nodane – det funksjonelle domenet til f-strukturen – som spesifiserer informasjonen som står i f-strukturen. Viss me har grunnlag for å lenkje to f-strukturar, vil me au ha grunnlag for å lenkje nodane som projiserte desse f-strukturane. Og omvendt vil det aldri vere grunnlag for å ha ei c-strukturlenkje som står i konflikt med f-strukturlenkjer, dvs. kor ϕ av kjeldenoden er lenkja til noko anna enn ϕ av målnoden (då burde kjeldenoden vore lenkja til dette andre). Det at to nodar er lenkja på c-strukturnivå må i det minste implisere at informasjonen dei projiserer korresponderer. I utgangspunktet bør krevje følgjande:

(12) to c-strukturnodar n_s og n_t kan berre lenkjast om $\phi(n_s)$ og $\phi(n_t)$ er lenkja på f-strukturnivå

Det enklaste ville vere å berre seie at alle nodane i dei to funksjonelle domena er mange-mange-lenkja med kvarandre, men denne lenkja vil ikkje gi oss meir informasjon enn at sjølve f-strukturane er lenkja; ei lenkje på c-strukturnivå bør kunne gi meir nyansert informasjon.

Det viktige forholdet på c-strukturnivå er *dominans*; hovudgrunnen til at me snakkar om c-struktur er at me vil skildre den hierarkiske inndelinga av frasestrukturen i setninga, der ein node på høgare nivå *dominerer* mengder av nodar på lågare nivå. Ei lenkje mellom to c-strukturnodar må altså implisere at det dominerte materialet korresponderer.

I figur 1.3 er dei funksjonelle domena til *regnar/regent* lenkja, og det same med *det/het* og *kjøttbollar/gehaktballen*. Viss me føreset at subjekt-NP-ane er lenkja med kvarandre, og at objekt-NP-ane er lenkja med kvarandre, på c-strukturnivå, vil det vere ønskeleg å ein-ein-lenkje IP-nodane, I'-nodane og I-nodane. Me skal

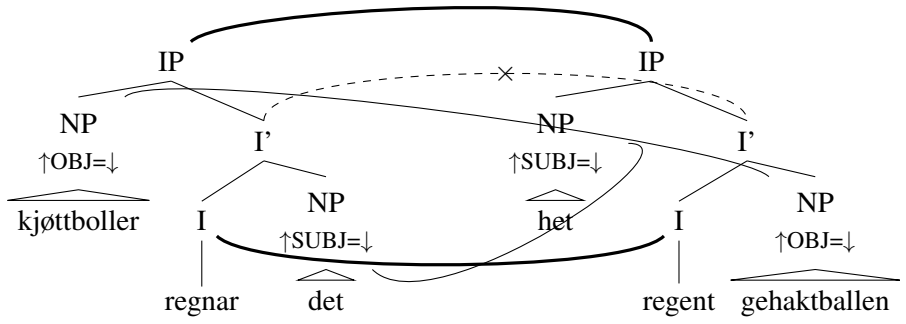


Figur 1.3: Enkel lenkjing av c-strukturnodar mellom norsk og nederlandsk; IP til IP, I' til I' og I til I.

sjå kvifor.

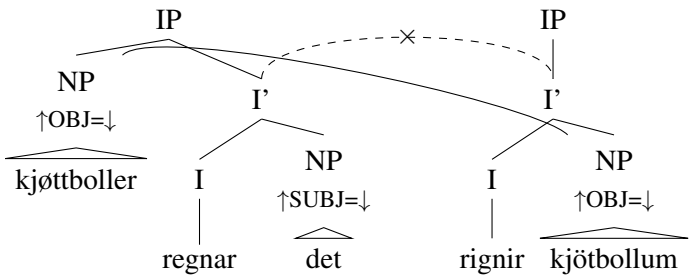
IP-nodane bør lenkjast sidan dei dominerer alt innanfor dei lenkja funksjonelle domena; det finst ikkje ein gong nodar som står utanfor det dei dominerer. Dei nodane som står nedanfor det funksjonelle domenet til IP-ane er i tillegg lenkja med kvarandre. Det vil seie at det ikkje finst informasjon på kjeldespråket som ikkje er uttrykt på målspråket (eller omvendt) innanfor det IP-ane dominerer.

I'-nodane dominerer ikkje subjekta i figur 1.3. Ei lenkjing av I'-nodane impliserer at det som står under desse korresponderer, men au at nodane står i liknande omgivnader. Det er lett å sjå føre seg eit døme der det ikkje ville vore ønskeleg med ei lenkje mellom I'-nodane. I figur 1.4 vil me t.d. ikkje lenkje desse nodane, på norsk dominerer I' subjektet, som er lenkja til subjektet på nederlandsk, men på nederlandsk står ikkje subjektet under I', og omvendt for objektet. Ei lenkje mellom I'-nodane ville sagt at nodane dei dominerte projiserte korresponderande informasjon, det gjer dei ikkje i figur 1.4. (I 1.3, derimot, står dei lenkja objekta under I', medan dei lenkja subjekta er utanfor.) Men merk at IP-nodane likevel kan lenkjast, dei dominerer begge både subjekt og objekt, sjølv om dei kjem i ulik følgje under. I-nodane dominerer berre verba, og kan au lenkjast.



Figur 1.4: C-strukturlenkjer kan ikkje gå på tvers av dominerte lenkjer (norsk og nederlandsk)

Sjølv om subjektet sto ulenkja, t.d. ved lenkjing inn i eit pro-drop-språk eller liknande, ville me fått same situasjon; I'-nodane i figur 1.5 kan ikkje lenkjast sidan I' på islandsk dominerer objektet, medan I' på norsk ikkje gjer dette, og objekta er lenkja med kvarandre (her både på c- og f-strukturnivå). Ei lenkje mellom desse I'-nodane ville sagt at dei dominerer korresponderande materiale, men det gjer dei ikkje.



Figur 1.5: C-strukturlenkjer kan ikkje gå på tvers av dominerte lenkjer (norsk og islandsk)

Når treet deler seg i to som i desse figurane, får me ei mogleg oppdeling av kjeldene til f-strukturinformasjonen. Me vil ikkje lenkje nodar som ikkje gir same tilskot til f-strukturen, på same måte som me ikkje vil lenkje på tvers av f-strukturlenkjer.

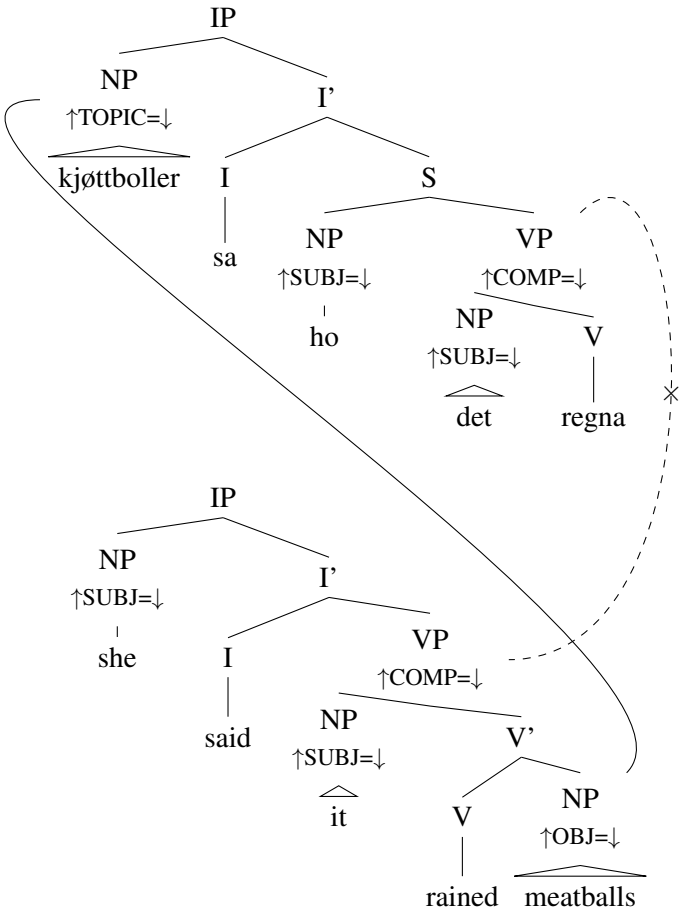
I både figur 1.4 og figur 1.5 er det slik at det I'-nodane dominerer gir ulike tilskot til f-strukturen, dei kan difor ikkje lenkjast. Likevel må me tillate litt slingringsmonn her, nodane skal ikkje trenge projisere heilt like f-strukturar. Det som er relevant er det som blir lenkja i f-strukturen.

Som desse døma viser må me nyansere prinsippet om å ikkje lenkje c-strukturnodar på tvers av f-strukturlenkjer, til å ta innover seg dominans: me vil ikkje lenkje c-strukturnodar viss *det dei dominerer* kjem i konflikt med f-strukturlenkjer.

I visse tilfelle kan det hende at sjølv toppnodane i det funksjonelle domenet ikkje bør lenkjast. I døma over dominerer toppnoden i det funksjonelle domenet, IP, alt som står under $\phi(IP)$ i f-strukturen. I figur 1.6, derimot, er objektet til *regna* ikkje dominert av toppnoden i det funksjonelle domenet til *regna*, VP-en; men det er lenkja til objektet i funksjonelle domenet til *rained*. F-strukturane til dei to VP-ane er lenkja, men toppnodane i dei funksjonelle domena kan ikkje lenkjast.

Me vil altså lenkje ein node n_s med n_t berre viss $\phi(n_s)$ er lenkja på f-strukturnivå med $\phi(n_t)$, og det ikkje finst nodar under n_s som er lenkja med nodar utanfor det funksjonelle domenet til n_t, og omvendt.

Men, kva om det finst nodar under n_s som ikkje er lenkja på c-strukturnivå (kanskje fordi det ikkje finst tilsvarande nodar på målspråket, t.d. ved lenkjing inn i pro-drop-språk), men som har ei lenkje på f-strukturnivå? Her finst det fleire alternative løysingar, som eg ser på nedanfor.



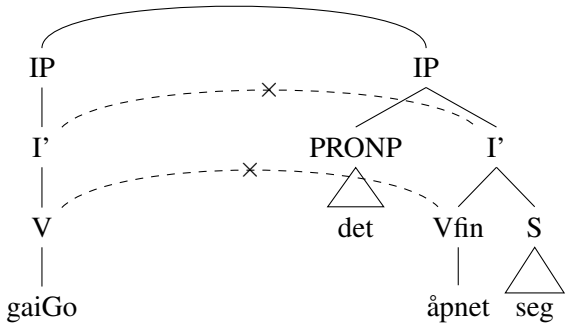
Figur 1.6: Sjølv toppnodane i eit funksjonelt domene kan stå ulenkja; her kan ikkje VP-nodane lenkjast sidan det norske TOPIC er objektet til *regna*, lenkja til objektet under VP på engelsk

1.7.1 Ulenkja c-strukturnodar

I figur 1.7 kan iallfall IP-nodane lenkjast, dei dominerer alle orda på begge setningane, og f-strukturane er lenkja

For å gjere dette meir konkret kan me sjå på dømet i figur 1.7. IP-nodane er her lenkja, men me kan ikkje lenkje I'-nodane i same funksjonelle domene. PRONP-noden, spesifikator på den norske sida, er ikkje lenkja med nokon spesifikator på den georgiske sida. Den informasjonen (her reint syntaktisk) som ordet *det* tilfører IP, ligg under I' på georgisk. Me kunne lenkja I' om me hadde ein georgisk spesifikator som var lenkja til den norske PRONP. Me kan heller ikkje lenkje Vfin til V, her manglar endå meir informasjon (både frå *det* og frå *seg*).

CURRENT TODO (bilete av f-lenkjer? iallfall fjerne ein del i denne delen, og liste opp alternativ LL vs LL_c)



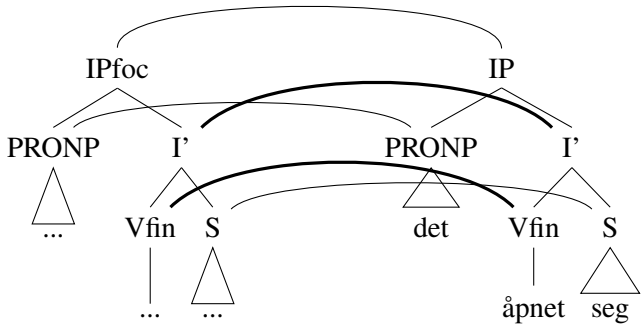
TODO: teikne inn f-domene her òg

Figur 1.7: Umogleg lenkjing av underordna c-strukturnodar mellom georgisk og bokmål

Hadde det georgiske treet hatt spesifikator og komplement som kunne lenkjast til spesifikator og komplement på norsk, kunne me ha lenkja I' og Vfin. For å tillate desse lenkjene, men ikkje dei i figur 1.7, ville det vore nok å krevje at søsternodane var lenkja. Hadde me hatt situasjon i figur 1.8 nedanfor ville dette stemt, Vfin-nodane kan lenkjast fordi S-nodane er lenkja, I'-nodane fordi PRONP-nodane er lenkja.

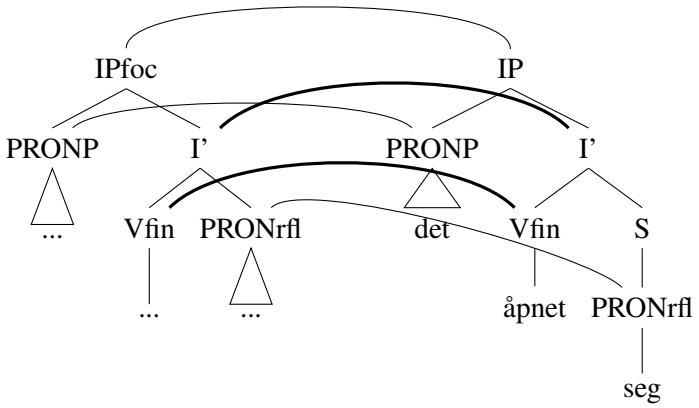
Men det blir for strengt å krevje at søsternodar er lenkja; S-noden i den norske analysen er ein del av det funksjonelle domenet til IP, medan komplementet i andre språk kanskje går rett på eit nytt funksjonelt domene. Figur 1.9 demonstrerer denne situasjonen. Her kan ikkje S lenkjast til PRONrfl sidan dei ikkje er i same funksjonelle domene, men me vil jo likevel lenkje Vfin-nodane.

Me treng altså eit litt meir nyansert krav. Dyvik et al. (2009, s. 77) definerer i denne samanhengen omgrepet *lenkja leksikalske nodar*, *LL*, kor $LL(n)$ er mengda av nodar dominert av n som har ei ordlenkje. For å lenkje c-strukturnodane n_s og n_t, som er i lenkja funksjonelle domene, må alle nodane i mengda $LL(n_s)$ vere lenkja til nodar i $LL(n_t)$. Ulenkja nodar under n_s og n_t står ikkje i vegen for lenkjing av n_s og n_t, men dei to mengdene kan ikkje vere tomme. For at me skal unngå å lenkje I' og Vfin i figur 1.7 må altså *det* og *seg* òg vere ordlenkja til *gaiGo*. Dette blir



TODO: teikne inn f-domene her òg

Figur 1.8: Ei enkel lenkjing av c-strukturnodar



TODO: teikne inn f-domene her òg

Figur 1.9: Her vil me lenkje Vfin-nodane utan å lenkje søstrene deira.

ei ein-til-mange-lenkje på ordnivå, som må representerast som fleire lenkjer som alle byrjar i *gaiGo*. Viss georgisk er kjeldespråket (n_s, norsk: n_t) blir $LL(IP_s) = LL(I'_s) = LL(V_s) = \{(\text{det}, \text{gaiGo}), (\text{åpnet}, \text{gaiGo}), (\text{seg}, \text{gaiGo})\} = LL(IP_t)$. Mengdene $LL(I'_t) = \{(\text{åpnet}, \text{gaiGo}), (\text{det}, \text{gaiGo})\}$ og $LL(Vfin_t) = \{(\text{åpnet}, \text{gaiGo})\}$ på den norske sida har ikkje korresponderande mengder på georgisk og blir ikkje lenkja[16].

Men viss me vil unngå å referere til ordlenkjer, går det au an å definere kravet i form av f-strukturlenkjer på preterminale nodar[17]:

(13) For å lenkje dei underordna c-strukturnodane n_s og n_t (kor toppnodane i dei funksjonelle domena er lenkja ved krav (1-c)):
La $L_c(n_s)$ vere alle f-strukturlenkjer frå $\phi(n'_s)$ for alle preterminale n'_s som er dominert av n_s. Ulenkja nodar n'_s er ikkje med i L_c. Nodane n_s og n_t kan då lenkjast viss

a. mor av n_s, m_s, har same funksjonelle domene som n_s, og mor av n_t, m_t, har same funksjonelle domene som n_t,
b. m_s og m_t er lenkja,
c. og $L_c(m_s) - L_c(n_s) = L_c(m_t) - L_c(n_t)$.

(13-a og -b) seier til saman det same som (12). Det (13-c) seier er at me må ha same *informasjonstap* når me går nedover i c-strukturtreet.

Figur 1.9 illusterer dette kravet. IPfoc og IP er toppnodar, lenkja ved krav (1-c). I'-nodane er i lenkja funksjonelle domene og har lenkja mødre. I tillegg har dei same informasjonstap:

$L_c(IPfoc_s) = \{(\textbf{PanJara}, \textbf{den}), (\textbf{ga-Geba}, \textbf{åpne})\}$, $L_c(I'_s) = \{(\textbf{ga-Geba}, \textbf{åpne})\}$
$L_c(IP_t) = \{(\textbf{PanJara}, \textbf{den}), (\textbf{ga-Geba}, \textbf{åpne}), (\textbf{pro}, \textbf{pro})\}$, $L_c(I'_t) = \{(\textbf{ga-Geba}, \textbf{åpne}), (\textbf{pro}, \textbf{pro})\}$
$L_c(IPfoc_s) - L_c(I'_s) = \{(\textbf{PanJara}, \textbf{den})\} = L_c(IP_t) - L_c(I'_t)$

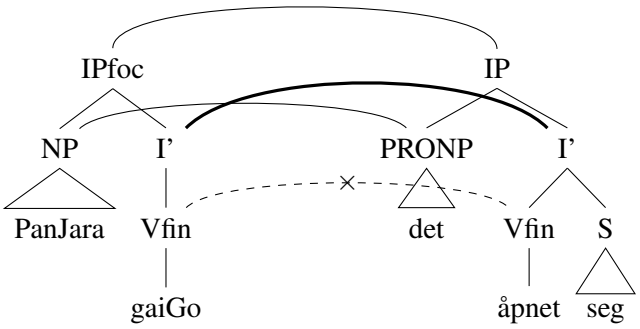
Vfin-nodane er au i same funksjonelle domene, og har lenkja mødre, men her har dei ulikt informasjonstap:

$L_c(I'_s) = \{(\textbf{ga-Geba}, \textbf{åpne})\}$, $L_c(Vfin_s) = \{(\textbf{ga-Geba}, \textbf{åpne})\}$
$L_c(I'_t) = \{(\textbf{ga-Geba}, \textbf{åpne}), (\textbf{pro}, \textbf{pro})\}$, $L_c(Vfin_t) = \{(\textbf{ga-Geba}, \textbf{åpne})\}$
$L_c(I'_s) - L_c(Vfin_s) = \{\}$
$L_c(I'_t) - L_c(Vfin_t) = \{(\textbf{pro}, \textbf{pro})\}$

Me ikkje har altså ein c-strukturnode som direkte projiserer f-strukturen til *pro* på georgisk, lenkja til *seg* på norsk; f-strukturen til *pro* er spesifisert via verbet.

[16] Eg har ikkje gjort noko forskjell på nodar som dominerer like mengder på same språk, dette kan tolkast som at alle dei ikkje-terminale nodane på georgisk er lenkja til IP på norsk.

[17] Då kan me au representere mange-til-mange-ordlenkjer som «udelelege», $(\{\text{gaiGo}\}, \{\text{det,åpnet,seg}\})$ blir den einaste ordlenkja i dømet over, sidan me ikkje må samanlikne ordlenkjene frå IP_t, I'_t og $Vfin_t$.



TODO: teikne inn f-domene her òg

Figur 1.10: Delvis mogleg lenkjing av underordna c-strukturnodar mellom georgisk og bokmål

Men sjølv om me ikkje kan sjå ein slik node, kan me sjå at det er *noko* som er spesifisert via dette fragmentet av den georgiske c-strukturen, og at det er lenkja til *seg*, difor mister me informasjon når me går frå I' til Vfin.

1.7.2 Alternativ formulering, utan å sjekke informasjonstap

F-strukturen til *gaiGo* i dømet over har pro-elementet (lenkja til *seg*) som argument, difor burde det au vere mogleg å finne f-strukturlenkja til dette argumentet når ein finn $L_c(I'_s)$. Om ein legg til denne lenkja i L_c, er det nok at $L_c(n_s) = L_c(n_t)$ for at n_s og n_t skal lenkjast (dvs. ein treng ikkje sjekke forskjellen til mornodane). Men skal ein formulere eit slikt krav må ein passe på å spesifisere kor djupt ned i f-strukturen ein skal kikke, og avgjere om adjunkt-lenkjer skal med, osb. I implementasjonen min har eg valt å følgje krav (13) for å unngå å måtte kode denne typen informasjon inn i systemet.

1.7.3 Funksjonelle c-strukturnodar

Ikkje alle ord tilsvarer PRED-element i f-strukturen, dette gjeld typisk funksjonsord (t.d. *som*, *at*). Ved endosentrisitetsprinsippa til Bresnan (2001) er komplementet til funksjonelle kategoriar (C, I, P) ein funksjonell ko-kjerne, det er altså komplementet som gir PRED-elementet i dette funksjonelle domenet.

Problemet med å nytte metoden frå del 1.7 i dette tilfellet er at nodar over funksjonsord er i det same funksjonelle domenet som komplementet, og nodane over funksjonsorda tilføyer ikkje ei ny PRED-lenkje som kan dele opp treet slik me gjorde tidlegare. Så me må utvide prinsippa for å dele opp c-strukturtreet i buntar med likt informasjonstap.

Ord som ikkje projiserer PRED-lenkjer kan likevel ha LPT-korrespondanse og bestå krava på ordnivå, men når me skal lenkje desse på c-strukturnivå må me sjekke ordkrava direkte (me kan ikkje gå via nokon f-strukturlenkjing). LPT-kravet

gir oss eit utgangspunkt for lenkjing.

Viss begge språk har funksjonsord, men funksjonsord som ikkje kan sjåast på som moglege omsetjingar (t.d. *fordi* og *whether*), bør me ikkje ein gong lenkje komplementa[18]. Samtidig vil me ikkje at eit manglande funksjonsord på det eine språket skal hindre lenkjing av komplementa, sidan det kan hende at funksjonsordet ikkje er krevd på det språket (eventuelt kjem dette fram som korrespondansar i f-strukturtrekk, eg har ikkje teke høgd for korrespondansar mellom element som ikkje er PRED i denne oppgåva).

Me krev at komplementa er lenkja for å sikre at me ikkje lenkjer nodar som står i ulike konstekstar (me vil ikkje lenkje *at* i «han såg at det gjekk bra» med *that* i «he saw that she drew a picture»), jamfør kravet om lenkja argument for lenkja predikat i del 1.6.1.

Då får me følgjande:

(14) Krav for lenkjing av funksjonelle kategoriar i c-strukturen:

a. Gitt ei mogleg lenkjing av FP og GP, kor F og G er funksjonelle kategoriar der komplementa elles kan lenkjast, introduserer me eit «falskt informasjonstap» mellom FP og F' og mellom GP og G'; orda under F' og G' må ha LPT-korrespondanse for at FP og GP skal kunne lenkjast, då kan me au lenkje F' og G'.

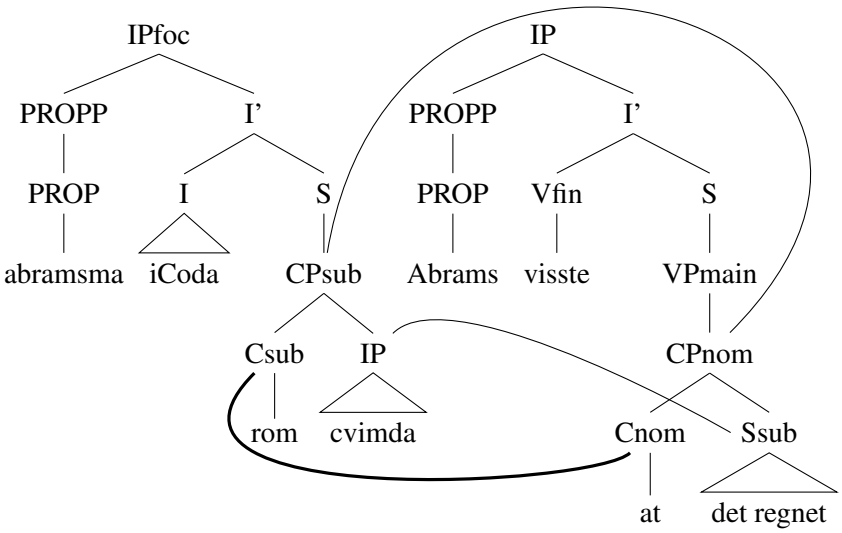
b. Gitt ei mogleg lenkjing av FP og XP, der F er ein funksjonell kategori, medan X er ein ikkje-funksjonell kategori, ignorerer me den funksjonelle kategorien i c-strukturlenkjinga. Sidan det ikkje er noko forskjell i informasjonstap mellom FP og F', er F' medlem av nodemengden som blir lenkja til XP.

Om (14-a) er oppfylt, kan me få samanstillinga vist i figur 1.11. Her vil dei funksjonelle domena til CPsub og CPnom kvar kunne delast opp i to deler, kor den funksjonelle delen har LPT-korrespondanse medan komplementa er lenkja på f-strukturnivå. Det er ingen informasjonstap frå CPsub til Csub som ikkje er reflektert i CPnom til Cnom, og det er ingen informasjonstap frå CPsub til IP som ikkje er reflektert i CPnom til Ssub.

(Alle nodane under S vist i dei to trea er i same funksjonelle domene, så om dei funksjonelle domena er lenkja, vil krav (12) vere oppfylt kva gjeld CP-komplementa – lenkjinga går ikkje ut over dei funksjonelle domena, medan krav˜(13) er dekkja for IP og Ssub med unntaket over.)

Der det eine språket har eit funksjonsord og det andre språket ikkje krever det, bryr me oss ikkje om funksjonsordet. For å sjekke noko slikt må me som nemnt sjå på andre trekk enn PRED i f-strukturane, noko som blir utanfor denne oppgåva; men om me hadde sjekka slike f-strukturkorrespondansar kunne me unngått

[18] Skal ein lenkje ordet *som* (utan PRED) med ordet *which* (med PRED)? Viss krava elles er oppfylt, kan det kanskje vere informativt med ein type «defekt» lenkje, sjølv om berre det eine ordet blir rekna for å vere eit innhaldsord. Frasane til deira funksjonelle domene vil uansett vere samanstilt via toppnodane (t.d. CP).



Figur 1.11: Mogleg samanstilling av funksjonelle c-strukturnodar mellom georgisk og norsk (bokmål)

kravet om LPT-korrespondanse og i staden nytta informasjon frå f-strukturane til lenkjing av funksjonelle kategoriar. Utan å ha slike mekanismar på plass blir f-strukturlenkjinga avhengig av c-strukturforhold, og i implementasjonen min har eg difor lagt mindre vekt lenkjing av funksjonelle kategoriar.

1.8 SKRIV Rangering

(meir om dette i del **??**)

Kapittel 2

Avslutning

Referansar

Bresnan, J. (2001). *Lexical-Functional Syntax*. Oxford, UK: Blackwell Publishers. Tilgjengeleg frå http://books.google.com/books?id=7elu0CcxQWkC (ISBN: 0631209743)

Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. & Mercer, R.L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311. Tilgjengeleg frå http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.8919

Butt, M. (1998). Constraining Argument Merger Through Aspect. I E. Hinrichs, A. Kathol & T. Nakazawa (red.), *Complex predicates in nonderivational syntax* (vol. 30, kap. 1). New York: Academic Press.

Butt, M., Dyvik, H., King, T., Masuichi, H. & Rohrer, C. (2002). The Parallel Grammar Project. I *COLING-02 on Grammar engineering and evaluation* (vol. 15, s. 1–7). Morristown, NJ: Association for Computational Linguistics. Tilgjengeleg frå http://portal.acm.org/citation.cfm?id=1118783.1118786

Cheung, L., Lai, T., Luk, R., Kwong, O., Sin, K., Tsou, B. et al. (2002). Some Considerations on Guidelines for Bilingual Alignment and Terminology Extraction. , 1–5. Tilgjengeleg frå http://www.aclweb.org/anthology-new///W/W02/W02-1802.pdf

Dyvik, H., Meurer, P., Rosén, V. & Smedt, K.D. (2009). Linguistically motivated parallel parsebanks. I M. Passarotti, A. Przepiórkowski, S. Raynaud & F.V. Eynde (red.), *Proceedings of the eighth international workshop on treebanks and linguistic theories* (s. 71–82). Milan, Italy: EDUCatt. Tilgjengeleg frå http://tlt8.unicatt.it/allegati/Proceedings_TLT8.pdf#page=83

Hearne, M., Ozdowska, S. & Tinsley, J. (2008). Comparing Constituency and Dependency Representations for SMT Phrase-Extraction. I *Actes de la 15e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN '08)*. Avignon, France. Tilgjengeleg frå http://www.computing.dcu.ie/~mhearne/publications.html

Koehn, P., Och, F. & Marcu, D. (2003). Statistical phrase-based translation. I *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (s. 48–54). Morristown, NJ, USA. Tilgjengeleg frå http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/phrase2003.pdf

Meurer, P. (2008, March). *A Computational Grammar for Georgian.* Tilgjengeleg frå http://maximos.aksis.uib.no/~paul/articles/Tbilisi2007-LNAI.pdf

Munday, J. (2001). *Introducing Translation Studies: Theories and Applications*. London: Routledge.

Piao, S. & McEnery, T. (2001). Multi-word Unit Alignment in English-Chinese Parallel Corpora. I P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja (red.), *Proceedings of the Corpus Linguistics 2001 Conference* (s. 466–475). Lancaster, UK. Tilgjengeleg frå http://personalpages.manchester.ac.uk/staff/scott.piao/research/papers/mwu_align4.pdf

Pullum, G. & Scholz, B. (2001). On the Distinction between Model-Theoretic and Generative-Enumerative Syntactic Frameworks. *Logical Aspects of Computational Linguistics: 4th International Conference, Lacl 2001, Le Croisic, France, June 27-29, 2001, Proceedings*. Tilgjengeleg frå http://portal.acm.org/citation.cfm?id=645668.665062

Riezler, S. & Maxwell, J. (2006). Grammatical Machine Translation. I M. Butt, M. Dalrymple & T.H. King (red.), *Intelligent Linguistic Architecture: Variations on themes by Ronald M. Kaplan* (s. 35–52). Stanford, CA: CSLI Publications. Tilgjengeleg frå http://www.parc.com/research/publications/details.php?id=5675

Rosén, V., Meurer, P. & Smedt, K. de. (2009). LFG Parsebanker: A Toolkit for Building and Searching a Treebank as a Parsed Corpus. I F.V. Eynde, A. Frank, G. van Noord & K.D. Smedt (red.), *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7)* (s. 127–133). Utrecht: LOT. Tilgjengeleg frå http://ling.uib.no/~desmedt/papers/tlt7rosen-submitted.pdf

Samuelsson, Y. & Volk, M. (2006). Phrase Alignment in Parallel Treebanks. I *Proceedings of Treebanks and Linguistic Theories (TLT '06).* Prague. Tilgjengeleg frå http://ling16.ling.su.se:8080/new_PubDB/doc_repository/229_align.pdf

Samuelsson, Y. & Volk, M. (2007). Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment. I *Proceedings of Treebanks and Linguistic Theories (TLT '07).* Bergen, Norway.

Thunes, M. (2003). *Ekserpering av leksikalske oversettelsekorrespondanser fra parallelltekst.* Tilgjengeleg frå `http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/marthaex.pdf`

Tinsley, J., Hearne, M. & Way, A. (2007). Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. I *Proceedings of Treebanks and Linguistic Theories (TLT '07).* Bergen, Norway.

Unhammer, K.B. (2009). *Do arguments and adjuncts ever align? LINGMET semester assignment.* Tilgjengeleg frå `http://www.student.uib.no/~kun041/doc/argstr.pdf`

XPar. (2008). *XPAR: Language diversity and parallel grammars.* (Submitted to the Research Council of Norway.)

[fn:2] Tilgjengeleg frå http://github.com/unhammer/lfgalign som fri og open programvare under GNU General Public License.

[fn:5] Formatet er dokumentert på http://www2.parc.com/isl/groups/nltt/xle/doc/xle.html. Importeringa til Lisp-strukturar handterer «pakka representasjonar» og kjenner igjen ekvivalensforhold (t.d. der fleire ϕ-variablar refererer til same f-struktur, eller fleire Prolog-variabler refererer til same analyseval); men filene eg har testa utnyttar ikkje det fulle spennet til formatet, så det finst ganske sikkert feil.

[fn:16] Dette språkvalet kan gjere eventuell integrering med andre LFG-system lettare (Common Lisp er m.a. nytta i LFG Parsebanker (Rosén et al., 2009)).

[fn:17] Når eg her skriv at to f-strukturar har LPT-korrespondanse, meiner eg sjølvsagt at ordformene til PRED-verdien til kvar f-struktur har LPT-korrespondanse.

[fn:18] Eigentleg eit slag avgjerdstre; kvart element er eit par, kor første element er lenkja mellom dei yttarste f-strukturane, og andre element er dei moglege samanstillingane for dei indre strukturane. Denne strukturen kan vere nyttig for å rangere samanstillingar, og `f-align` blir mykje meir oversiktleg av å jobbe med eit slikt tre. Ein funksjon `flatten` omformar det ferdige treet til ei enkel liste med samanstillingar, kor kvar samanstilling er ei flat liste med lenkjer mellom f-strukturar.

[fn:19] Ved c-struktur-f-strukturavbildinga ϕ, ein funksjon som tek ein c-strukturnode og returnerer ein (delvis) f-struktur.

[fn:20] Dette er ein litt enklare måte å definere kravet på; ei *lenkje* refererer til både kjelde og mål, dimed blir det mogleg å seie at ein node på kjeldespråket kan dominere same mengd som ein node på målspråket.

[fn:8] Som nemnt i del 1.7.3 kan funksjonsord gjere f-strukturlenkjinga avhengig av forhold i c-strukturen, evt. krevje meir nyansert f-strukturlenkjing. Dette har eg ikkje teke høgd for i implementasjonen, så der eit funksjonsord burde blokkert ei f-strukturlenkje vil `lfgalign` gi feil samanstilling.