Syntaktisk informert frasesamanstilling

Kevin Brubeck Unhammer

22/09, 2010

Innhald

Kapittel 1

Innleiing

Denne masteroppgåva utforskar kva det vil seie at to uttrykk er omsetjingar av kvarandre, og korleis me automatisk kan generere og evaluere samanstilling (*alignment*) av uttrykk som står i eit slikt omsetjingsforhold.

TODO: abstract/samandrag

Omsetjingsforhold finn me mellom setningar i kontekst på ulike språk, men me kan au finne ulike typar ekvivalensforhold (samanstillingar) mellom frasar innanfor setningane, og mellom lingvistiske skildringar av setningane. I samanheng med XPar-prosjektet (XPar, 2008) har eg sett på metodar for automatisk frasesamanstilling – å finne omsetjingsforhold mellom grupper av fleire ord. Resultatet blir ein *annotasjon*, endå ei lingvistisk skildring av tekstene.

Det at me kan finne korrespondansar mellom lingvistiske skildringar (t.d. trekkstrukturane til dei grammatiske rammeverka HPSG eller LFG) gjer det tydeleg at me arbeider med ein *modell* av språket; ulike skildringar kan vere sanne innanfor modellen, utan at modellen er lik språket. Sjølve omsetjingsforholdet er au ein teoretisk storleik, og me kan leggje ulike kriterium til grunn for å kalle to uttrykk omsetjingar av kvarandre.

Kriteria avheng av formålet. Samanstillingsannotasjon kan t.d. nyttast som grunnlag for statistisk eller eksempelbasert maskinomsetjing, i tillegg til oppbygging av parallelle korpora for meir teoretiske språkstudium. For statistisk maskinomsetjing vil alle uttrykk vere omsetjingar av kvarandre med eit visst sannsyn (kanskje null), ein har vanlegvis ikkje kriterium som krev lingvistisk analyse. Når samanstillinga skal nyttast i parallelle korpora for lingvistiske undersøkingar vil ein kanskje ha krav om at uttrykk som skal lenkjast er «like» på eit eller anna mål, utover at dei har opptredt saman ofte; i den manuelle samanstillinga i Samuelsson & Volk (2006) har dei t.d. ein del reint semantiske kriterium for å opprette fraselenkjer i ein parallell trebank, men dei har ikkje krav om syntaktisk likskap.

Xpar-prosjektet, som denne masteroppgåva er ein del av, har mellom anna som mål å oppdage forhold mellom grammatiske funksjonar, tematiske roller og kasusmarkering, ved hjelp av parallelle trebankar annotert med djupe grammatiske analysar. For å lenkje to frasar i dette prosjektet krev me ein viss syntaktisk likskap i omgivnadene til frasane, sjølv om frasane internt kanskje er syntaktisk uli-

ke. Grammatikkane som gir analysane er utvikla med tanke på at *like syntaktiske fenomen på ulike språk skal få like analysar*, frasesamanstillinga bør då kunne tene på at omsetjingar som har ein syntaktisk likskap vil få liknande analysar.

Dei grammatiske analysane er gjort i leksikalsk-funksjonell grammatikk, LFG (Bresnan, 2001). Ei grammatisk analyse i LFG involverer både konstituentstruktur (c-struktur) og funksjonell struktur (f-struktur). Konstituentstrukturen liknar på frasestrukturtrea frå andre grammatiske tradisjonar. Dei funksjonelle strukturane er trekkstrukturar, som mellom anna representerer avhengnadsforhold mellom syntaktiske funksjonar som predikat, subjekt og objekt, i tillegg til å halde informasjon om grammatiske trekk som genus, tal eller kasus. Nodar i c-strukturen kan spesifisere informasjon på ulike stader i f-strukturen[1].

I XPar-prosjektet vil ein finne ut om metodar for frasesamanstilling kan tene på det at LFG-grammatikkane for dei ulike språka er skrivne med same prinsipp lagt til grunn; to parallellstilte setningar bør ha f-strukturar som er like nok til at me kan samanstille frasar ved hjelp av likskapen mellom f-strukturane. I Dyvik et al. (2009, s. 72) finn me følgjande hypotese:

> On the basis of monolingual treebanks constructed from a parallel corpus by means of parallel grammars it will be possible to achieve automatic word and phrase alignment with significantly higher precision and recall than hitherto achieved through other means.

kor «parallel grammars» her tydar at grammatikkane har ein viss parallellisme i både f-struktur og c-struktur.

Men i tillegg til at ein kanskje kan få betre skåre på desse kvantitative måla, vil lenkjer mellom f-strukturar gi informasjon som er kvalitativt forskjellig frå det ein kan få med å berre sjå på lenkjer mellom ord, N-gram eller konstituentar.

I denne masteroppgåva spesifiserer eg kva for lenkjer mellom f-strukturar og c-strukturnodar me ønskjer, implementerer eit program `lfgalign` som automatisk finn samanstillingar med slike lenkjer, evaluerer resultatet av å køyre programmet mitt, og samanliknar dette med kva me kan få frå andre metodar.

Programmet `lfgalign` opprettar frasesamanstillingar med hjelp av f-strukturinformasjonen gitt av dei parallelle grammatikkane, og bottom-up-informasjon om kva for ordsamanstillingar som er moglege. F-strukturane avgrensar igjen kva for ordsamanstillingar som er moglege, og kva for c-strukturnodar (syntaktiske frasar) som kan lenkjast.

1.1 Vegkart

I neste kapittel ser eg på andre metodar for frasesamanstilling.

I kapittel 3 går eg gjennom kva me ønskjer av ei frasesamanstilling når formålet m.a. er å oppdage relasjonane mellom syntaktiske funksjonar, kasusmarkering og

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

tematiske roller med hjelp av ein parallell trebank. Dette ender opp i ei liste med «krav» som samanstillingane må fylle for å vere lovlege, og som implementasjonen av den automatiske frasesamanstillinga (kapittel 4) må følgje.

Eg evaluerer samanstillingane som kjem ut av denne metoden i kapittel 5, og samanliknar dei med det som er mogleg der me berre har konstituentstruktur (syntaktiske tre) i tillegg til ordsamanstilling.

Eg nyttar språka georgisk og norsk i evalueringa, hovudsakleg fordi dei er svært ulike syntaktisk og morfologisk; Georgisk er mellom anna eit pro-drop-språk, med friare ordfølgje og rikare morfologi enn norsk.

fortelje om georgisk i kap.5 heller

Sidan eg ikkje har tilgang på ferdig setningssamanstilt georgisk-norsk parallelltekst, blir det vanskeleg å køyre den statistiske ordsamanstillinga som er vanleg som første steg i N-grambaserte metodar (utan ein god del forarbeid). Difor konsentrerer eg meg i evalueringa om eit testkorpus kor eg manuelt gjer ordsamanstillinga. Eg veit heller ikkje enno om nokon statistisk parsar av høg kvalitet for georgisk, men testkorpuset er ferdig parsa med LFG-parsaren frå Meurer (2008), c-strukturnodane avgrenser då kva som er ein syntaktisk konstituent.

Kapittel 2

Bakgrunn og omgrepsavklaring

2.1 SKRIV Relaterte metodar

Automatisk frasesamanstilling er eit nytt felt. Det finst allereie veldig gode system for automatisk setningssamanstilling, og automatisk samanstilling av ord har komme langt, men nivåa mellom ord og setning ser ut til å by på fleire problem. «by på fleire problem» – weasel wording, TODO omskriv. Dei ulike tilnærmingane som finst er prega av formåla til utviklarane. Det er verdt å merkje seg at ordet «frase» ofte blir nytta i litteraturen om strenger av ord (N-gram) som ikkje treng vere syntaktiske konstituentar, igjen avhengig av formålet med metoden.

Innanfor korpuslingvistikken har t.d. Piao & McEnery (2001) nytta enkel kollokasjonsinformasjon for å først finne sannsynlege nominale frasar på engelsk og kinesisk (dvs. «chunking»), og så samanstille desse; her er evalueringsgrunnlaget rett og slett ein manuell gjennomgang av dei mest sannsynlege omsetjingane dei får.

Den manuelle frasesamanstillinga i Samuelsson & Volk (2006), nemnt over, blei nytta som evalueringsstandard for den automatiske metoden i Samuelsson & Volk (2007). Her kjem frasesamanstillinga frå ei ordsamanstilling, der berre N-gram som svarer til ein syntaktisk node blir lenkja som frasar (meir om denne metoden nedanfor). Formålet er å lage ein parallell trebank, kor det altså er unyttig å lenkje «frasar» som *ikkje* er konstituentar.

fleire slike? meir om dette, algoritmen

Sjølv om fraselenkjer kan vere nyttige i korpuslingvistikken er det hovudsakleg innanfor statistisk maskinomsetjing at ein har forska på samanstilling av frasar. Koehn et al. (2003) gir ei grundig evaluering av ulike statistiske metodar for frasesamanstilling til bruk i stokastisk maskinomsetjing. Dei nyttar BLEU-skåren til å rangere resultata (Papineni et al., 2001, i Koehn et al., 2003, s. 51), som gir ei rangering ved (N-grambasert) samanlikning med ferdig omsett tekst.

Den første metoden, *AP*, er reint N-grambasert. Dei nyttar verktøyet Giza++ (Och og Ney, 2000, i Koehn et al., 2003, s. 50) til å indusere ordsamanstilling frå eit setningssamanstilt korpus (vha. «modell 4» for ordsamanstilling, utvikla ved IBM av Brown et al. (1993)). Denne samanstillinga er 1-til-n (t.d. eitt engelsk ord til to

franske), så dei finn ordsamanstilling for begge retningar og tek så snittet av alle moglege N-gramsamanstillingar som ikkje er i konflikt med ordsamanstillingane. Dei føyer så på ord frå unionen av desse vha. nokre enkle heuristikkar.

Den andre metoden, *Syn*, tek berre med dei frasane som står under syntaktiske nodar i eit parsa korpus; frasesamanstillinga til *Syn* er ein delmengd av den i *AP*. Denne syntaktisk informerte modellen gav ein mykje dårlegare BLEU-skåre enn den reint N-grambaserte modellen (faktisk dårlegare enn omsetjingane frå den opphavlege modell 4 for ordsamanstilling, utan frasesamanstilling). Dei forklarer dette med den store mengda uttrykk som ikkje utgjer syntaktiske konstituentar i følgje parsaren deira, men likevel konsekvent blir omsett til visse uttrykk på det andre språket (t.d. «es gibt» på tysk til «there is» på engelsk).

Seinare resultat har vist at ein *kombinasjon* av syntaktisk informerte metodar med reint N-grambaserte modellar (dvs. i motsetning til å berre fjerne samanstillingar mellom ikkje-konstituentar) kan auke skåren i ein maskinomsetjingsevaluering, både om ein som i *Syn*-modellen nyttar frasestrukturinformasjon, men i endå større grad om ein nyttar dependendsinformasjon (Tinsley et al., 2007; Hearne et al., 2008). Dette er interessant med tanke på at LFG-analysane gir begge typar informasjon.

Riezler & Maxwell (2006) utvikla ein metode for å kombinere frasebasert statistisk maskinomsetjing med LFG-basert setningsgenerering. Dei finn ei n-til-m-ordsamanstilling med Giza++ som i metodane over, men parsar i tillegg setningane i LFG. Dei to moglege f-strukturane som liknar mest blir valt ut, og frå ordsamanstillinga finn dei mange-til-mange-korrespondansar mellom substrukturane i f-strukturane. Ved å leggje til LFG-basert generering fekk det kombinerte systemet betre resultat på langdistanseavhengnader og generalisering til nye uttrykk med strukturell likskap til tidlegare observerte uttrykk.

Så langt har eg ikkje komme over metodar som går i motsett retning, altså prøver å finne eller betre på frase- og ordsamanstilling ut frå ein LFG-parse – det er dette som er strategien til programmet `lfgalign` i kapittel 4 – men det er stor overlapp mellom krava som kjem i kapittel 3 og dei gitt i den første publiseringa i XPar-prosjektet, Dyvik et al. (2009).

Dette er motsett retning av det mitt program gjer, nemne seinare?

2.2 SKRIV Eit kort oversyn over leksikalsk-funksjonell grammatikk og terminologi

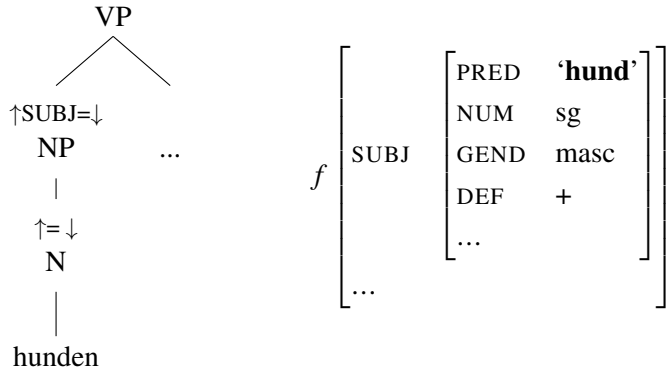
I dei følgjande kapitla nyttar eg ein del terminologi frå LFG, Leksikalsk-Funksjonell Grammatikk. Difor gir eg her eit kort oversyn over det som kan vere nytt for dei som er meir vand med andre grammatiske rammeverk, i tillegg til å avklare eit par eigne termar eg nyttar i teksta.

LFG er eit **modellteoretisk**, ikkje-derivasjonelt, rammeverk for grammatikk. Pullum & Scholz (2001) gir ein god gjennomgang av forskjellen mellom derivasjonelle (au kalla enumerative) grammatikkar og modellteoretiske grammatikkar. Derivasjonelle grammatikkar, som transformasjonsgrammatikkane til Chomsky, defi-

nerer eit språk som *ei mengd av uttrykk* ved avleiing frå eit startsymbol. Ein modellteoretisk grammatikk, derimot, gir skildringar av *enkeltuttrykk*, kor eitt uttrykk kan ha fleire moglege skildringar (språket er ikkje definert som ei mengd).

Ein modellteoretisk grammatikk kan i tillegg skildre strukturen (eller dei moglege strukturane) til *fragment* av setningar, og denne strukturen er lik det bidraget som fragmentet tilfører analysen av heile setninga. Det tilsvarande er ikkje mogleg å gjere derivasjonelt. Pullum & Scholz (2001, s. 32–33) gir t.d. eit fragment som kjem midt i eit høgreforgreina tre; ei derivasjonell skildring ville måtte skildre treet over eller under, men utan informasjon om kva som kjem til høgre eller venstre kan me ikkje (på ein ikkje-vilkårleg måte) skildre subtreet utanfor fragmentet heilt fram til terminal- eller startsymbol.

I LFG har analysane ulike *nivå*, *strukturar*. Konstituentforhold er skildra i **c-strukturen** («constituent structure»), medan forhold mellom syntaktiske funksjonar og grammatiske trekk kjem til syne i **f-strukturen** («functional structure»), ein trekkstruktur. Ein trekkstruktur er ei mengd attributt og verdiar, kor ein verdi kan vere atomær eller peike på ein ny trekkstruktur. Figur 2.1 illusterer eit enkelt døme for eit fragment.



Figur 2.1: Konstituentstruktur og funksjonell struktur

Konstituentstrukturen liknar på tradisjonelle frasetre, kor dominans mellom nodane viser frasehierarkiet i analysen av setninga. Men i tillegg har kvar node ei kopling til f-strukturen, via c-struktur-f-strukturavbildinga ϕ. Nodar i c-strukturen kan spesifisere informasjon på ulike stader i f-strukturen (me seier at nodane **projiserer** f-strukturar, eller deler av dei). I dette tilfellet går ϕ av VP her til f-strukturen f, VP projiserer f. NP-noden er annotert med ↑SUBJ=↓, dette les me som at «denne noden projiserer subjektet til ϕ av mornoden», altså projiserer NP-en SUBJ av f. NP er ikkje åleine om å gjere dette, N-noden har ↑=↓ som vil seie at N projiserer same f-struktur som NP. Dette subjektet har fleire trekk i f-strukturen, t.d. $\text{NUM}_{\{\}}$ og $\text{GEND}_{\{\}}$ som har atomære verdiar og seier at dette er i eintal og maskulinum. Viss eit anna ord i setninga må samsvare med dette for å vere grammatisk, kan me krevje i grammatikken at me kan **unifisere** visse trekk; for atomære trekk som dette kan me alltid unifisere dei viss atomet er formmessig likt. Me kan au unifisere heile

trekkstrukturar så lenge dei ikkje har trekk som ikkje kan unifiserast; dei unifiserte strukturane er då blitt *ein* struktur, og alle referansar til dei to peiker no på same struktur.

PRED er eit spesielt trekk, verdien her er ein *semantisk form*. Desse er alltid *unike*, og kan ikkje unifiserast sjølv om dei har lik form.

endosentrisitetsprinsippa ...

$\overline{\text{X}}$ ...

diskontinuerlege konstituentar ...

ϕ c-struktur-f-strukturavbildinga ϕ ...

ϕ^{-1} Det funksjonelle domenet til ein f-struktur er gitt ved ϕ^{-1}, inversen av c-til-f-strukturavbildinga, og tilsvarer dei nodane i c-strukturen som projiserer denne f-strukturen, t.d. ein VP-node med dominerande IP og CP (Bresnan, 2001, s. 126). Sidan dette er inversen av ein funksjon, kan me ha diskontinuerlege konstituentar i same funksjonelle domene (på same måte som ulike argument til ein funksjon kan gi same verdi).

fraselenkjer vs frasesamanstilling Eg nyttar her termane *lenkjing* og *samanstilling* i omtrent same tyding som dei engelske termane *link* og *alignment*, kor ei samanstilling er ei mengd lenkjer. Merk at ei enkeltlenkje treng ikkje å vere ein-til-ein. Lenkjer og samanstillingar er ekvivalensforhold som me kan finne mellom lingvistiske *representasjonar* (f-struktur, c-struktur) eller *uttrykk* (ord, setningar). Lenkjing mellom dei siste er meir ateoretisk / datanært – grunnlaget for å opprette ei lenkje mellom to c-strukturnodar er at uttrykka i kontekst som dei representerer er omsetjingar, og har lik nok syntaks (i følgje dei to grammatiske analysane) til at me kan lenkje nodane.

Kapittel 3

Krav til frasesamanstilling

3.1 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 4. 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.

3.2 Formål med frasesamanstilling

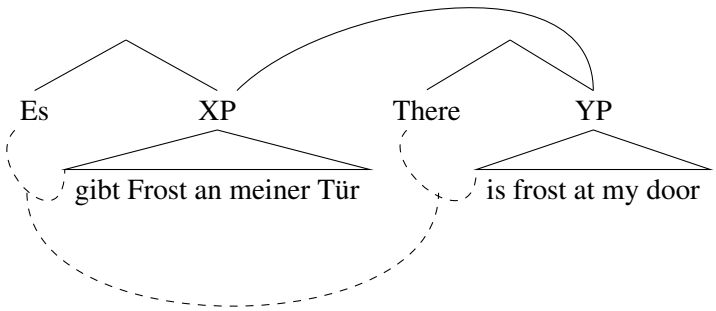
Ei frasesamanstilling er ein slag annotasjon av eit korpus. På same måte som oppbygginga av eit korpus avheng av formålet til korpuset, kan ein ikkje definere den ideelle annotasjonen av eit korpus utan å ta høgd for kva ein skal nytte annotasjonen til.

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 3.1 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 3.1: 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 3.5 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.

3.3 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 2.2 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 3.4 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).

3.4 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 3.1 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).

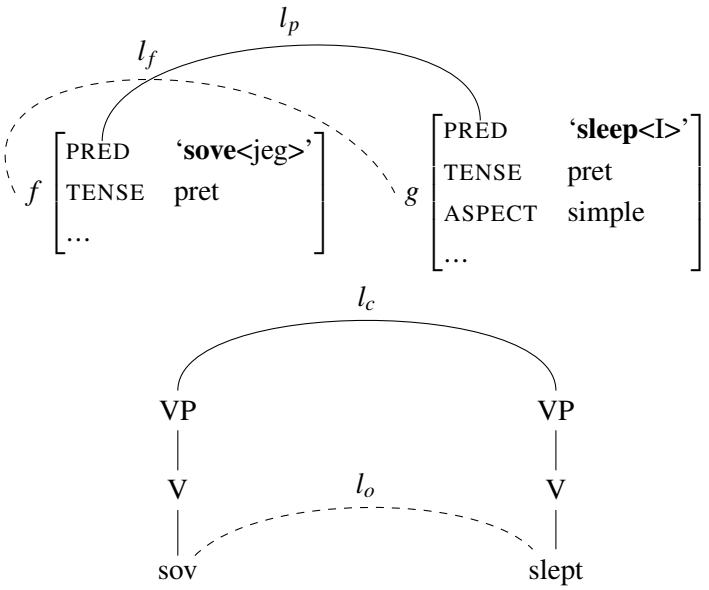
Samtidig er det au eit omsetjingsforhold mellom trekka i same f-struktur som dei lenkja PRED-elementa, og me ville kanskje ikkje ha omsett dei to PRED-elementa i andre f-strukturkontekstar. Difor bør me au sjå på ei PRED-lenkje som

[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 5.

[7] I del 3.7.3 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 2.2), får me følgjande samanheng, illustrert i figur 3.2:

(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 3.2: 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 3.2 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 3.7 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 3.7.3.

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 3.5) er oppfylte for å lenkje to PRED-element. TODO: litt brå avslutning

3.5 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 3.6.1 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 3.2) 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].

3.5.1 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.

3.6 Krav på f-strukturnivå

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

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.

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

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.

3.6.1 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 3.5 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 når me prøver å lenkje, men argument-argument-lenkjer bør prioriterast viss alt anna er likt.

3.6.2 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

$$\left[\text{PRED} \quad \text{‘}\textbf{gefallen}\text{<Tonfall, ich}_i\text{>’ ...}\right]$$

$$\leftrightarrow$$

[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.

3.6.3 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 3.8. 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 3.6.5.

3.6.4 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.

3.6.5 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.

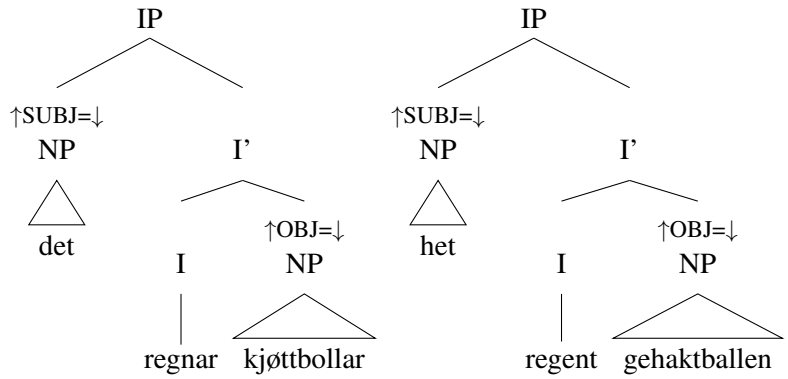
3.7 Krav på c-strukturnivå

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.



Figur 3.3: Enkel lenkjing av c-strukturnodar mellom norsk og nederlandsk; IP til IP, I' til I' og I til I.

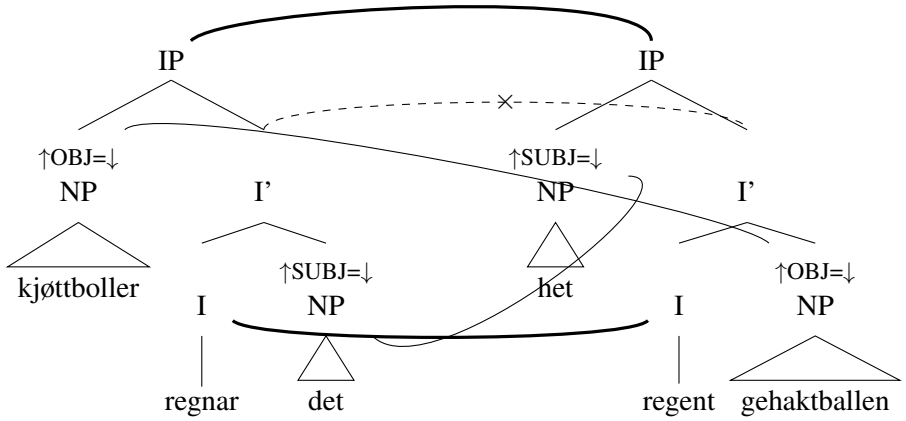
I figur 3.3 er dei funksjonelle domena til *regnar/regent* lenkja[16], og det same med *det/het* og *kjøttbollar/gehaktballen*. Viss me føreset at subjekt-NP-ane er lenkja med kvarandre, og at objekt-NP-ane er lenkja med kvarandre, på c-strukturnivå, vil det vere ønskeleg å ein-ein-lenkje IP-nodane, I'-nodane og I-nodane. Me skal 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 3.3. 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 3.4 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 3.4. (I 3.3, 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.

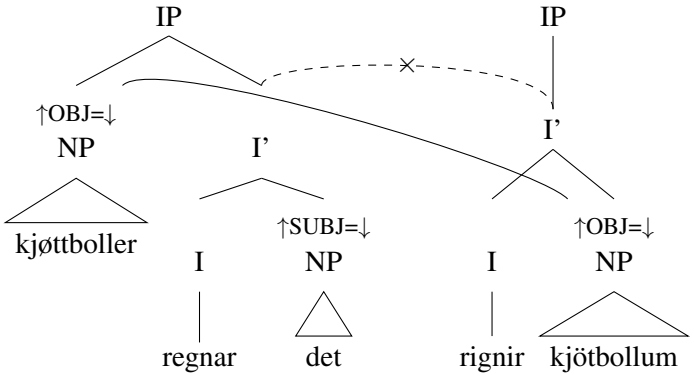
Sjølv om subjektet sto ulenkja, t.d. ved lenkjing inn i eit pro-drop-språk eller liknande, ville me fått same situasjon; I'-nodane i figur 3.5 kan ikkje lenkjast sidan I' på islandsk dominerer objektet, medan I' på norsk ikkje gjer dette, og objekta er lenkja med kvarandre (her både på c- og f-strukturnivå). Ei lenkje mellom desse

[16] I desse trea har eg annotert f-strukturforhold på visse nodar; der eg ikkje har teikna inn dette gjeld ↑=↓, altså at noden er i same funksjonelle domene som mornoden. Eg har i tillegg forenkla kategorinamna ein del frå dei som kjem direkte frå ParGram/XPar-analysane, det bør ikkje ha noko å seie for framstillinga her.



Figur 3.4: C-strukturlenkjer kan ikkje gå på tvers av dominerte lenkjer (nynorsk og nederlandsk)

I'-nodane ville sagt at dei dominerer korresponderande materiale, men det gjer dei ikkje.



Figur 3.5: C-strukturlenkjer kan ikkje gå på tvers av dominerte lenkjer (nynorsk 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 3.4 og figur 3.5 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 3.6, 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 sidan dei to toppnodane dominerer materiale som inneheld ulike lenkjer på f-strukturnivå – ei slik c-strukturlenkje ville stått i konflikt med f-strukturlenkjene. Intuitivt synest det au feil med ei lenkje mellom konstituentane *det regner* og *it rained meatballs*. Dei kan iallfall ikkje reknast som omsetjingar av kvarandre åleine; i ein større kontekst kan dei inngå i ein korrespondanse, men denne større konteksten har me jo lenkja allereie ved IP-nodane.

I det minste bør me difor krevje følgjande av lenkjer på c-strukturnivå:

(13) Ein node n_s kan lenkjast med ein node n_t berre viss:

a. $\phi(n_s)$ er lenkja på f-strukturnivå med $\phi(n_t)$, og
b. det ikkje finst nodar under n_s som er lenkja med nodar utanfor det funksjonelle domenet til n_t, og
c. det ikkje finst nodar under n_t som er lenkja med nodar utanfor det funksjonelle domenet til n_t.

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.

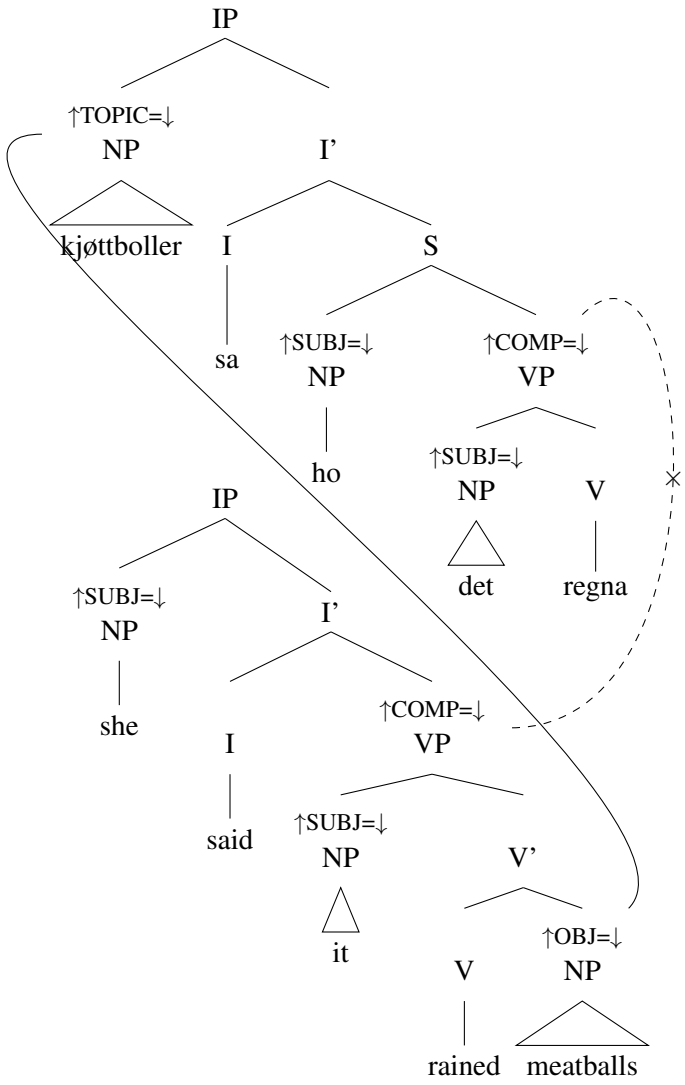
3.7.1 Lenkja f-strukturar utan c-strukturnodar

I figur 3.7 kan iallfall IP-nodane lenkjast, dei dominerer alle orda på begge setningane, og f-strukturane er lenkja. Men NP-subjektet på den norske sida, er ikkje lenkja med noko i det georgiske treet; dette subjektet er lenkja med eit pro-element på f-strukturnivå. Den informasjonen (her reint syntaktisk) som ordet *det* tilfører IP, ligg under I' på georgisk. Ved I-nodane manglar det norske treet i tillegg den informasjonen som *seg* tilfører.

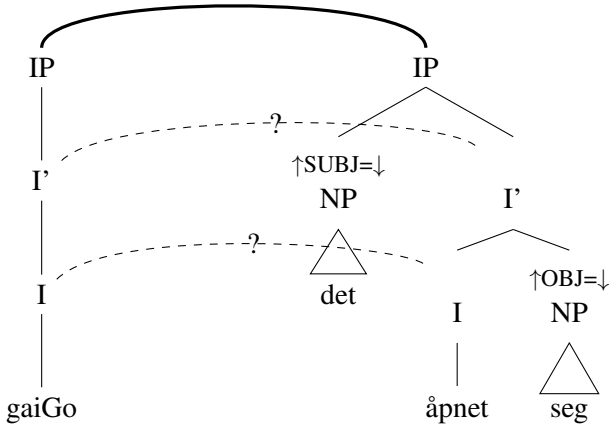
Hadde det georgiske treet hatt spesifikator og komplement som kunne lenkjast til spesifikator og komplement på norsk, ville det ha vore uproblematisk å lenkje I' og I. Men om me berre har krav (13) å halde oss til, er det uspesifisert kva me skal gjere i ein situasjon kor nodar lenkja på f-strukturnivå ikkje er lenkja på c-strukturnivå.

Det finst (iallfall) to alternativ.

Det eine alternativet er å seie seie at I- og I'-nodane ikkje skal lenkjast, sidan *det* og *seg* er lenkja på f-strukturnivå (til subjekt og objekt av gaiGo), då tolker me det slik at I' og IP dominerer ulikt lenkja materiale. Det at det *ikkje* finst ei lenkje



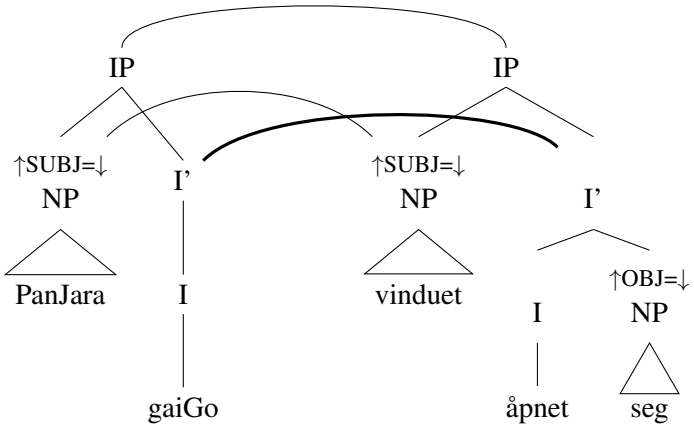
Figur 3.6: Sjølv toppnodane i eit funksjonelt domene kan stå ulenkja; her kan ikkje VP-nodane lenkjast sidan det norske TOPIC er objektet til *regna*, lenkja til objektet under VP på engelsk



Figur 3.7: Skal ulenkja søsternodar hindre lenkjing? (Georgisk og bokmål)

mellom I'-nodane, men mellom IP-nodane, vil då opplyse oss om at I'-nodane dominerer ulike f-strukturlenkja informasjonstilskot på dei ulike språka; likeins for I-nodane. Eg kjem tilbake til korleis ein kan formalisere dette kravet i del 3.7.2.

Det andre alternativet er å ikkje gjere forskjell på IP, I' og I når det gjeld c-strukturlenkjinga. Grunnen til å gjere dette er at *gaiGo* både korresponderer med heile frasen *det åpnet seg*, men au med berre *åpnet seg*. I figur 3.8 ser me t.d. at I'-nodane kan lenkjast (utan å sjå på anna enn krav (13)), det vil altså vere mogleg å lenkje I'-nodane i andre omgivnader. Det finst ein slags dobbeltheit mellom korrespondansen *gaiGo-det åpnet seg* og korrespondansen *gaiGo-åpnet seg* og me kan uttrykkje dette ved å ikkje gjere forskjell på IP og I' i figur 3.7 (**?**).har eg forstått dette rett? (korleis er dette ein korrespondanse på tokennivå?)



Figur 3.8: Delvis mogleg lenkjing av underordna c-strukturnodar mellom georgisk og bokmål

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.

Dette kravet gjer at ein ikkje treng krav (13), og vil gi ei mange-mange-lenkje mellom alle nodane i dei to funksjonelle domena til *gaiGo-åpnet* i figur 3.7. Viss me skriv ei f-strukturlenkje som eit ordna par mellom PRED-verdien på kjeldesida (georgisk, med subskript $_s$) og PRED-verdien på målsida (norsk, med subskript $_t$) får me $LL(IP_s) = LL(I'_s) = LL(I_s) = \{(\textbf{ga-Geba}, \textbf{åpne})\} = LL(IP_t) = LL(I'_t) = LL(I_t)$ kor *det* og *seg* er ulenkja på både c-strukturnivå og ordnivå[17].

TODO: diskutere litt meir forskjellane på desse alternativa

3.7.2 Eit strengare lenkjingskriterium

Sidan det er mogleg å ønskje seg å ikkje lenkje I'- og I-nodane i 3.7, gir eg her ein måte å formalisere dette på.

For å tillate lenkjene i figur 3.3, men ikkje dei stipla lenkjene i figur 3.7, ville det vore nok å krevje at søsternodane var lenkja. I figur 3.3 kan I-nodane lenkjast fordi objekta er lenkja, I'-nodane fordi subjekta er lenkja. I figur 3.7 kan dei norske I'- og I-nodane ikkje lenkjast med noko fordi søstrene deira ikkje er lenkja. Men dette blir for strengt. Det kan t.d. vere gode uavhengige grunnar til å ha ein mellomliggande S-node før objektet på norsk, kor S er i same funksjonelle domene som IP, medan det kanskje finst uavhengige grunnar for å *ikkje* gjere dette på andre språk. Figur 3.9 demonstrerer denne situasjonen. Her kan ikkje S lenkjast til objektet sidan dei ikkje er i same funksjonelle domene, men me vil jo likevel lenkje I-nodane; så eit krav om lenkja søsternodar blir for strengt.

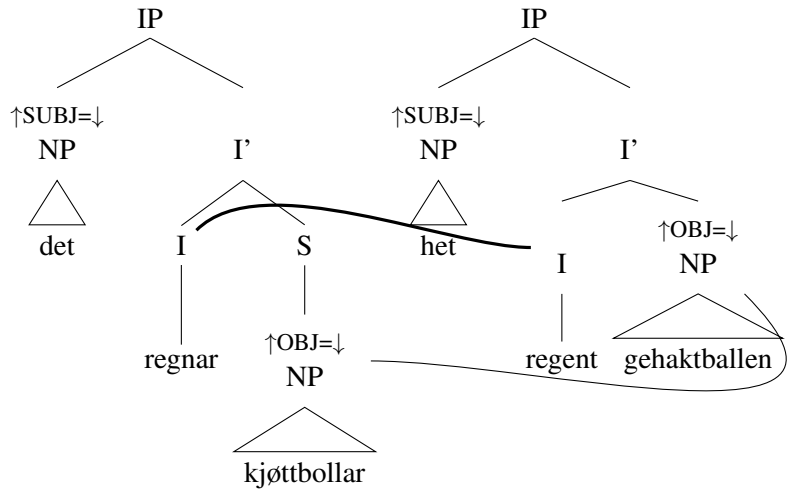
Me treng altså eit litt meir nyansert krav. Som nemnt i fotnote 17 går det an å få til dette ved ein kombinasjon av konseptet om lenkja leksikalske nodar og å krevje at orda *det*, *åpnet* og *seg* i figur 3.7 er mange-mange-lenkja på ordnivå til *gaiGo*, sidan dei er lenkja til subjekt, predikat og objekt av *gaiGo* på f-strukturnivå.

Men viss me vil unngå å referere til ordlenkjer, går det au an å definere kravet i form av f-strukturlenkjer på preterminale nodar[18]:

(14) For å lenkje c-strukturnodane n_s og n_t:

[17] Merk at orda *det* og *seg* måtte definerast som ulenkja for at denne definisjonen skulle fungere, noko som krev at ein nyanserer krav (1-b) litt. Viss me hadde definert *det* og *seg* som mange-ein-lenkja (med *åpnet*) inn i *gaiGo*, ville me fått same resultat som det krav (14) i del 3.7.2 gir. Viss georgisk er kjeldespråket (n_s, norsk: n_t) blir $LL(IP_s) = LL(I'_s) = LL(I_s) = \{(\text{gaiGo}, \text{det}), (\text{gaiGo}, \text{åpnet}), (\text{gaiGo}, \text{seg})\} = LL(IP_t)$. Mengdene $LL(I'_t) = \{(\text{gaiGo}, \text{åpnet}), (\text{gaiGo}, \text{det})\}$ og $LL(I_t) = \{(\text{gaiGo}, \text{åpnet})\}$ på den norske sida har då ikkje korresponderande mengder på georgisk og blir ikkje lenkja.

[18] 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 I_t.



Figur 3.9: I-nodane bør lenkjast sjølv om søsternodane ikkje er lenkja (norsk og nederlandsk)

> La $l_c(f)$ vere mengda som inneheld f-strukturlenkja til f, *og* f-strukturlenkjene til alle argument a av f som ikkje har c-strukturnodar, dvs. kor $\phi^{-1}(a) = \emptyset$. La $L_c(n)$ vere mengda av $l_c(\phi(n'))$ for alle f-strukturlenkja preterminale n' som er dominert av n. n_s og n_t kan lenkjast om $L_c(n_s) = L_c(n_t)$.

I figur 3.8 har me då følgjande situasjon:

$L_c(IP_s) = \{(\textbf{PanJara}, \textbf{vindu}), (\textbf{ga-Geba}, \textbf{åpne}), (\textbf{pro}, \textbf{seg})\} = L_c(IP_t)$
$L_c(I'_s) = \{(\textbf{ga-Geba}, \textbf{åpne}), (\textbf{pro}, \textbf{seg})\} = L_c(I'_t)$
$L_c(I_s) = \{(\textbf{ga-Geba}, \textbf{åpne}), (\textbf{pro}, \textbf{seg})\} \neq \{(\textbf{ga-Geba}, \textbf{åpne})\} = L_c(I_t)$

Dette vil seie at krav (14) gir lenkjer mellom IP-nodane og I'-nodane, men ikkje mellom I-nodane. I figur 3.7 vil ikkje ein gong I'-nodane få ei lenkje, sidan den norske I'-node dominerer $\{(\textbf{ga-Geba}, \textbf{åpne}), (\textbf{pro}, \textbf{seg})\}$ medan den georgiske I'-node dominerer $\{(\textbf{pro}, \textbf{det}), (\textbf{ga-Geba}, \textbf{åpne}), (\textbf{pro}, \textbf{seg})\}$, det same som IP-nodane.

Merk at om me omdefinerer $l_c(f)$ til å ikkje innehalde f-strukturlenkjer til argument av a, vil krav (14) gi same c-strukturlenkjer som kravet frå Dyvik et al. (2009), men definert i form av f-strukturlenkjer på preterminale nodar.

3.7.3 Funksjonelle c-strukturnodar

Ikkje alle ord tilsvarer PRED-element i f-strukturen, dette gjeld typisk funksjonsord (t.d. *som*, *at*). ...og desse vil me lenkje kvifor? i kva situasjonar? TODO diskuter. 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 krava nemnt over 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 som dominerer same mengd med lenkjer.

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 nok ikkje ein gong lenkje komplementa, sidan funksjonsorda då gjer at komplementa speler ulike roller i omgivnadene[19]. 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 andre f-strukturelement enn PRED i denne oppgåva).

Me kan krevje 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 3.6.1.

Desse ønskene kan me formalisere slik:

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

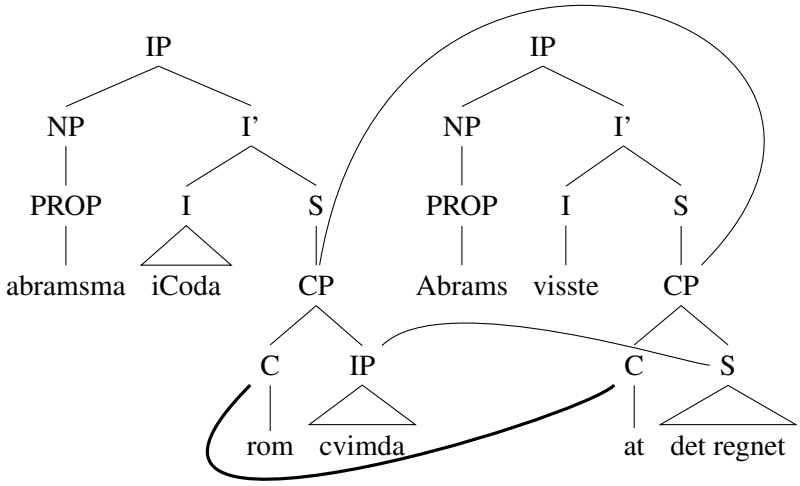
a. Gitt ei mogleg lenkjing av FP og GP, kor F og G er funksjonelle kategoriar der komplementa elles kan lenkjast, tolk LPT-korrespondansen mellom orda under F' og G' som eit medlem av lenkjemengda L_c (evt. *LL*), kor denne må vere lik for at FP og GP skal kunne lenkjast, då kan me au lenkje F' og G'.

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 nokon forskjell i L_c (evt. *LL*) mellom FP og F', er F' medlem av nodemengden som blir lenkja til XP.

Om (15-a) er oppfylt, kan me få samanstillinga vist i figur 3.10. Her vil dei funksjonelle domena til CP og CP kvar kunne delast opp i to deler, kor den funksjonelle delen har LPT-korrespondanse medan komplementa er lenkja på f-strukturnivå. Lenkjemengdene under CP-nodane er like, og dei under C-nodane er like.

[19] 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 kunne lenkjast viss dei andre krava er oppfylte.

(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 (14) er dekkja for S-nodane med unntaket over.)



Figur 3.10: Mogleg samanstilling av funksjonelle c-strukturnodar mellom georgisk og norsk (bokmål)

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 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.

3.8 SKRIV Rangering

(meir om dette i del 4.2)

3.9 SKRIV Oppsummering av krava

Kapittel 4

Implementasjonen av `lfgalign`

For å finne ut av kor godt krava i forrige kapittel fungerer til å avgrense kva for lenkjer som er moglege, har eg implementert dei etter beste evne i eit Lisp[1]-program.

intro TODO, kanskje noko om kva eg faktisk har fått ut av implementasjonen

Ei implementering gjer det svært synleg om det finst manglar i eit formelt krav, eller om noko ikkje er godt nok spesifisert.

Programmet `lfgalign`[2] tek inn LFG-analysane av to setningar som me av uavhengige grunnar trur er omsetjingar av kvarandre. LFG-analysane må vere disambiguerte og i Prolog-formatet frå XLE[3]. Programmet les inn dei to filene og opprettar ein intern representasjon av LFG-analysen.

Me kan i tillegg gi programmet informasjon om kva for ord-omsetjingar me ser på som lingvistisk prediktable. Intensjonen er at dette kan vere informert av omsetjingstabellen frå eit automatisk ordsamanstillingsprogram, eller av handskrivne omsetjingsordbøker.

treng eg ein eigen del om LPT i dette kapittelet? Implementasjonen er jo veldig enkel iallfall.

Programmet byrjar lenkjinga med f-strukturane. Ei f-struktur*samanstilling* er ei mengd med *lenkjer* mellom individuelle f-strukturar. Resultatet av lenkjinga på dette nivået kan vere tvitydig: sidan det ofte finst fleire måtar å lenkje argument og adjunkt på, får me i første omgang mange samanstillingar mellom kjelde- og mål-f-strukturar.

Difor rangerer me f-struktursamanstillingane, og den beste sender me vidare til c-struktursamanstillinga. Denne delen av programmet gir ut éi, utvitydig mengd med mange-til-mange-lenkjer mellom c-strukturane (her treng me ingen rangering). Nodane i kvar av desse mange-til-mange-lenkjene definerer no den endelege frasesamanstillinga.

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

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

[3] 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.

Nedanfor går eg gjennom detaljane rundt dei relevante delene av programmet.

4.1 Lenkjer mellom f-strukturar

Hovudalgoritmen for lenkjing mellom f-strukturar er vist i kodefigur 1. Funksjonen `f-align` returnerer ei mengd med moglege samanstillingar. Kvar samanstilling er ei mengd med par av f-strukturar[4]. Eit par (F_s, F_t) representerer ei lenkje frå ein f-struktur på kjeldespråket, til ein f-struktur på målspråket. Me går ut frå at dette paret har LPT-korrespondanse[5], dette blir sjekka før alle kall på `f-align`. Der me ikkje har informasjon om LPT-korrespondanse mellom to ord (orda er ukjende), er lenkjing lov. Pro-element og substantiv kan alltid lenkjast med kvarandre.

TOGROK: kan me i tillegg sjekke om funksjonelle nodar i same f-domene som PRED-ordformen har LPT-korrespondanse?

Hjelpefunksjonen `argalign` (som igjen kallar `argalign-p`, vist i kodefigur 2) gir alle moglege «argumentpermutasjonar», dvs. moglege kombinasjonar av lenkjer mellom argumenta til F_s og F_t som tilfredsstiller kravet om LPT-korrespondanse, men utan å sjekke at desse argumenta igjen kan samanstillast. Funksjonen prøver å lenkje kvart argument til eit argument eller eit adjunkt, men gir ingen lenkjer mellom to adjunkt (sjå del 4.1.1 nedanfor om dette). Funksjonen gir heller ikkje kombinasjonar der minst eitt argument ikkje er lenkja – alle kombinasjonane må inkludere alle argument frå F_s og F_t, jf. krav (iii) og (iv) i Dyvik et al. (2009, s. 75). Elles er krav (i) er tautologisk oppfylt, medan me som nemnt føreset at krav (ii) er oppfylt før alle kall på `f-align`.

Eit døme: viss F_s har argumenta SUBJ og OBJ og ingen adjunkt, og F_t har argumentet SUBJ og eitt adjunkt ADJ, der alle ord-omsetjingar er moglege, vil `argalign` gi dei to samanstillingane $\{(\text{SUBJ}, \text{SUBJ}), (\text{OBJ}, \text{ADJ})\}$ og $\{(\text{SUBJ}, \text{ADJ}), (\text{OBJ}, \text{SUBJ})\}$. Viss adjunktet til F_t ikkje fantest, eller ikkje hadde LPT-korrespondanse med nokon av argumenta til F_s, ville me ikkje fått nokon samanstillingar; medan viss paret $(\text{SUBJ}, \text{SUBJ})$ ikkje hadde LPT-korrespondanse og alt anna var likt, ville me berre fått den siste samanstillinga.

Funksjonen `f-align` går så gjennom kvar lenkje i kvar argumentpermutasjon, og prøver å kalle `f-align` på alle lenkjene. Sidan lenkjene som `argalign` gir har LPT-korrespondanse, vil alle f-strukturane i dei rekursive kalla i `f-align` ha LPT-korrespondanse. Eit rekursivt kall kan gi nye samanstillingar i dei indre f-strukturane, viss dei relevante krava er oppfylte.

Det er mogleg at ei lenkje frå éi samanstilling kan finnast i andre samanstillingar, me unngår dobbeltarbeid ved å lagre alle delvise samanstillingar i tabellen *aligntable*. Dette føreset at `f-align`(s,t) er uavhengig av konteksten rundt; t.d. må mengda av samanstillingar som kjem ved å lenkje subjektet til F_s mot subjektet

[4] 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.

[5] 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.

til F_t vere uavhengig av om objektet til F_s er lenkja mot eit objekt eller eit adjunkt osb. av F_t.

TODO: nemne føresetnaden om uavhengnad i kapittel 3

```
alignments ← ∅ ;
forall the argperm in argalign(F_s, F_t) do
    p ← ∅ ;
    forall the A_s, A_t in argperm do
        if not(aligntable[A_s,A_t]) then
            aligntable[A_s,A_t] ← f-align(A_s, A_t);
        if aligntable[A_s,A_t] then
            add aligntable[A_s,A_t] to p;
        else
            add (A_s,A_t) to p
    end
    add p to alignments ;
    forall the adjperm in adjalign(argperm, F_s, F_t) do
        a ← copy-of(p) ;                    // optional adjunct links
        forall the A_s, A_t in adjperm do
            if not(aligntable[A_s,A_t]) then
                aligntable[A_s,A_t] ← f-align(A_s, A_t);
            if aligntable[A_s,A_t] then
                add aligntable[A_s,A_t] to a;
            else
                add (A_s,A_t) to a
        end
        add a to alignments ;
    end
end
 // loop through adjalign if no arguments exist
if alignments = ∅ then return ∅ ;  // Fail
else return ((F_s,F_t),alignments) ;
```

Funksjon 1: f-align(F_s, F_t)

Sjølv om det er krav om LPT-korrespondanse mellom kvart argument og eit argument/adjunkt for å lenkje F_s og F_t, er det ikkje noko krav om at alle para i ein argumentpermutasjon tilfredsstiller alle lenkjingskrava. Viss `f-align`(OBJ, ADJ) frå dømet over gir null, og ikkje kan lenkjast (t.d. fordi ADJ hadde eitt argument, og OBJ ingen argument/adjunkt), medan `f-align`(SUBJ, SUBJ) kan lenkjast, vil `f-align` likevel returnere samanstillinga som inneheld (OBJ, ADJ) og (SUBJ, SUBJ). Me kan sjå i *aligntable* for å finne ut av om kvar av f-strukturane kunne lenkjast; i dette tilfellet vil *aligntable*[OBJ, ADJ] vere tom.

Om me i tillegg krev at substrukturar kan samanstillast kan me utelukke len-

TODO: forskjellen mellom LPT-krav og rekursjons-krav på argument må inn i kapittel 3

usage: Kalt av argalign slik:
argalign-p(arguments(F_s), adjuncts(F_s), arguments(F_t), adjuncts(F_t))

$a \leftarrow \emptyset$;
if $args_s$ **then**
  $s \in args_s$;
  forall the $t \in args_t$ ***where*** *LPT(s,t)* **do**
    forall the $p \in$ *argalign-p*($args_s - \{s\}$, $adjs_s$, $args_t - \{t\}$,$adjs_t$) **do**
    add $\{(s,t)\} \bigcup p$ to a;
  end
  forall the $t \in adjs_t$ ***where*** *LPT(s,t)* **do**
    forall the $p \in$ *argalign-p*($args_s - \{s\}$, $adjs_s$, $args_t$,$adjs_t - \{t\}$) **do**
    add $\{(s,t)\} \bigcup p$ to a;
  end
  return a;
else if $args_t$ **then**
  if $adjs_s$ **then**
    $s \in adjs_s$;
    forall the $t \in args_t$ ***where*** *LPT(s,t)* **do**
      forall the $p \in$ *argalign-p*($args_s$, $adjs_s - \{s\}$, $args_t - \{t\}$,$adjs_t$) **do** add $\{(s,t)\} \bigcup p$ to a;
    end
    return a;
  else
    return $\emptyset$; `// Fail`

else
  return $\{\emptyset\}$; `// End`

Funksjon 2: argalign-p($args_s$, $adjs_s$, $args_t$, $adjs_t$)

kjing av f-strukturane F_s og F_t i (1) under:

(1) a. F_s [PRED '**planlegge**<*eg*,[1:*gi*]>' XCOMP $_1$[PRED '*gi (opp)*']]

b. F_t [PRED '**plan**<*I*,[2:*give*]>' XCOMP $_2$[PRED **give**<*I*,*him*,*it*>']]

Men det kan vere at me ikkje *vil* krevje dette i alle moglege tilfelle. Ei tryggare løysing er å rangere ulike løysingar i etterkant, ved å spørje etter dei argumentsamanstillingane som har flest medlem i *aligntable*, dette kjem eg tilbake til i 4.2 nedanfor.

4.1.1 Overflødige adverbial

Argumentpermutasjonane frå `argalign` prøver som nemnt ikkje reine adjunkt-adjunkt-lenkjer, sidan me ikkje vil forkaste lenkjing av F_s og F_t berre på grunn av at ikkje alle adjunkt kunne lenkjast. Men når me har prøvd ein argumentpermutasjon, kan me lage ein kopi av denne som i tillegg inneheld lenkjer mellom «overflødige» adverbial, altså dei adjunkt-adjunkt-lenkjene som `argalign` ikkje prøver. Hjelpefunksjonen `adjalign` (ikkje vist her TODO: implementere :->) konstruerer moglege permutasjonar av lenkjer mellom adjunkt som ikkje er inkludert i *argperm*, og `f-align` prøver desse rekursivt på same måte som med argumentlenkjene. Lenkjene blir lagt til ein *kopi* av argumentpermutasjonane, sidan det ikkje er sikkert at me ønskjer å lenkje alle adjunktdøtre. Viss me har to overflødige adjunkt på kvar side, og kravet om LPT-korrespondanse er dekkja for alle fire moglege par, får me seks moglege permutasjonar, sidan me inkluderer dei fire permutasjonane der eitt adjunktpar er ulenkja.

og så er det spørsmålet om me kan lenkje adjunkt på ulike nivå i f-strukturane

Viss F_s og F_t ikkje hadde argument i det heile teke, går me au gjennom moglege permutasjonar av adjunktdøtre, på same måte (ikkje vist i kodefigur 1).

4.1.2 SKRIV Funksjonsord utan LPT-korrespondanse bør eigentleg hindre f-strukturlenkjing

4.1.3 SKRIV Når f-lenkjene ikkje er 1-1

kausativ

preposisjonsobjekt

Dette må 1. spesifiserast (kap.3), og 2. implementerast...

"sigaretten" og "sigaretze" er ikkje på same nivå i dei respektive f-strukturane nedanfor:

```
0[ PRED vedde<28,29,27,30>
   29[ PRED sigarett<> ] ]
```

```
0[ PRED da-najleveba<37,10,46>
   ADJUNCT { 2 }
   2[ ze<5>
      OBJ 5[ sigareti ] ] ]
```

Men dette er fordi sigaretze er adjunkt, og i følgje fotnote 3 i Dyvik et al. (2009) er P(ADJ) lik den semantiske formen til preposisjonsobjektet til ADJ. (Er dette heilt enkelt? Kan me ha adjunkt med fleire «preposisjonsargument»?)

I motsetning til å endre på P(ADJ) slik at pred-verdien til eit adjunkt hoppar over preposisjonar og inn i preposisjonsobjekt, har eg i implementasjonen min gjort det slik at funksjonen som henter ut *f-strukturen* til adjunkt-døtre av ein f-struktur hoppar over preposisjonsobjekt (m.a. fordi ved eit kall P(X) kan ikkje P vite om X er eit adjunkt eller argument).

Sjå au del **??**.

4.1.4 Kan me gjere f-struktursamanstillinga bottom-up?

Ein alternativ metode for lenkjing av f-strukturane er å byrje med alle logisk moglege permutasjonar av LPT-korrespondansar, og så sile ut dei som ikkje svarer til krava. Prosessen ville nok blitt mykje meir oversiktleg på denne måten, sidan det då berre er snakk om å sjekke krav for kvar enkelt lenkje. Men ein slik metode er vanskeleg i praksis; når avskjeringa skjer så seint, blir det alt for mange moglege kombinasjonar for lengre setningar med mange ukjende ord til at ein vanleg datamaskin kan halde styr på dei.

Me må i alle tilfelle vere klar for ei setning der alle ord er ukjende (me har ingen informasjon om LPT-korrespondanse), slik at kvart kjeldeord kan lenkjast til kvart målord. Viss begge setningane er 4 ord, får me 16 moglege samanstillingar der alle ord er med i nøyaktig éi lenkje (2^l, kor l er setningslengd). Men ofte har me null-lenkjer, me må altså i tillegg tillate samanstillingar der minst eitt ord er ulenkja, utan at me treng å vite kva for ord det er; med desse kortare listene inkludert får me endå fleire moglege samanstillingar per setning (4 ord gir 26, 8 ord gir 2186 moglege samanstillingar). Sjølv om me heile tida vel dei samanstillingane som lenkjar flest ord, ville maskinen raskt fått problem. I tillegg har me problemet med 1-mange-lenkjer, som skaper endå fleire moglege samanstillingar.

Ein sideverknad av å byrje med ytre lenkjer og gå innover (prosessen skildra i del 4.1) er at me automatisk unngår å prøve «kryssande» lenkjer, t.d. å lenkje F_s med XCOMP av F_t, og XCOMP av F_s med F_t (denne kombinasjonen av lenkjer vil jo vere ein del av alle logisk moglege permutasjonar). Me får au prioritert å lenkje ytre element, som jo er sikrare lenkjer: gitt to f-strukturar for setningar der alt me veit om lenkjinga er at *setningane* er omsetjingar av kvarandre, vil dei to ytre f-strukturane ha størst sjanse for å korrespondere med kvarandre. For kvart steg du går innover må du multiplisere inn sjansen for å trå feil i argumentpermutasjonane.

4.2 SKRIV Rangering

Rangering er ikkje implementert enno (akkurat no gir rank(f-alignments) berre ut første samanstilling.) Rangering foregår etter ulike kriterium. Her er eit par forslag:

4.2.1 rekursivt lenkja > ulenkja, men LPT

aligntable seier om noko er rekursivt lenkja eller ikkje, plusspoeng viss me har klart å lenkje rekursivt.

4.2.2 argument-argument > argument-adjunkt

Plusspoeng for argument-argument-lenkjer, burde vere eit bra kriterium, men me får sjølvsagt problem viss LPT ikkje seier noko i døme (5) med
`da-najleveba<Abrams,Browne,regne>` adjunkt: `sigarett`
`bet<Abrams,sigarett,regne>` adjunkt: `Browne`

4.2.3 arg1-arg1 arg2-arg2 > arg1-arg2 arg2-arg1 (følgje)

Dette kjem til å gi problem når me vil lenkje «behage» og «like», viss me ikkje har motstridande LPT-informasjon (og argumentfølgje i leksikon ikkje er basert på semantikk, men syntaks). Men elles er det vel OK.

Enklaste implementasjon: Levenshtein-avstand. Men burde visse argument vektast? (T.d. vekte subjekt om alt anna er likt.)

Andre forslag: http://en.wikipedia.org/wiki/Edit_distance)

4.2.4 flest lenkja adjunkt

Usikker på dette... avheng av om me tillèt lenkjer på tvers av f-strukturar.

4.2.5 Prioritet på rangeringskriterium

Dette bør sjølvsagt testast empirisk, blir kanskje utanfor denne oppgåva (diskusjonsdel?), men kan jo prøve meg litt rundt.

4.3 Lenkjing av c-strukturnodar

Samanstilling mellom f-strukturar treng i `lfgalign` ikkje informasjon om c-strukturen, medan lenkjing av c-strukturnodar skjer på grunnlag av f-struktursamanstillinga. Programmet utfører difor samanstilling av c-strukturar sist[6].

Funksjonen `c-align` har som inndata c-strukturanalysane av kjelde- og målsetninga, og éi f-struktursamanstilling; utdata er ei mengd med lenkjer. Ei lenkje er

[6]Som nemnt i del 3.7.3 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.

eit par der første element er ei mengd c-strukturnodar på kjeldespråket, og andre element ei mengd nodar på målspråket. Det er ingen overlapp mellom medlem av lenkjer (ein node er aldri med i meir enn eitt par).

I Dyvik et al. (2009, s. 77) er kravet for å lenkje to c-strukturnodar er at dei dominerer same mengd med ordlenkjer[7]. Ein node n dominerer ei mengd lenkjer l viss unionen av lenkjene dominert av døtrene til n er lik l. I `lfgalign` opererer eg ikkje med *ordlenkjer* i seg sjølv; f-struktursamanstillinga er basert på LPT-korrespondansar, som definerer moglege ordlenkjer utan å sjå på kontekst, og f-struktursamanstillinga avgrenser vidare moglege ordlenkjer gitt f-strukturinformasjon. Preterminale nodar er dei mest ordnære nodane som kan ha ei f-strukturlenkje (ved ϕ); når formålet er å lenkje c-strukturnodar kan me nytte f-strukturlenkja til den preterminale noden i staden for ordlenkjer.To problem (kva vil me ha med?)
1. me får *ikkje* med LPT-korrespondansar som er OK, men ikkje med i $f-alignment$;
2. me *får* med LPT-korrespondansar som er med i $f-alignment$ men ikkje $aligntable$ (ikkje er rekursivt lenkja).

Programmet `lfgalign` følgjer krav (14) og lenkjer øvste nodar i funksjonelle domene, og subordinate nodar som har same informasjonstap. Prosedyren `c-align` i kodefigur 3 implementerer dette kravet.

TODO: gammal implementasjon, er ganske annleis no.

c-alignments $\leftarrow \emptyset$;
$splits_s \leftarrow$ new table ;
add-links(f-alignment, $tree_s, splits_s$) ;
$splits_t \leftarrow$ new table ;
add-links(f-alignment, $tree_t, splits_t$) ;
forall the *links being the keys in* $splits_s$ **do**
  if *(links in* $splits_t$*)* **then**
    add $(splits_s[links], splits_t[links])$ to c-alignments ;
end
return c-alignments ;

Funksjon 3: c-align(f-alignment, $tree_s$, $tree_t$)

Hjelpeprosedyren `add-links` (kodefigur 4) utfører hovudjobben. Inndata er rotnoden til c-strukturtreet for eitt av språka, og f-samanstillinga. Prosedyren kappar opp treet i nodemengder, kor kvar nodemengd dominerer same lenkjemengd (som definert over). Nodemengdene blir lagra i ein tabell, indeksert på lenkjemengdene. Prosedyren går rekursivt gjennom treet frå rot til lauv; lenkjemengden for kvar node er unionen av lenkjemengdene returnert av `add-links` kalt på kvar av døtrene. Viss ein node dominerer ei lenkjemengd *links*, legg me til denne noden i tabellen $splits[links]$. Merk at kvar c-strukturnode berre opptrer éin gong i tabellen.

[7] 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.

```
links ← ∅;
if node then
    if preterminal?(node) then
        let link ∈ f-alignment s.t. ϕ(node) ∈ link ;
        if link then links ← {link} ;
        if *pro-args-affect-c-links* then
            forall the a ∈ args(ϕ(node)) s.t. ϕ⁻¹(a) = ∅ do
                let link_a ∈ f-alignment s.t. a ∈ link_a ;
                if link_a then links ← {link_a} ;
            end
    else
        links ←add-links(f-alignment, left-branch(node)) ⋃
        add-links(f-alignment, right-branch(node)) ;
    add node to splits[links] ;
return links ;
```

Funksjon 4: add-links(f-alignment, $node, splits$)

Sidan `c-align` kallar `add-links` for kvar av sidene, får me to tabellar $splits_s$ og $splits_t$. Me hentar så ut alle dei lenkjemengdene som er i begge tabellane (dvs. snittet av oppslagsnøklene til tabellen); nodane som er lagra med same mengd med f-strukturlenkjer (same nøkkel i tabellen) skal lenkjast på c-strukturnivå. Alle desse mange-til-mange-lenkjene blir til slutt returnert av `c-align`. Om brukarvariabelen `*pro-args-affect-c-links*` er sann, vil me leggje til lenkja pro-argument i lenkjemengdene; denne variabelen styrer forskjellen mellom dei to alternative løysingane på ulenkja c-strukturnodar diskutert i del 3.7.1.

Prosessen er no ferdig, mange-til-mange-lenkjene mellom c-strukturnodar definerer frasesamanstillinga

til diskusjonsdel:*Det er ikkje berre ei N-gramsamanstilling; sidan lenkjene er mellom c-strukturnodar kor kvar node dominerer ein konstituent, kunne me kalt det ei* konstituentsamanstilling..

4.3.1 SKRIV viss me har LPT, men ikkje rekursiv f-lenkje

Dette bør kanskje vere valfritt i programmet, for å sjå kva det fører til: vil du ta med LPT-korrespondansar som ikkje har f-lenkjer i `add-links`?

Og omvendt, finst det LPT-korrespondansar som ikkje kjem med i f-alignment i det heile teke, men som likevel burde ha noko å seie for c-strukturlenkjinga? (Men burde dei ikkje då vere med i f-alignment au?)

Kan me **fjerne** visse f-samanstillingar pga. c-strukturinfo? dvs. disambiguere... (dette er vel heller stoff for diskusjonsdelen?)

Kapittel 5

Diskusjon, resultat av å automatisk samanstille norske og georgiske setningar

I denne delen gir eg ei evaluering av resultata frå å køyre `lfgalign` på LFG-analysar av norske og georgiske parallelle setningar. Eg ser på manglar ved implementasjonen i forhold til dei ideelle krava frå kapittel 3, og samanliknar dei resultata som er moglege å få frå `lfgalign` med dei som er mogleg å få med andre metodar, då spesielt reint N-grambaserte metodar. I tillegg diskuterer eg kort ulike bruksområde for desse samanstillingane.

5.1 Ressursar

Kjeldematerialet mitt er ei mengd med omtrent hundre med mindre det finst fleire? LFG-analyserte testsetningar på norsk og georgisk, frå eit testsett kor setningane er valt for å illustrere ei vid rekkje ulike syntaktiske situasjonar. Analysane er manuelt disambiguerte.

Dette materialet er sjølvsagt alt for lite for statistisk samanstilling, så når eg samanliknar med N-grambaserte metodar gir eg desse «perfekte» lenkjer (så gode som det vil vere mogleg å få).

5.2 N-grambaserte metodar

Och & Ney (2003, s. 20–21) definerer ei samanstilling på det mest generelle som ei delmengd av det kartesiske produktet av ordposisjonane i to setningar. Viss $f_1^J = f_1, ..., f_j, ..., f_J$ er orda i setninga på kjeldespråket og $e_1^I = e_1, ..., e_i, ..., e_I$ er orda i setninga på målspråket[1], er ei N-gramsamanstilling $\mathcal{A}$ gitt ved $\mathcal{A} \subseteq \{(j,i) : j =$

[1] e for engelsk og f for fransk, sjølvsagt.

$1,...,J; i = 1,...,I\}$. Eitt kjeldeord kan altså vere lenkja til eitt eller fleire målord, og omvendt.

I praksis vil spesifikke modellar prøve å avgrense dette, t.d. ved å krevje maksimalt eitt målord per kjeldeord, kor ein nyttar «det tomme ordet» for kjeldeord som ikkje er lenkja med noko målord. I litteraturen om statistiske metodar er dette ei *ordsamanstilling*; når mange-mange-lenkjer er med har me ei *frasesamanstilling*.

Ei samanstilling er her altså ei mengd par av ordposisjonar. Det finst mange ulike måtar å komme fram til samanstillinga på. Dei vanlege metodane er basert på statistiske modellar, kor den beste samanstillinga a er den som får høgast skåre på $p(f_1^J, a|e_1^I)$, dvs. sannsynet for samanstillinga a og kjeldesetninga gitt målsetninga og parametrane i modellen. Verdien til parametrane finn ein ved å trene modellen på eit parallellkorpus, kor ein god treningsmetode aukar den totale p for heile korpuset (dvs. produktet av p for alle setningane). For å finne p er det vanleg å nytte ein skjult Markov-modell, dvs. at p for heile setninga blir dekomponert til p for eitt enkelt kjeldeord, kor det siste er basert på p av orda som kjem før. I tillegg vil ein ofte vekte på sannsynet for ei viss setningslengd gitt målsetninga (sidan lange setningar ofte blir omsett til like lange setningar, modulo språkforskjellar). Det er mogleg å forenkle Markov-modellen til at p for eit enkeltord berre er avhengig av ordet som kjem rett før, noko som gjer utrekningane enklare (og lèt ein generalisere på bakgrunn av mindre data), men sjølvsagt gir ein mindre nyansert modell. Om me har forenkla modellen slik at p berre er avhengig av dei N orda som kjem før, har me ein N-grambasert modell.

5.3 Evaluering av `lfgalign`

5.4 Samanlikning med tremetodar og n-grammetodar

I tillegg til at ein kanskje kan få betre skåre på kvantitative mål som presisjon og gjenkjenning, vil lenkjer mellom f-strukturar gi informasjon som er kvalitativt forskjellig frå det ein kan få med å berre sjå på lenkjer mellom ord, n-gram eller konstituentar.

5.4.1 c->f er mange-til-ein

Avbildinga frå c-strukturnodar til f-struktur er mange-til-ein, kan me t.d. innanfor eitt tre ha fleire N-gram per f-strukturhovud; ein metode som berre ser på enkle N-gramlenkjer vil ikkje registrere desse relasjonane (t.d. metoden i Samuelsson & Volk (2007)).

5.5 Bruksområde

5.5.1 Oppdage argumentstrukturalternasjon

I døme TODO i kapittel TODO viste eg at f-strukturar og LPT-korrespondanse kanskje ikkje har nok informasjon til å kunne handtere ulik følgje i argumentstruktur. Programmet mitt vil her gi begge løysingar, ei rangering basert på lik argumentfølgje vil gi feil løysing på topp. ref

Kanskje kan me nytte data frå fleire førekomstar med andre subjekt og objekt til å lære slike argumentstrukturalternasjonar. Om me observerer *sie gefällt mir/jeg liker henne* vil me jo ha f-strukturinformasjon som kan nyttast til å informere argumentstrukturalternasjon (*sie/henne* er hokjønn, etc.), om det var substantiv der ville LPT-korrespondanse kunne informere dette.

Kapittel 6

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.

Och, F.J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19-51. Tilgjengeleg frå http://dblp.uni-trier.de/db/journals/coling/coling29.html#OchN03

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.)