

Syntaktisk informert frasesamanstilling

Kevin Brubeck Unhammer

13/08, 2010

Innhald

1	Innleiing	4
1.1	Vegkart	5
2	Bakgrunn og omgrepsavklaring	7
2.1	SKRIV Relaterte metodar	7
2.2	SKRIV Eit kort oversyn over leksikalsk-funksjonell grammatikk og terminologi	8
3	Krav til frasesamanstilling	10
3.1	Innleiing	10
3.2	Formål med frasesamanstilling	10
3.3	Frasesamanstilling i ein LFG-trebank	12
3.4	Kva kan lenkjast?	14
3.5	Krav på ordnivå	16
3.5.1	Ordklasse	17
3.6	SKRIV Krav på f-strukturnivå	17
3.7	Krav om lik argumentstruktur	17
3.7.1	TODO Sitere eigen korpusundersøking av variasjon i arg-str?	19
3.7.2	TODO Ulik følgje i argumentstruktur	19
3.8	SKRIV Kan adjunkt lenkjast til nodar <u>undermor</u> -lenkja?	20
3.8.1	1. Kausativar og inkorporering	21
	TOGROK adjunkt bør ikkje samanføyast? eller?	22
3.8.2	2. Adposisjonsobjekt	22
3.9	SKRIV Underordna c-strukturnodar	22
3.10	Funksjonelle c-strukturnodar	24
3.11	SKRIV Rangering	26
4	Implementasjonen av lfgalign	27
4.1	Lenkjer mellom f-strukturar	28
4.1.1	Overflødige adverbial	31
4.1.2	SKRIV Funksjonsord utan LPT-korrespondanse bør eigentleg hindre f-strukturlenkjing	31
4.1.3	SKRIV Når f-lenkjene ikkje er 1-1	31

	kausativ	31
	preposisjonsobjekt	31
4.1.4	Kan me gjere f-struktursamanstillinga bottom-up?	32
4.2	SKRIV Rangering	33
4.2.1	rekursivt lenkja > ulenkja, men LPT	33
4.2.2	argument-argument > argument-adjunkt	33
4.2.3	arg1-arg1 arg2-arg2 > arg1-arg2 arg2-arg1 (følge)	33
4.2.4	flest lenkja adjunkt	33
4.2.5	Prioritet på rangeringskriterium	33
4.3	Lenkjing av c-strukturnodar	33
4.3.1	SKRIV viss me har LPT, men ikkje rekursiv f-lenkje . . .	35
5	Avslutning	36

List of Corrections

TODO: abstract/samandrag	4
fortelje om georgisk i kap.5 heller	6
TODO (innleiing?):	
- reine N-gram-samanstillingar, dependensbaserte	
- ulike formål for samanstilling gir ulike metodar	
- kort oversyn over LFG-termar	
.	7
«by på fleire problem» – weasel wording, TODO omskriv.	7
fleire slike? meir om dette, algoritmen	7
Dette er motsett retning av det mitt program gjer, nemne seinare?	8
todo: referere til den faktiske parsaren? det var Bikel kanskje?	11
Diskutabelt, TODO:	14
TODO: teikne inn f-domene	15
der ADJUNKT ikkje er realisert, lenkjer me ikkje PRED. skal me då	
ikkje lenkje ord heller?	16
finst det tilfelle der ordlenkjer ikkje impliserer PRED-lenkjer?	
(hypotese: det er alltid slik at ordlenkjing av predikerande ord =>	
PRED-lenkje)	
PRED->ord :: iallfall	
PRED<-ord :: ?	
PRED<->ord	
PRED, ord	16
TODO: når?	16
TODO: litt brå avslutning	16
TODO: Er det mogleg å presisere LPT-kravet meir? Skal det berre vere	
eit rangeringskrav??	16
utgangspunkt i det som står i kap.4	17
LCS, dorr	19
intro TODO, kanskje noko om kva eg faktisk har fått ut av implementa-	
sjonen	27
treng eg ein eigen del om LPT i dette kapittelet? Implementasjonen er jo	
veldig enkel iallfall.	27
TOGROK: kan me i tillegg sjekke om funksjonelle nodar i same f-domene	
som PRED-ordformen har LPT-korrespondanse?	28

TODO: nemne føresetnaden om uavhengnad i kapittel 3	29
TODO: forskjellen mellom LPT-krav og rekursjonskrav på argument må inn i kapittel 3	29
TODO: implementere :->	31
og så er det spørsmålet om me kan lenkje adjunkt på ulike nivå i f- strukturane	31
Dette må 1. spesifiserast (kap.3), og 2. implementerast...	31
Rangering er ikkje implementert enno (akkurat no gir rank(f-alignments) berre ut første samanstilling.)	33
To problem (kva vil me ha med?)	
1. me får <i>ikkje</i> med LPT-korrespondansar som er OK, men ikkje med i <i>f – alignment</i> ;	
2. me får med LPT-korrespondansar som er med i <i>f – alignment</i> men ikkje <i>aligntable</i> (ikkje er rekursivt lenkja).	34
til diskusjonsdel: <i>Det er ikkje berre ei N-gramsamanstilling; sidan len- kjene er mellom c-strukturknodar kor kvar node dominerer ein kon- stituent, kunne me kalt det ei konstituentsamanstilling.</i>	34
Kan me fjerne visse f-samanstillingar pga. c-strukturinfo? dvs. disambi- guere... (dette er vel heller stoff for diskusjonsdelen?)	35

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 omsetje 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 språkstudie. 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 relasjonar mellom grammatiske funksjonar, tematiske roller og kasusmarkering, ved hjelp av parallelle trebankar annotert med djupe grammatiske analysar. Samanstillinga planlagt der krev i større grad syntaktisk likskap for å kalle to uttrykk omsetjingar, men kan difor au tene på det at dei grammatiske ana-

lysane er utvikla med tanke på å vere så parallelle som mogleg. 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-strukturen¹.

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 krev 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-strukturknodar 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-strukturknodar (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 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 følge.

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

Eg evaluerer samanstillingane som kjem ut av denne metoden i kapittel ??, 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.

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 avgrensar då kva som er ein syntaktisk konstituent.

fortelje om
georgisk i
kap.5 heller

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 kolokasjonsinformasjon for å først finne sannsynlege nominale frasar på engelsk og kinesisk (dvs. «chunking»), og så samanstillе 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.

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

TODO
(innleiing?):
- reine
N-gram-
samanstillingar. ■
dependensba-
serte
- ulike formål
for
samanstilling
gir ulike
metodar
- kort oversyn
over
LFG-termar

fleire slike?
meir om dette,
algoritmen

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 parsar 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 forklarar dette med den store mengda uttrykk som ikkje utgjer syntaktiske konstituentar i følge 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 dependensinformasjon (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 LFG-terminologi (i tillegg til eit par eigne termar). Difor gir eg her eit kort oversyn over det som kan vere nytt for dei som er meir vand med andre grammatiske rammeverk.

modellteoretisk (vs derivasjonelt) LFG er eit modellteoretisk, ikkje-derivasjonelt, rammeverk for grammatikk. Pullum & Scholz (2001) gir ein god gjennomgang av forskjellen mellom derivasjonelle (enumerative) grammatikkar og modellteoretiske grammatikkar, kor førstnemnde definerer *mengder av ut-*

trykk ved avleiing frå startsymbol, medan sistnemnde gir skildringar av *enkeltuttrykk*. 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 skildringa 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; ein 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.

f-struktur ...

c-struktur ...

endosentrisitetsprinsippa ...

\bar{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 tilsvarende 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 termene *lenkjing* og *samanstilling* i omtrent same tyding som dei engelske termene *link* og *alignment*, kor ei samanstilling er ei mengd lenkjer. Dette er ekvivalensforhold som me kan finne mellom lingvistiske *representasjonar* (f-struktur, c-struktur) eller *uttrykk* (ord, setningar); lenkjing mellom dei siste altså er meir ateoretisk / datanært.

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

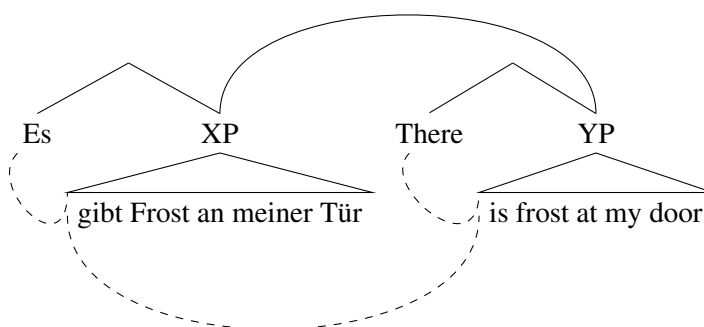
Ei frasesamanstilling er ein slag annotasjon av eit korpus. På same måte som oppbygginga av eit korpus avheng av formålet til korpuset, kan ein ikkje definere den ideelle annotasjonen av eit korpus utan å ta høgd for kva ein skal nytte annotasjonen til.

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 omsetjingssteget¹ 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 semantiske² omsetjingsskifte, t.d.

¹Både 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).

²Sidan 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), osv. (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 konstituentar³ 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 ordsamanstilling⁴, hovudsakleg for dei opne klassene, og er definert ved å vise til kva for rolle eit argumentord spelar, eller kva for rolletildeling eit predikat eller modifierande ord gir. Så for å t.d. samanstill to verb må dei ha like mange semantiske argument (men argumenta treng ikkje alle realiserast syntaktisk) og dei

³LFG 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 *tildeler same roller*; medan argumenta må *spele same rolle*, og både argument og adjunkt må vere *koreferente*. Lenkja ord må vere del av frasar som spelar 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 spelar 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) avleidd 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 LFG⁵ (f-struktur, c-struktur, endosentrisitetsprinsipp, \bar{X} -tre, osv.) 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 sjølvstøtt 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.

⁵I 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-strukturknodar.

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.

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* 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 setningane⁶. 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. **Diskutabelt, TODO:** PRED-elementa er i tillegg gjerne enklare å knyte direkte opp mot den konkrete, observerte tekststrengen, medan t.d. aspekt kanskje er umogleg å skilje frå tempus i affikset.

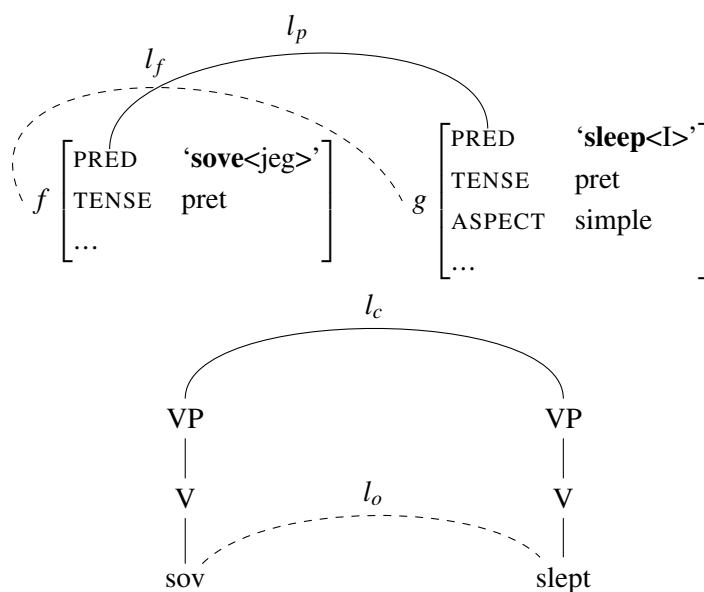
Samtidig er det au eit omsetjingsforhold mellom trekka i same f-struktur som dei lenkja PRED-elementa, og me ville kanskje ikkje ha omsett dei to PRED-elementa i andre f-strukturkontekstar. Difor bør me au sjå på ei PRED-lenkje som ei lenkje mellom *f-strukturane til desse PRED-elementa*⁷. Med dette 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,

⁶I del 3.10 kjem eg tilbake til spørsmålet om me vil inkludere visse f-strukturar utan PRED-element i kandidatane for samanstilling.

⁷Eventuelt kunne me ha definert lenkjingskandidatane på f-strukturnivå som alle PRED-haldande f-strukturar, resultatet blir det same.

- b. orda i setningane som projiserer PRED-elementa tek del i ei lenkje (kor andre ord kan vere involvert), og at
- c. iallfall dei øvste nodane i $\phi^{-1}(f)$ og $\phi^{-1}(g)$, dei funksjonelle domena til f-strukturane f og g , er lenkja



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, vil VP-ane som heilskap vere lenkja (både VP-nodane som dominerer dei lenkja funksjonelle domena, og f-strukturane frå ytre PRED til verba), det er dette at heile VP-ane 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 øvste node dominerer alle nodar i det funksjonelle domenet; det er det funksjonelle domenet som spesifiserer informasjonen i f-strukturane, toppnodane dominerer heile dette domenet og bør difor lenkjast viss og berre viss f-strukturane er lenkja.

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 3.9 ser eg på kva som må til for å lenkje ikkje-øvste nodar i det funksjonelle domenet. I tillegg finst det nodar over ord som ikkje projiserer PRED-element, desse kjem eg tilbake til i del 3.10.

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.

3.5 Krav på ordnivå

Ord som skal lenkjast må i Thunes (2003) vere del av frasar som spelar 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*; forholde 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 omsetjingar⁸. Dyvik et al. (2009, s. 74) krev dei at orda generelt, utan kontekst, må vere semantisk plausible omsetjingar, eller at målordet er eit medlem av mengda av *linguistically predictable translations* av kjeldeordet. Målordet har då *LPT-korrespondanse* med kjeldeordet. Informasjon om slik LPT-korrespondanse kan komme frå ei djup semantisk dekomponering av kvart ord, då blir det eit krav på f-strukturnivå (eller på ein semantisk struktur), eller han kan komme bottom-up, typisk frå automatisk ordsamanstilling. Bottom-up-informasjonen viser då om orda generelt (i ulike kontekstar) blir nytta som omsetjingar av kvarandre. Nedanfor reknar eg LPT-kravet som eit krav på ordnivå.

⁸Eigentleg burde slike setningar ikkje vere lenkja på setningsnivå ein gong, men som me skal sjå i del 3.7 treng me kravet om lik tyding sjølv innanfor setninga.

der
 ADJUNKT
 ikkje er
 realisert,
 lenkjer me
 ikkje PRED.
 skal me då
 ikkje lenkje
 ord heller?
 finst det
 tilfelle der
 ordlenkjer
 ikkje
 impliserer
 PRED-
 lenkjer?
 (hypotese: det
 er alltid slik at
 ordlenkjing av
 predikerande
 ord =>
 PRED-lenkje)
 PRED->ord ::
 iallfall
 PRED<-ord ::
 ?
 PRED<->ord
 PRED, ord
 TODO: litt
 brå avslutning

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 {han, bjeffa} og {iqePa}. Generelt må me ha slike lenkjer der eitt ord projiserer fleire PRED-element⁹.

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 omsetjing¹⁰.

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.

3.6 SKRIV Krav på f-strukturnivå

3.7 Krav om lik argumentstruktur

utgangspunkt
i det som står i
kap.4

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

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 sa-

⁹Me 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.

¹⁰Munday (Catford (1965), i 2001, s. 61) gir ein gjennomgang av slike *klasseskifte*, og andre typar omsetjingsskifte.

For å gjere dette konkret kan me sjå på setning 7 i MRS-suiten (?)¹¹:

- ¹²Analysane 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).

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

$$\left[\begin{array}{ll} \text{PRED} & \text{'wetten<Abrams, sigarett>'} \\ \text{ADJUNCT} & \{ \text{Browne, sigarett} \} \end{array} \right]$$

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

$$\left[\begin{array}{ll} \text{PRED} & \text{'bet<Abrams, sigarett, regne>'} \\ \text{ADJUNCT} & \{ \text{Browne} \} \end{array} \right]$$

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 treng bottom-up-informasjon om kva for adjunkt og argument som samsvarer.

Det same gjeld forøvrig lenkjing av adjunkt til adjunkt. Adjunkt plukker ut si eiga rolle der argument får rolla tildelt frå verbet, og f-strukturane har ingen hierarkisk inndeling av desse slik me har for verb og argument, dei er i staden representert som *uordna mengder*.

3.7.1 TODO Sitere eigen korpusundersøking av variasjon i arg-str?

Ei undersøking av den frasesamanstilte trebanken SMULTRON (Samuelsson & Volk, 2006) mot LFG-grammatikkane for engelsk og tysk fann at 2 av 15 korresponderande verbtoken¹³ for høgfrequente innhaldsverb fekk analysar kor argument korresponderte med adjunkt (?).

LCS, dorr

3.7.2 TODO 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 opplevart som objekt og tema som subjekt vil ha opplevart nedanfor tema i argumentstrukturen, medan ei omsetjing av dette verbet kan ha tema nedanfor:

- (6) a. sie_j gefallen ihnen_i

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

$$\left[\text{PRED} \quad \text{'gefallen'} \langle \text{de}_j, \text{de}_i \rangle \right]$$

\leftrightarrow

b. de_i liker dem_j

$$\left[\text{PRED} \quad \text{'like'} \langle \text{de}_i, \text{de}_j \rangle \right]$$

Argumentstrukturane i (6) har omvendt intern følgje, og som vist ved dette dømet er det heller ikkje noko f-strukturinformasjon me kunne nytta til å sikre lenkjinga *sie/dem* og *ihnen/de*. Igjen ser det ut til at bottom-up-informasjon trengst.

3.8 SKRIV Kan adjunkt lenkjast til nodar under mor-lenkja?

Krav (vi) i Dyvik et al. (2009, s. 75) krev at viss F_s og F_t er lenkja, så kan ingen adjunkt D_s til F_s vere lenkja til nodar utanfor F_t . Men kan ein D_s lenkjast til ei dotternode av argument eller adjunkt til F_t ?

R_t er dotter til F_t , og må då vere lenkja til ei dotter av F_s , A_s . Då må au alle argument til R_t vere lenkja til døtre av A_s , så D_s kan ikkje lenkjast til argument av dotternodar til F_t . Kva med adjunkt? Om me finn eit ulenkja adjunkt til R_t kan me heller ikkje lenkje dette til D_s ved krav (vi) igjen, sidan D_s står utanfor A_s .

Men om D_t er ei ulenkja *adjunktdotter* av F_t , så vil døtre av D_t kunne lenkjast til D_s , så lenge D_t forblir ulenkja. Me kan altså sjå ned i adjunktdøtre av F_t for å lenkje D_s .

På same måte bør ein kunne rekursivt sjå ned i ulenkja adjunktdøtre av R_t , men ein bør kanskje ikkje kunne lenkje så djupt uansett? Ikkje automatisk, uansett.

Programmet mitt vil, gitt to initielle f-strukturar med LPT-korrespondanse, finne alle moglege kombinasjonar av lenkjer som inneheld alle argument og kanskje adjunkt, dvs. om me har

$F_s \quad [\quad \text{PRED} \quad p \langle 1, 2 \rangle \quad \text{ADJUNCT} \quad \{ \quad 3 \quad \} \quad]$

$F_t \quad [\quad \text{PRED} \quad p \langle 4 \rangle \quad \text{ADJUNCT} \quad \{ \quad 5, 6 \quad \} \quad]$

vil dette vere logisk moglege samanstillingar av «f-strukturdøtre»:

$$\begin{aligned} &(((1 \ . \ 4) \ (2 \ . \ 5)) \ ((1 \ . \ 4) \ (2 \ . \ 6)) \ ((1 \ . \ 5) \ (2 \ . \ 4)) \\ &((1 \ . \ 5) \ (2 \ . \ 6) \ (3 \ . \ 4)) \ ((1 \ . \ 6) \ (2 \ . \ 4)) \ ((1 \ . \ 6) \ (2 \ . \ 5) \ (3 \ . \ 4))) \blacksquare \end{aligned}$$

Me luker ut kombinasjonar som bryt med LPT-korrespondanse. Med full informasjon bør me sjølvstødig berre ende opp med éin kombinasjon, t.d. $((1 \ . \ 4) \ (2 \ . \ 5))$.

Så langt bør altså krav (i-iv) frå Dyvik et al. (2009) vere dekkja.

Me kan krevje at f strukturane-til f strukturdøtre-kan lenkjast rekursivt for at F_s og F_t skal lenkjast, t.d. både (1 . 4) og (2 . 5). Men her kjem det (iallfall) to problem.

3.8.1 1. Kausativar og inkorporering

Om me har

```
F_s [ PRED p<SUBJ, 1, 2> XCOMP 2[ PRED q<1> ] ]
```

```
F_t [ PRED pq<SUBJ,OBJ> ]
```

kor pq er t.d. ein kausativ som tilsvarear $p<\dots, q>$, så vil me ikkje kunne lenkje F_s og F_t sidan det bryt med krav (iii), F_s har eit argument for mykje. Men her vil det kanskje vere naturleg å ha ei ein-mange-lenkje:

```
((F_s 2) . F_t)
```

No kan me sjå på unionen av argument av F_s (minus XCOMP) og argument av XCOMP, alle argument i denne unionen må då ha LPT-korrespondanse med argument/adjunkt av F_t , og alle argument av F_t må ha LPT-korrespondanse med argument/adjunkt av unionen.

Det same bør kanskje skje ved vanleg inkorporering av substantiv, då må det altså vere mogleg å føye saman t.d. verb og objekt; ein kombinasjon av dette og kausativ bør vel vere mogleg, t.d.

```
F_s [ PRED la<SUBJ, 1> XCOMP 2[ få<1, 3:pengar> ] ]
```

```
F_t [ PRED belønn<SUBJ, 1> ]
```

Igjen ser me på argument frå unionen av $(F_s 2 3)$ minus 2 og 3, og om det er mogleg å lenkje dei til argument/adjunkt av F_t , og omvendt.

Men det bør kanskje vere grenser for kor langt samanføyning kan gå... eg kan ikkje tenkje meg at me vil lenkje $((F_s 2) . F_t)$ eller $((F_s 1 2) . F_t)$ her:

```
F_s [ PRED p<..., 1> XCOMP 1[ PRED q<..., 2> XCOMP 2[ PRED r<...> ] ] ] ■
```

```
F_t [ PRED pr<...> ]
```

... men det kan jo hende det finst situasjonar der dette au vil vere rett. Problemet er altså kor me skal setje grensene i implementasjonen. Om me skal prøve å samanføye på alle moglege måtar (altså, der me ikkje har informasjon om LPT), i tillegg til «vanlege» lenkjer, blir det fort komputasjonelt vanskeleg. Me kan sjølvstundt snu på LPT-kravet her, og seie at dette er berre lov der me har positiv informasjon om LPT-korrespondanse, i staden for at det ikkje er lov om me har motstridande LPT-informasjon, det vil nok hjelpe, men det er vanskeleg å finne prinsippelle avgrensingar her.

TOGROK adjunkt bør ikkje samanføyast? eller?

Det einaste eg kan tenkje meg er at adjunkt ikkje bør vere kandidatar for samanføyning (i såfall burde dei vel heller vore analysert som argument?).

3.8.2 2. 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 $L'(F_t) = \{\text{sigaretze}, \text{ze}\}$ (evt. $\{\text{sigaret}, \text{ze}\}$), då kan me lenkje (F_s . F_t), medan 1 er ulenkja.

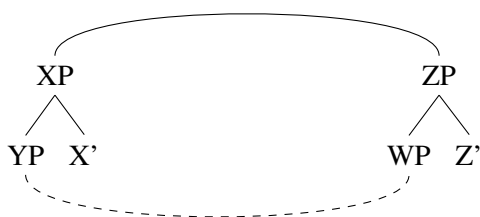
Eller me kan lenkje (F_s . 1), kor me har skikkeleg LMT-korrespondanse, men då må me slakke på (iii) og (iv), og altså ha lov til å «hoppe over» ein f-struktur for å lenkje «veddet» og «daenajleva». F_t er då ulenkja. Det er løysinga valt i Dyvik et al. (2009, s. 75, fotnote 3), og den løysinga eg følgjer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

3.9 SKRIV Underordna c-strukturnodar

Toppnodane i eit lenkja funksjonelt domene i c-struktur (XP på språk 1, ZP på språk 2) vil ha ein informasjonsmessig korrespondanse, og kan samanstillast. Men det er mogleg å samanstille to toppnodar i funksjonelle domene i c-strukturen utan at nodane under (X' , Z') er samanstilt. Ein grunn til å ikkje samanstille desse underordna nodane, vil vere viss spesifikator til X ikkje spelar same rolle i tolkinga som spesifikator til Z, dvs. viss YP og WP i figur 3.3 ikkje er lenkja.

Me kan utelukke lenkjing av ikkje-konstituentar som *there is* ved å krevje at ei fullstendig samanstilling mellom to frasar må vere slik at heile substrukturen au er samanstilt. *There is* og *Es gibt* i figur 3.1 kan då ikkje samanstillast åleine, men berre som del av ei ytre frasesamanstilling. Så når *kan* me samanstille nodane som står under øvste node i det funksjonelle domenet?

I figur 3.3 der XP og ZP er lenkja, vil YP og WP – i kraft av å vere toppnodar i sine domene – måtte ha ei lenkje i f-strukturen for at c-strukturnodane kan lenkjast



Figur 3.3: Lenkjing av underordna c-strukturknodar

(det kunne jo t.d. hende at f-strukturen projisert av YP samsvarte med den projisert av Z', eller ein struktur under Z').

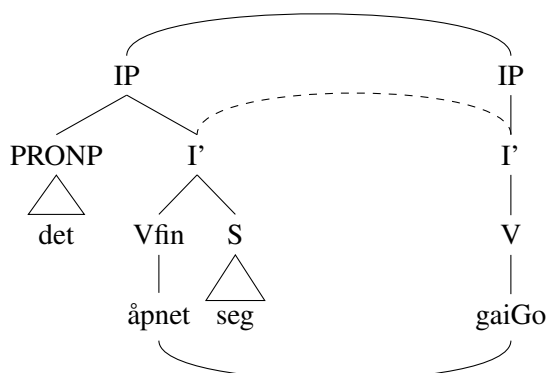
Om me skal lenkje Z' og X' i figuren over må dei respektive spesifikatornodane vere lenkja. Me får då følgjande krav:

- (7) Krav for lenkjing av underordna c-strukturknodar:
- c-strukturknodar som ligg under øvste node i to funksjonelle domena kan berre samanstillast med nodar som ligg innanfor desse domena,
 - c-strukturknodar kan berre samanstillast om deira funksjonelle domene er lenkja på f-strukturnivå,
 - om ein c-strukturnode X' som ikkje er toppnode i det funksjonelle domenet har ein søsternode YP, må YP vere samanstilt med ein søsternode til Z' for å samanstillast X' og Z'

(7-a) seier at om XP og ZP er samanstilt, der XP er t.d. OBJ til IP, kan ikkje Z' samanstillast med SUBJ til IP osb., men berre til nodar innanfor OBJ-domenet. (7-c) påført figur 3.3 seier altså at spesifikatornodane må vere lenkja for at X' og Z' skal lenkjast (manglande søsternode på den eine sida vil au hindre samanstilling).

I figur 3.5 er alle nodane under S vist i dei to trea i same funksjonelle domene (kvar node under S er annotert med $\uparrow=\downarrow$), så om dei funksjonelle domena er samanstilt (som krev at *rom cvimda* og *at det regner* er samanstilt), vil (7-a og -b) vere oppfylt kva gjeld CP-komplementa – lenkjinga går ikkje ut over dei funksjonelle domena. Sidan Csub og Cnom er funksjonelle kategoriar er dei au samanstilt via samanstillinga av S-nodane og føringane i (8), og (7-c) er då oppfylt. (7) står altså ikkje i vegen for å samanstillast IP-en over *cvimda* og Ssub.

I figur 3.4 derimot (?), kan me ikkje samanstillast I'-nodane. PRONP-noden, spesifikator på den norske sida, er ikkje lenkja med nokon spesifikator på den georgiske sida. Den informasjonen (her reint syntaktisk) som ordet *det* tilfører IP, ligg under I' på georgisk. Om me skulle lenkja I', måtte me altså hatt ein georgisk spesifikator som var lenkja til den norske PRONP.



Figur 3.4: Umogleg samanstilling av funksjonsord mellom bokmål og georgisk

3.10 Funksjonelle c-strukturnodar

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

Problemet med å nytte metoden frå del 3.9 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-strukturteet i buntar med likt informasjonstap.

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

Viss begge språk har funksjonsord, men funksjonsord som ikkje kan sjåast på som moglege omsetjingar (t.d. *fordi* og *whether*), bør me ikkje ein gong lenkje komplementa¹⁴. Samtidig vil me ikkje at eit manglande funksjonsord på det eine språket skal hindre lenkjing av komplementa, sidan det kan hende at funksjonsordet ikkje er krevd på det språket (eventuelt kjem dette fram som korrespondansar i f-strukturtrekk, eg har ikkje teke høgd for korrespondansar mellom element som ikkje er PRED i denne oppgåva).

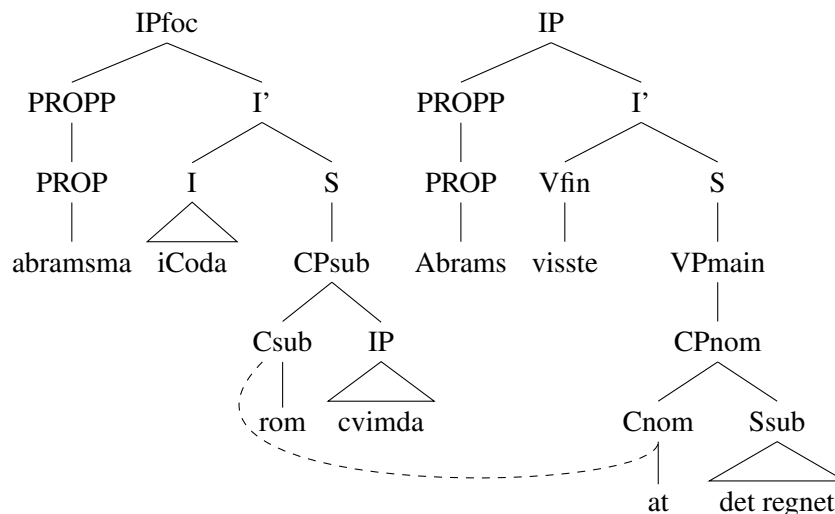
Me krev at komplementa er lenkja for å sikre at me ikkje lenkjer nodar som står i ulike kontekstar (me vil ikkje lenkje *at* i «han såg at det gjekk bra» med *that* i «he saw that she drew a picture»), jamfør kravet om lenkja argument for lenkja predikat i del 3.7.

¹⁴Skal ein lenkje ordet *som* (utan PRED) med ordet *which* (med PRED)? Viss krava elles er oppfylt, kan det kanskje vere informativt med ein type «defekt» lenkje, sjølv om berre det eine ordet blir rekna for å vere eit innhaldsord. Frasane til deira funksjonelle domene vil uansett vere samanstilt via toppnodane (t.d. CP).

Då får me følgjande:

- (8) Krav for lenkjing av funksjonelle kategoriar i c-strukturen:
- Gitt ei mogleg lenkjing av FP og GP, kor F og G er funksjonelle kategoriar der komplementa elles kan lenkjast, introduserer me eit «falskt informasjonstap» mellom FP og F' og mellom GP og G'; orda under F' og G' må ha LPT-korrespondanse for at FP og GP skal kunne lenkjast, då kan me au lenkje F' og G'.
 - Gitt ei mogleg lenkjing av FP og XP, der F er ein funksjonell kategori, medan X er ein ikkje-funksjonell kategori, ignorerer me den funksjonelle kategorien i c-strukturlenkjinga. Sidan det ikkje er noko forskjell i informasjonstap mellom FP og F', er F' medlem av nodemengden som blir lenkja til XP.

Om (8-a) er oppfylt, kan me få samanstillinga vist i figur 3.5. Her vil dei funksjonelle domena til CPsub og CPnom kvar kunne delast opp i to deler, kor den funksjonelle delen har LPT-korrespondanse medan komplementa er lenkja på f-strukturnivå. Det er ingen informasjonstap frå CPsub til Csub som ikkje er reflektert i CPnom til Cnom, og det er ingen informasjonstap frå CPsub til IP som ikkje er reflektert i CPnom til Ssub.



Figur 3.5: Mogleg samanstilling av funksjonsord 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.

3.11 SKRIV Rangering

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 Lisp¹-program.

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 XLE³. 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.

Programmet byrjar lenkjinga med f-strukturane. Ei f-struktursamanstilling 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.

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

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

³Formatet er dokumentert på <http://www2.parc.com/isl/groups/nlt/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-variablar refererer til same analyseval); men filene eg har testa utnyttar ikkje det fulle spennet til formatet, så det finst ganske sikkert feil.

intro TODO,
kanskje noko
om kva eg
faktisk har fått
ut av imple-
mentasjonen

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

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-strukturar⁴. 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-korrespondanse⁵, 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.

Hjelpfunksjonen *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*.

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

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 $\{(SUBJ, SUBJ), (OBJ, ADJ)\}$ og $\{(SUBJ, ADJ), (OBJ, SUBJ)\}$. 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 (SUBJ, 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 *alignable*. 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

⁴Eigentleg 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.

⁵Når eg her skriv at to f-strukturar har LPT-korrespondanse, meiner eg sjølvsagt at ordformen 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 .

```

alignments  $\leftarrow \emptyset$  ;
forall the argperm in argalign( $F_s$ ,  $F_t$ ) do
     $p \leftarrow \emptyset$  ;
    forall the  $A_s$ ,  $A_t$  in argperm do
        if not(aligntable[ $A_s$ ,  $A_t$ ]) then
            | aligntable[ $A_s$ ,  $A_t$ ]  $\leftarrow$  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
         $a \leftarrow$  copy-of( $p$ ) ; // optional adjunct links
        forall the  $A_s$ ,  $A_t$  in adjperm do
            if not(aligntable[ $A_s$ ,  $A_t$ ]) then
                | aligntable[ $A_s$ ,  $A_t$ ]  $\leftarrow$  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 =  $\emptyset$  then return  $\emptyset$  ; // Fail
else return (( $F_s$ ,  $F_t$ ), alignments) ;

```

Funksjon 1: f-align(F_s , F_t)

TODO:
nemne
føresetnaden
om
uavhengnad i
kapittel 3

Sjølvs 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 $p \in argalign-p(args_s - \{s\}, adj_s, args_t - \{t\}, adj_t)$ **do**
 add $\{(s,t)\} \cup p$ to a ;

end

forall the $t \in adj_t$ **where** $LPT(s,t)$ **do**

forall the $p \in argalign-p(args_s - \{s\}, adj_s, args_t, adj_t - \{t\})$ **do**
 add $\{(s,t)\} \cup p$ to a ;

end

return a ;

else if $args_t$ **then**

if adj_s **then**

$s \in adj_s$;

forall the $t \in args_t$ **where** $LPT(s,t)$ **do**

forall the $p \in argalign-p(args_s, adj_s - \{s\}, args_t - \{t\}, adj_t)$ **do**
 do add $\{(s,t)\} \cup p$ to a ;

end

return a ;

else

return \emptyset ; // Fail

else

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

Funksjon 2: $argalign-p(args_s, adj_s, args_t, adj_t)$

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

- (1) a.
$$F_s \left[\begin{array}{l} \text{PRED 'planlegge<eg,[1:gi]>'} \\ \text{XCOMP}_1 \left[\text{PRED 'gi (opp)'} \right] \end{array} \right]$$
- b.
$$F_t \left[\begin{array}{l} \text{PRED 'plan<I,[2:give]>'} \\ \text{XCOMP}_2 \left[\text{PRED 'give<I,him,it>'} \right] \end{array} \right]$$

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. Hjelpenfunksjonen *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

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

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

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

```

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

```

Men dette er fordi sigaretze er adjunkt, og i følge 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 3.8.

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 vek-
 tast? (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.

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 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-strukturknoder at dei dominerer same mengd med ordlenkjer⁶. 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 avgrensar vidare moglege ordlenkjer gitt f-strukturinformasjon.■

Preterminale nodar er dei mest ordnære nodane som kan ha ei f-strukturlenkje (ved \emptyset); når formålet er å lenkje c-strukturknoder kan me nytte f-strukturlenkja til den preterminale noden i staden for ordlenkjer.

```

c-alignments  $\leftarrow \emptyset$  ;
splitss  $\leftarrow$  new table ;
add-links(f-alignment, trees, splitss) ;
splitst  $\leftarrow$  new table ;
add-links(f-alignment, treet, splitst) ;
forall the links being the keys in splitss do
    if (links in splitst) then
        | add (splitss[links], splitst[links]) to c-alignments ;
    end
return c-alignments ;

```

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

Hjelpeprosedyren add-links utfører hovudjobben. Inndata er rotnoden til c-strukturreet 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].

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 oppslagsnøkklene til tabellen); nodane som er lagra med mengd med f-strukturlenkjer skal lenkjast på c-strukturnivå. Alle desse mange-til-mange-lenkjene blir til slutt returnert av c-align.

Proessen er no ferdig, mange-til-mange-lenkjene mellom c-strukturknoder definerer frasesamanstillinga

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

⁶Dette 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.

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).

```

links ← ∅;
if node then
  if preterminal?(node) then
    let link ∈ f-alignment s.t.  $\phi(\text{node}) \in \text{link}$  ;
    if link then links ← {link}
  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)

```

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. disambi-
guere... (dette
er vel heller
stoff for
diskusjonsde-
len?)

Kapittel 5

Avslutning

Referansar

- Alsina, A., Bresnan, J. & Sells, P. (red.). (1997). *Complex predicates*. Stanford, CA, USA: Center for the Study of Language and Information. Paperback.
- 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/~pkoeht/publications/phrase2003.pdf>
- Meurer, P. (2008, March). *A Computational Grammar for Georgian*. Tilgjengeleg frå <http://maximos.aksis.uib.no/~paul/articles/Tbilisi2007-LNAI.pdf>
- Munday, J. (2001). *Introducing Translation Studies: Theories and Applications*. London: Routledge.
- Piao, S. & McEnery, T. (2001). Multi-word Unit Alignment in English-Chinese Parallel Corpora. I P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja (red.), *Proceedings of the Corpus Linguistics 2001 Conference* (s. 466–475). Lancaster, UK. Tilgjengeleg frå http://personalpages.manchester.ac.uk/staff/scott.piao/research/papers/mwu_align4.pdf
- Pullum, G. & Scholz, B. (2001). On the Distinction between Model-Theoretic and Generative-Enumerative Syntactic Frameworks. *Logical Aspects of Computational Linguistics: 4th International Conference, Lacl 2001, Le Croisic, France, June 27-29, 2001, Proceedings*. Tilgjengeleg frå <http://portal.acm.org/citation.cfm?id=645668.665062>
- Riezler, S. & Maxwell, J. (2006). Grammatical Machine Translation. I M. Butt, M. Dalrymple & T.H. King (red.), *Intelligent Linguistic Architecture: Variations on themes by Ronald M. Kaplan* (s. 35–52). Stanford, CA: CSLI Publications. Tilgjengeleg frå <http://www.parc.com/research/publications/details.php?id=5675>
- Rosén, V., Meurer, P. & Smedt, K. de. (2009). LFG Parsebanker: A Toolkit for Building and Searching a Treebank as a Parsed Corpus. I F.V. Eynde, A. Frank, G. van Noord & K.D. Smedt (red.), *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7)* (s. 127–133). Utrecht: LOT. Tilgjengeleg frå <http://ling.uib.no/~desmedt/papers/tlt7rosen-submitted.pdf>
- Samuelsson, Y. & Volk, M. (2006). Phrase Alignment in Parallel Treebanks. I *Proceedings of Treebanks and Linguistic Theories (TLT '06)*. Prague. Tilgjengeleg frå http://ling16.ling.su.se:8080/new_PubDB/doc_repository/229_align.pdf

- Samuelsson, Y. & Volk, M. (2007). Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment. I *Proceedings of Treebanks and Linguistic Theories (TLT '07)*. Bergen, Norway.
- Thunes, M. (2003). *Ekserpering av leksikalske oversettelsekorrespondanser fra parallelltekst*. Tilgjengeleg frå <http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/marthaex.pdf>
- Tinsley, J., Hearne, M. & Way, A. (2007). Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. I *Proceedings of Treebanks and Linguistic Theories (TLT '07)*. Bergen, Norway.
- XPar. (2008). *XPAR: Language diversity and parallel grammars*. (Submitted to the Research Council of Norway.)