Syntaktisk informert frasesamanstilling

Kevin Brubeck Unhammer

15/04, 2010

Innhald

List of Corrections

Kapittel 1

Introduksjon (+ samandrag/abstract)

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

Det at me kan omsetje mellom slike skildringar (t.d. trekkstrukturane til 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 ulike kriterium kan leggast til grunn for å kalle to uttrykk omsetjingar av kvarandre. Samuelsson & Volk (2006) nyttar t.d. reint semantiske kriterium, utan krav om syntaktisk likskap, i deira manuelle samanstilling; medan samanstillinga planlagt i XPar-prosjektet (XPar, 2008), som kjem via f-struktur-parallellismen[1], i større grad krev syntaktisk likskap.

Automatiske metodar for tekstsamanstilling kan nyttast til ulike metodar for maskinell omsetjing, i tillegg til oppbygging av parallelle korpora for meir teoretiske språkstudie. I samanheng med XPar-prosjektet (XPar, 2008) har eg sett på metodar for automatisk frasesamanstilling, dvs. for å finne omsetjingsforhold mellom fleire ord. Dei første metodane for dette kom frå statistisk maskinomsetjing, der ein berre nytta sannsyn av N-gram-omsetjingar, utan nokon form for syntaktisk informasjon. Samuelsson & Volk (2007) skildrar ein metode kor frasesamanstilling blir oppnådd vha. ordsamanstilling på ein parallell trebank der berre N-gram som svarer til ein syntaktisk node blir samanstilt som frasar. Tinsley et al. (2007); Hearne et al. (2008) viser at slike syntaktisk motiverte metodar kan forbetre frasebasert stokastisk maskinomsetjing (*PBSMT*). I XPar vil ein finne ut om frasesamanstilling kan forbetrast ved å utnytte det at LFG-grammatikkane for dei ulike språka er skrivne med same prinsipp lagt til grunn; to parallellstilte setningar bør ha f-strukturar

[1] Eg går her ut frå at lesaren er kjend med grunnleggjande LFG-terminologi.

som er like nok til at me kan samanstille frasar ved hjelp av likskapen mellom f-strukturane. Sidan avbildinga frå c-strukturnodar til f-struktur er mange-til-ein, kan me innanfor eitt tre ha fleire N-gram per f-strukturhovud; slike relasjonar får me ikkje fram i metoden til Samuelsson & Volk (2007). Eg vil i masteroppgåva prøve å samanlikne desse metodane, m.a. i forhold til dei ikkje-kontinuerlege konstituentane til språk som georgisk, og utfordringane ved samanstilling av språk med stor typologisk avstand.

1.1 Framgangsmåte og ressursar

I XPar (2008, s. 5–6) finn me følgjande hypotese:

FiXme Note: siter heller den nye artikkelen

> 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 båe f-struktur og c-struktur. Eg vil konsentrere meg om å gjere ei samanlikning mellom, på den eine sida, ein metode som nyttar enkel frasestrukturannotasjon kombinert med ei ordsamanstilling for å finne frasesamanstillinga (*tremetoden*, basert på Samuelsson & Volk (2007)); og på den andre sida ein metode som i tillegg nyttar f-struktur-informasjon frå desse parallelle grammatikkane (*LFG-metoden*).

Eg kjem til å nytte språka georgisk og norsk i samanlikninga, hovudsakleg fordi dei er svært ulike syntaktisk og morfologisk. Georgisk har t.d. mykje friare ordfølgje og rikare morfologi (inkludert valensaukande mekanismar som *applikativ*).

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 tremetoden (utan ein god del forarbeid). Difor kjem eg til å konsentrere meg om eit testkorpus kor eg manuelt gjer ordsamanstillinga. Eg veit heller ikkje enno om nokon statistisk parser av høg kvalitet for georgisk, men testkorpuset vil vere ferdig parsa med LFG-parseren frå Meurer (2008), slik at c-strukturane kan fungere som den syntaktiske annotasjonen i tremetoden. Oppgåva blir altså å finne ut kva for bidrag informasjonen på f-strukturnivået kan gi til samanstillinga, og kva for problem ein støyter på.

Eg vil prøve å implementere testversjonar av LFG-metoden for frasesamanstilling i eit passande programmeringsspråk.

1.2 Frasesamanstilling frå f-struktur

Men om me har f-strukturane til to omsette setningar, burde det kanskje vere mogleg å finne ei f-struktursamanstilling først og så finne ordsamanstillinga ut frå denne. Tanken er at me frå to f-strukturar som skildrar omsette setningar, kan

1. lage ei samanstilling mellom relevante deler av f-strukturane,

2. nytte denne funksjonelle samanstillinga til å finne ei frasesamanstilling, ved å følgje avbildinga frå f-struktur til c-struktur (ϕ^{-1}).

Eitt problem som byr seg er: kva for «deler av f-strukturane»? I det minste må me kunne kople det opp mot c-strukturnodar; så PRED-element bør i det minste ha lenkjer, medan t.d. tempus og aspektuell informasjon kanskje er mindre viktig. Men kva kan ignorerast? Vil det oppstå tilfelle då me bør vekte visse element? (Dvs., må me nokon gong disambiguere med slike andre element?)

Vidare må me vite *korleis* me samanstiller desse delene. Me kan t.d. byrje med å kople ytterste PRED frå kvart språk, og så rekursivt kople PRED i dei relevante substrukturane[2]. Gitt ein funksjon i som returnerer indeksen til ein f-(sub)struktur, kan eit førsteutkast til ei *f-samanstilling*, samanstilling på f-strukturnivå, sjå slik ut:

$$falign(f_1, f_2) = \{(i(f_1(\text{PRED})), i(f_2(\text{PRED})))\} \cup \bigcup_{g_1, g_2 \in fpairs(f_1, f_2)} falign(g_1, g_2)$$

falign vil gi ei mengd av par av indeksar, kor kvart par altså er samanstilt. Ein føresetnad her er at me i tillegg veit kva for par av substrukturar som er «relevante» ($fpairs(f_1, f_2)$).

Sjølv om f-strukturar abstraherer frå skilnadene i korleis ulike språk nyttar ordgruppering og ordform til å kode syntaktiske forhold (Bresnan, 2001, s. 14), vil det likevel oppstå forskjellar i f-strukturane til to parallellstilte setningar i eit korpus; både pga. «omsetjarfridom» og det at ulike språk nyttar ulike syntaktiske funksjonar til å uttrykkje det same konseptet. I f-struktursamanstillinga til Riezler & Maxwell (2006, s. 40) får dei t.d. ei lenkje frå ein XCOMP på tysk til eit OBJ på engelsk. Skal ein algoritme gå frå f-strukturar til frasesamanstilling må han i det minste vere robust nok til å takle slik mangel på samsvar. Til å byrje med kan me tenkje oss at *fpairs* gir alle par av GF-ar som har same plass i argumentstrukturen[3] til predikatet, så viss 'sein⟨SUBJ,XCOMP⟩' står i f_1 og 'have⟨SUBJ,OBJ⟩' i f_2, vil *fpairs* i det minste returnere $\{(f_1(\text{SUBJ}), f_2(\text{SUBJ})), (f_1(\text{XCOMP}), f_2(\text{OBJ})), ...\}$. Men om me ikkje har slikt samsvar i argumentstrukturar, vil *fpairs* ha ein vanskelegare jobb.

Eit større problem er nok adverbial (elementa i $\text{ADJUNCT}_{\{\}}$), kor f-strukturane ikkje gir like greie hint om kva for substrukturar som høyrer saman[4]. Ein del av masteroppgåva vil altså vere å komme med forslag til funksjonen *fpairs*.

f-samanstillinga kan nyttast til å gi ein samanstilling av frasane dei representerer. ϕ^{-1} gir no ei samanstilling mellom funksjonelle domene i c-strukturane, me har t.d. ei lenkje mellom domenet $d_1 = \{X, Y, Z\}$ på språk 1 og $d_2 = \{U, V, W\}$

[2]Dette krev sjølvsagt at ytre PRED faktisk korresponderer i samanstilte setningar, ein ikkje-triviell påstand.

[3]Ved å nytte argumentplass kan me enkelt få til lenkjer mellom GF-ar med ulike namn, som vist i dømet.

[4]Det er mogleg at f-samanstillinga av adverbial kan tene på informasjon frå (og difor bør skje etter) samanstillinga av frasane som projiserer argumentfunksjonane.

på språk 2. Kvar node frå d_1 vil kunne (symmetrisk) samanstillast med ein (eller ingen) frå d_2.

Her kan me utnytte det at frasestrukturane i dei ulike grammatikkane er tufta på same X-bar-prinsipp. Ein $XP \in d_1$ skal sannsynlegvis samanstillast med ein $YP \in d_2$ (der X og Y gjerne er same symbol, men au kan vere t.d. V og I). I tillegg skal høge nodar sannsynlegvis samanstillast med andre høge nodar, der alt anna er likt, medan mangel på samsvar i samanstillinga til døtre kan føre til at mornodar ikkje skal samanstillast; ein formalisering dette steget, med diskusjon rundt problema, vil au inngå i masteroppgåva.

Kapittel 2

Bakgrunn og relaterte metodar

- reine N-gram-samanstillingar, dependensbaserte
- ulike formål for samanstilling gir ulike metodar
- kort introduksjon til LFG

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. Dei ulike tilnærmingane som finst er prega av formåla til utviklarane.

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

Men det er hovudsakleg innanfor stokastisk maskinomsetjing at ein har forska på samanstilling av frasar. Koehn et al. (2003) gir ein grundig evaluering av ulike statistiske metodar for frasesamanstilling til bruk i stokastisk maskinomsetjing. Dei nyttar BLEU-systemet 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 båe 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 originale modell 4, utan frasesamanstilling). Dei forklarer dette med den store mengda uttrykk som ikkje utgjer syntaktiske konstituentar i følgje parseren 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, båe om ein som i *Syn*-modellen nyttar frasestrukturinformasjon[1], men i endå større grad om ein nyttar dependendsinformasjon (Hearne et al., 2008). F-strukturane til LFG gir ein slags dependensinformasjon.

Riezler & Maxwell (2006) utvikla ein metode for PBSMT med LFG-basert generering på output-sida. Dei finn ei n-til-m-ordsamanstilling med Giza++ som i metodane over, men parser 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.

[1] Samuelsson & Volk (2007) evaluerer sitt *Syn*-liknande system ved samanlikning med ein manuelt frasesamanstilt gullstandard.

Kapittel 3

Den ideelle frasesamanstillinga

3.1 SKRIV LPT :ROTETE:

«a source word WS and a target word WT are taken to correspond translationally only if (i) WT can in general (out of context) be taken to be among the semantically plausible translations of WS, i.e., WT belongs to the set of 'linguistically predictable translations (LPT)' of WS, and (ii) WS and WT occupy corresponding positions within corresponding argument structures.»

«a source phrase PHS and target phrase PHT are taken to correspond if (i) they contain corresponding words, (ii) PHS contains no word or phrase corresponding to a target word or phrase outside PHT, and similarly (iii) PHT contains no word or phrase corresponding to a source word or phrase outside PH.»

«It remains to be considered whether we should add the requirement that PHS and PHT also occupy corresponding positions within translationally corresponding argument structures, as we assume on the level of word correspondences.»

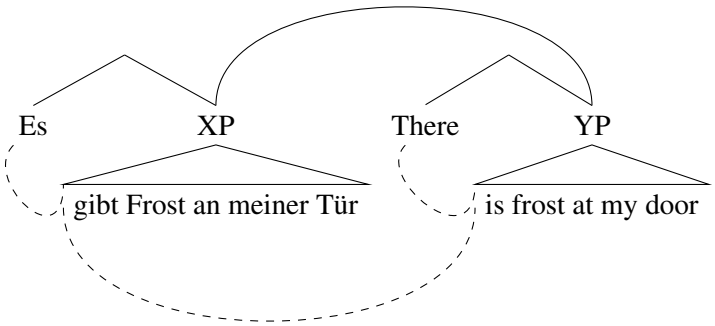
«possibly also eliminate some of the initial links.» – ie. non-monotonic phrase linking on top of the word linking.

3.2 Introduksjon

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, og tek utgangspunkt i visse krav for ordsamanstilling gitt i Thunes (2003). Eg kjem fram til at når formålet er utvikling av fasesamanstilte trebankar må ein revidere kravet om likskap i argumentstruktur, og gir eit forslag til krav for frasesamanstilling i trebankar.

3.3 Kva er formålet med ei frasesamanstilling?

I frasebasert statistisk maskinomsetjing (PBSMT) skal ei fraselenkje[1] forbetre 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-parser, ikkje var syntaktiske frasar (konstituentar). Dvs. at i figur 3.1 vil lenkja vist ved den prikkete lenkja bli fjerna frå mengda over moglege lenkjingar 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[2] er det for dei ingen grunn til å berre halde seg til samanstilling mellom syntaktiske konstituentar; dei har i utgangspunktet meir nytte av kollokasjonsinformasjon.



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? I einkvar større parallelltekst vil parallellstilte setningar ha visse syntaktiske og semantiske[3] omsetjingsskifte, t.d. leksikalisering av syntaktiske konstruksjonar eller omvendt, endring av ordklasse, presisering/depresisering, endringar i leksikale trekk (t.d. telleleg/utelleleg), osb. (Munday, 2001, s. 56–62), slik

[1] Eg nyttar her termane *lenkjing* og *samanstilling* om kvarandre, i same tyding som det engelske *alignment*; 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.

[2] 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).

[3] Sidan eg føreset setningssamanstilte data, kjem eg ikkje inn på diskurs-/pragmatiske verknader, med mindre det kan vere mogleg å handsame desse innanfor setningen.

at den einaste fullstendige, «perfekte» samanstillinga vil vere identitetsfunksjonen. Me må godta ein del mangel på samsvar; kor mykje me godtek blir då avgjort av formålet med samanstillinga.

Eg føreset her at eitt av formåla med samanstillinga 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 samanstille frasar med ulik kasusmarkering, men ha krav om lik tildeling av semantiske roller.)

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.

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

3.4 Krav / skrankar for frasesamanstilling i ein LFG-trebank

Samanstilte frasar bør ha nok semantisk likskap til å kunne opptre som omsetjingar i liknande omgivnader (**?**). Thunes (2003) gir nokre passande prinsipp for å fastslå det som kan kallast *omsetjingsmessig korrespondanse*, for ordsamanstilling. Dette er prinsipp som skal gjelde for eit litt forskjellig formål[4], 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 (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 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».

[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 leksikale 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.

Krava for ordsamanstillinga må au vere fylt for at desse frasane kan samanstillast. Ein ordsamanstilling er altså naudsynt for ein frasesamanstilling, og omvendt. Dette er berre motsetningsfylt 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, på same måte som dei modellteoretiske tolkingane av LFG og HPSG.

Pullum & Scholz (2001) gir ein god gjennomgang av forskjellen mellom derivasjonelle (enumerative) grammatikkar og skrankebaserte modellteoretiske grammatikkar, kor førstnemnde definerer *mengder av uttrykk* 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.

Sidan ei frasesamanstilling er ei skildring av forhold mellom setningsfragment vil det vere naturleg å skildre dei ønskelege forholda som skrankar på moglege samanstillingar. Dette let oss au setje skrankar på både frase- og ordsamanstilling sameleis, utan å måtte ha krav om at den eine samanstillinga er fullstendig avleiia av den andre; noko me ikkje har eit *a priori* grunnlag for å seie.

Sidan metoden er mynta på bruk i ein LFG-parsa trebank, og delvis vil nytte denne parsen 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.

Eg byggjar vidare på krava frå Thunes (2003) nedanfor, men kjem som nemnd med visse endringsforslag.

3.5 Kva kan samanstillast?

Viss to uttrykk er samanstilt på setningsnivå (slik at me dimed kan gå ut frå at dei er omsetjingar av kvarandre), og båe 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,

[5] I tillegg finst andre positive biverknader av ein LFG-basert frasesamanstilling for bruk i denne samanhengen, som at ein kan oppdage 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. mellom c-strukturnodar.

Alle ord i setninga er *kandidatar* for samanstilling med ord i omsetjinga, men *a priori* kan me ikkje utelukke at eit ord ikkje har ei lenkjing, og me kan heller ikkje utelate mange-til-mange-lenkjing. Det same gjeld nodane i c-strukturen.

FiXme Note: i tillegg vil samanstilling av andre trekk vere endå eit steg lenger vekk frå observerte data

Når det gjeld f-strukturane er det ganske mange element me teoretisk sett kunne ha samanstilt, t.d. enkelttrekk som bestemtheit eller dei uordna mengdene med adjunkt, men det som er mest *nyttig* er nok å berre gjere samanstillingar 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[6] alle desse PRED-elementa i f-strukturane til setningane. PRED-element representerer semantiske bidrag som oftare er naudsyne på båe 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 konkrete tekststrengen, medan t.d. aspekt kanskje er umogleg å skilje frå tempus i affikset.

Eg føreslår følgjande føringar:

(1) Ei samanstilling av to PRED-element i f-strukturane tilseier at:
   a. f-strukturane til desse er lenkja,
   b. orda i setningane som projiserer PRED-elementa tek del i ei samanstilling med kvarandre (kor andre ord kan vere involvert), og at
   c. iallfall dei øvste nodane i det funksjonelle domenet[7] til f-strukturen er samanstilt.

(Underordna nodar i det funksjonelle domenet kan berre lenkjast om visse krav, gitt nedanfor, er oppfylt. Me kan altså gjerne ha c-strukturnodar som ikkje er lenkja til andre nodar.)

FiXme Note: backe det med eksemplar i trebank; kople til adj-arg-lenkje

Påstandane over må forsvarast. Punkt (1-a) og (1-c) over seier at viss PRED-elementa projisert av t.d. to verb i verbfrasar er lenkja, vil *heile* VP-ane vere lenkja (både VP-nodane som dominerer dei lenkja funksjonelle domena og f-strukturane frå ytre PRED til verba), det er dette som gjer det til ei fraselenkje; medan i følgje punkt (1-b) vil denne fraselenkja leie til at sjølve verba au er lenkja, ein sterkare påstand sidan dette tilseier at *PRED-samanstilling impliserer ordsamanstilling*. 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:

[6] I del 3.8 kjem eg tilbake til spørsmålet om me vil inkludere visse f-strukturar utan PRED-element i kandidatane for samanstilling.

[7] 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 (fleire funksjonsargument som gir same verdi).

(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...*.

Som nemnd over; ordsamanstillinga treng ikkje vere ein-til-ein, det punkt (1-b) seier er at desse orda iallfall er ein del av ein samanstilling med kvarandre (i (2) altså VP-samanstillinga). Kanskje er dette ei mange-til-mange-lenkjing som ikkje *kan* reduserast til ein-til-ein-lenkjingar; eller kanskje er det som i (2) mogleg å skilje ut delsamanstillingar, som *have/har*. Eg kjem tilbake til dette i del 3.15 om argumentstruktur og adjunkt.

Alle nodar i c-strukturen (alle syntaktiske *frasar/konstituentar* i setninga) som kan koplast til PRED-haldande f-strukturar, vil altså vere kandidatar for samanstilling på c-strukturnivå (dette inkluderer diskontinuerlege konstituentar), men ikkje alle vil bli samanstilt.

FiXme Note: der ADJUNKT ikkje er realisert, lenkjer me ikkje PRED. skal me då ikkje lenkje ord heller?
FiXme Note: PRED->ord :: iallfall
PRED<-ord :: ?
PRED<->ord
PRED, ord
FiXme Note: avsnittet over er litt rotete
TODO

3.5.1 TOGROK finst det tilfelle der ordlenkjer ikkje impliserer PRED-lenkjer?

hypotese: det er alltid slik at
ordlenkjing av predikerande ord => PRED-lenkje

3.6 TOGROK kva med ekspletivar? ingen PRED men heller ikkje C/F/I :ROTETE:

Kandidatane på f-strukturnivå må jo inkludere desse au...

3.7 TODO Gi enkelt døme kor alt fungerer :ROTETE:

3.8 Funksjonsord

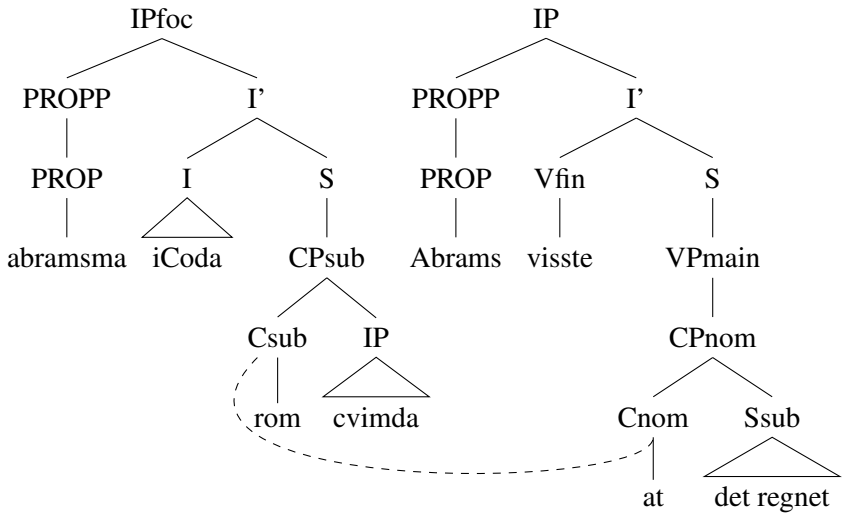
I tillegg kan me ha ord i setninga som ikkje tilsvarer PRED-element i f-strukturen, typisk funksjonsord (t.d. *som*, *at*). Ved endosentrisitetsprinsippa til Bresnan (2001) er komplementet til funksjonelle kategoriar (C, I, P) ein funksjonell ko-kjerne.

(3) Skal nodar for ord som ikkje projiserer PRED-element[8] samanstillast, må følgjande krav vere oppfylt:
a. det funksjonelle domenet (gitt ved komplementet) må vere samanstilt, og

[8] Skal ein lenkje ordet *som* (utan PRED) med ordet *which* (med PRED)? Viss båe står under C i treet, 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).

b. dei er båe c-strukturhovud.

Om (3-a og -b) er oppfylt, kan me få samanstillinga vist i figur 3.2, og i dette tilfellet er (3-b) oppfylt og (3-a) vil vere oppfylt om me kan samanstille *cvimda* med *det regnet*.



Figur 3.2: Mogleg samanstilling av funksjonsord mellom georgisk og norsk (bokmål)

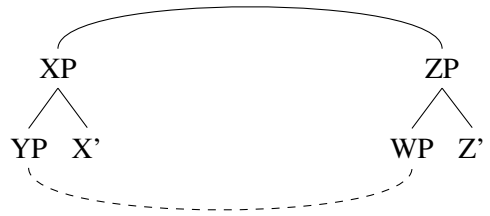
3.8.1 TOGROK cvimda<PRO> men regne<>expletive – lenkje? :ROTETE:

3.9 Lenkjing av 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 speler 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 f-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-strukturnodar

(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:

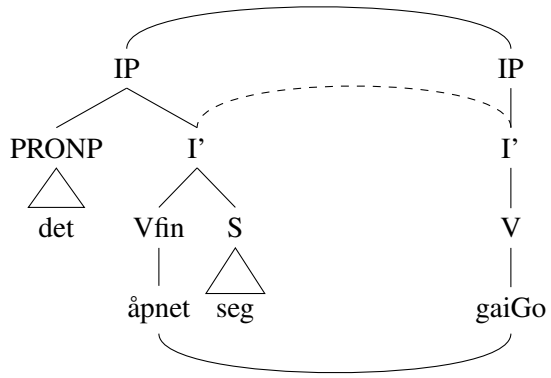
(4) Krav for lenkjing av underordna c-strukturnodar:

- a. c-strukturnodar som ligg under øvste node i to funksjonelle domena kan berre samanstillast med nodar som ligg innanfor desse domena,
- b. c-strukturnodar kan berre samanstillast om deira funksjonelle domene er lenkja på f-strukturnivå,
- c. 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 samanstille X' og Z'

(4-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. (4-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 3.2 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 (4-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 (3), og (4-c) er då oppfylt. (4) står altså ikkje i vegen for å samanstille IP-en over *cvimda* og Ssub.

I figur 3.4 derimot (**?**), kan me ikkje samanstille 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

3.9.1 SKRIV døme! :ROTETE:

3.9.2 TOGROK me$_{\text{OBJ}}$ gusta X$_{\text{SUBJ}}$ // I$_{\text{SUBJ}}$ like X$_{\text{OBJ}}$?? :ROTETE:

3.9.3 TOGROK korleis finn me *there is*-lenkjer då? :ROTETE:

(og kva skal me med dei?)

«Til gjengjeld vil me få lenkjer sjølv om me har mellomståande ord (*There* never *is*) som opptrer utanfor N-grammet på det andre språket.»

3.10 TOGROK mange-til-mange-lenkjing i f-strukturane? :ROTETE:

Eg er litt usikker på om me skal ha slike mange-til-mange-korrespondansar i f-strukturane; eg har rekna med at ei f-strukturlenkje *impliserer* ei slags lenkjing mellom det som er innanfor f-strukturane; men i Riezler & Maxwell (2006) er det i staden berre eit krav om at desse f-strukturane er lenkja i same transfer-regel.

Riezler & Maxwell (2006, s. 40–41) tillet mange-til-mange-lenkjing mellom f-strukturar, så lenge alle f-strukturane som blir lenkja til slutt opptrer i same transfer-regel. Frå følgjande setningspar:

(5) Dafür bin ich zutiefst dankbar
I have a deep appreciation for that

lenkjar dei {*zutiefst*} med { *a, deep, appreciation* }, men sidan {*appreciation*} er samanstilt med {*dankbar*}, må transfer-regelen inkludere { *zutiefst, dankbar* } på den eine sida og { *a, deep, appreciation* } på den andre.

3.10.1 SKRIV Kva inneber ei mange-til-mange-lenkjing? :ROTETE:

3.11 SKRIV Mangel på samsvar i syntaks og semantikk :ROTETE:

(Kruijff-Korbayova et al., 2006, s. 5) gir følgjande døme:

(6) nikdy nebyl
never was.not
'has never been'

nebyl blir «svakt» samanstilt med *never*, men «sterkt» samanstilt med *has ... been* i deira system. I tillegg er det ein sterk samanstilling mellom *never* og *nikby*.

3.12 TOGROK Diskontinuerlege einingar :ROTETE:

- diskontinuerlege einingar (Cheung et al., 2002, s. 4) @books.google – skal dei eigentleg samanstillast? Kva for problem gir dei i forhold til c-strukturnivåsamanstilling?

3.12.1 TODO døme på diskontinuerlege konstituentar som er lenkja :ROTETE:

3.13 TOGROK Er «compounds» frasar? :ROTETE:

(Giegerich, 2006, p. 1)

3.14 Lik ordklasse?

Ulike språk leksikaliserer same konsept på ulike måtar. Cheung et al. (2002, s. 3) skriv at det engelske ordet *fulfilment* meir naturleg blir omsett til eit verb på kinesisk. Det same gjeld t.d. *solitude* omsett til norsk. Eit georgisk verbalsubstantiv (*masdar*) kan bli omsett til eit verb i infinitiv på norsk[9]. Slike skifte mellom ordklassar er svært vanlege i omsetjing[10].

Me kan opne for ordklasseoverskridande lenkjer der det er samsvar mellom visse *trekk*, t.d. kan to predikerande ord lenkjast, eller to «nominale» ord. Ein annan måte å gjere dette på er rett og slett å krevje ein viss likskap i argumentstruktur.

[9] Det georgiske verbalsubstantivet (*masdar*) er i følgje Aronson (1990, kap. 2.5) ein *nominal* form, det kan i motsetning til norske verbalsubstantiv og engelske gerundium ikkje ta objekt, men kan ha modifiserande substantiv i genitiv.

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

3.15 Krav om lik argumentstruktur

Thunes (2003) gir som nemnd eit krav om at *predikat må ha tilsvarande 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 samanstille predikata, og argument med argument, mykje enklare. Men kan me stille så sterke krav?

Sett at ein 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 samanstilte om dei opptrådde åleine. Om dei uttrykkjer same proposisjon og *speler same rolle i verbsituasjonen*, synest det naturleg å lenkje desse.

Omsetjingsrelasjonar gir data for verbsituasjon, 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 (om det er mogleg å finne akseptable syntaktiske grunnar for å kalle noko ein adjunkt heller enn eit argument).

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

(7) 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[12]. 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 (8) nedanfor har eg representert dei omsetjingsmessig korresponderande frasane i f-strukturane med dei norske omsetjingane for å illustrere dette:

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

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

[11] Setningane i første og tredje linje i døma er direkte henta frå MRS-suiten, med mindre anna er opplyst.

[12] 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).

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

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

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

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

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

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

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 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*.

3.15.1 forsvare «tilsvarande» :ROTETE:

Tilsvarande på engelsk: [13]

3.15.2 TODO Sitere eigen korpusundersøkjing av variasjon i arg-str?

Ei undersøkjing av den frasesamanstilte trebanken SMULTRON (Samuelsson & Volk, 2006) mot LFG-grammatikkane for engelsk og tysk fann at 2 av 15 korresponderande verbtoken[14] for høgfrekvente innhaldsverb fekk analysar kor argument korresponderte med adjunkt (**?**).

FiXme Note: LCS, dorr

[13] "wagered * with * that *" på Google gir 215 treff, kor 9 av dei første 10 følgjer det intenderte mønsteret.

[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).

3.15.3 SKRIV kvifor lik arg-str er bra, så kvifor det er eit problem :ROTETE:

3.15.4 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 opplevar som objekt og tema som subjekt vil ha opplevar nedanfor tema i argumentstrukturen, medan ei omsetjing av dette verbet kan ha tema nedanfor:

(9) 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

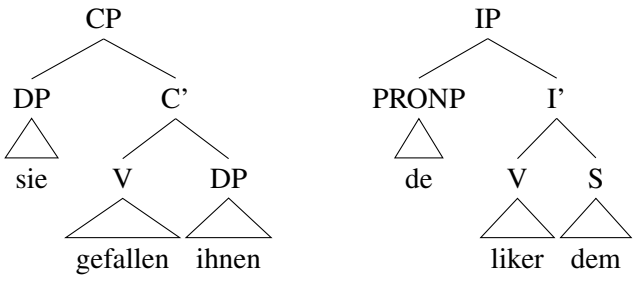
$$\left[\text{PRED} \quad \text{`}\textbf{like}\text{<de}_i\text{, de}_j\text{>'}\right]$$

Argumentstrukturane i (9) 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.

TODO Flytte til kapittel om metodar for å oppdage lenkjer?:

Kanskje me kan 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.).

c- og f-strukturar for dømet over :ROTETE:



$$
\begin{bmatrix}
\text{PRED } \text{‘}\textbf{gefallen}\text{<}[1{:}\textit{pro}],[2{:}\textit{pro}]\text{>’} \\
\text{TOPIC}_{1}\begin{bmatrix}
\text{PRED } \text{‘}\textit{pro}\text{’} \\
\text{NTYPE}_{7}\begin{bmatrix}\text{NSYN PRONOUN}\end{bmatrix} \\
\text{PRON-TYPE PERS, PRON-FORM SIE, PERS 3,} \\
\text{NUM PL, CASE NOM}
\end{bmatrix} \\
\text{TNS-ASP}_{4}\begin{bmatrix}\text{TENSE PRES, MOOD INDICATIVE}\end{bmatrix} \\
\text{OBJ-TH}_{2}\begin{bmatrix}
\text{PRED } \text{‘}\textit{pro}\text{’} \\
\text{NTYPE}_{10}\begin{bmatrix}\text{NSYN PRONOUN}\end{bmatrix} \\
\text{PRON-TYPE PERS, PRON-FORM SIE, PERS 3,} \\
\text{NUM PL, CASE DAT}
\end{bmatrix} \\
\text{SUBJ } \begin{bmatrix}1\end{bmatrix} \\
\text{VTYPE MAIN, STMT-TYPE DECL,} \\
\text{PASSIVE -, CLAUSE-TYPE DECL}
\end{bmatrix}_{0}
$$

$$
\begin{bmatrix}
\text{PRED } \text{‘}\textbf{like}\text{<}[10{:}\text{DE}],[11{:}\text{DE}]\text{>NULL’} \\
\text{TNS-ASP}_{13}\begin{bmatrix}\text{TENSE PRES, MOOD INDICATIVE}\end{bmatrix} \\
\text{TOPIC}_{10}\begin{bmatrix}
\text{PRED } \text{‘}\textbf{de}\text{’} \\
\text{NTYPE}_{18}\begin{bmatrix}\text{NSYN PRONOUN}\end{bmatrix} \\
\text{DEF +, CASE NOM, REF +,} \\
\text{PRON-TYPE PERS, PRON-FORM DE, PERS 3,} \\
\text{NUM PL}
\end{bmatrix} \\
\text{OBJ}_{11}\begin{bmatrix}
\text{PRED } \text{‘}\textbf{de}\text{’} \\
\text{NTYPE}_{45}\begin{bmatrix}\text{NSYN PRONOUN}\end{bmatrix} \\
\text{REF +, PRON-TYPE PERS, PRON-FORM DE,} \\
\text{PERS 3, NUM PL, DEF +,} \\
\text{CASE OBL}
\end{bmatrix} \\
\text{SUBJ } \begin{bmatrix}10\end{bmatrix} \\
\text{VTYPE MAIN, VFORM FIN, STMT-TYPE DECL}
\end{bmatrix}_{0}
$$

3.15.5 SKRIV døme med wager/3 og vedde/4 og gewettet/3 :ROTETE:

3.15.6 SKRIV (reinskriv) :ROTETE:

Same globale tyding krev i det minste at, i situasjonen verbet denoterer, speler deltakarane same rolle. Men dette er endå meir abstrakt/semantisk enn (semantisk) argumentstruktur...

Problem: ikkje-komposisjonell omsetjing. Same globale tyding. Det treng ikkje vere berre pragmatisk forskjell–type *kan du lukke døra* vs *lukk døra*, kor situasjon gjer setningane like–sidan me kan ha konvensjonaliserte konstruksjoner på L1 kor heile tilsvarer enkeltord på L2, a la japansk *viss eg ikkje går på skulen så kan det ikkje vere ˜= eg må gå på skulen*.

Ein føresetnad eg har, er at setningar som er samanstilte faktisk har ein omsetjingsmessig korrespondanse (dette er min data). Så om eit par av ytre predikat ikkje korresponderer er det au ein type data; nemleg at me har ein omsetjingsmessig korrespondanse der det var ein mismatch i ytre argumentstruktur. (Algoritmen bør då lagre slike mismatches eksplisitt, ikkje berre la vere å lenkje, for det kan vere andre grunnar til at det ikkje kom ei lenkjing. A la ekspertsystem: forklare resonnementet.)

Alternativt ein konstruksjonslenkjing...

Kan au ha eit krav om at argstr til $PRED_{L1}$ er ein slags delmengd av argstr til $PRED_{L2}$.

3.15.7 SKRIV True Arguments vs True Adjuncts, Pustejovsky :ROTETE:

- Treng døme først...
- Er «with Browne» eit Default Argument for «wager»?
  - D-ARG: he built a house out of bricks
- Adjunkt plukker ut sine eigne roller, per definisjon, ved vedde/4 og wager/3 har me ein slik situasjon:

```
vedde <----------wager >------<------wetten
       \____with_/     \__dass
```

  Bottom-up-informasjon vil au vere naudsynt for dei 3 rollene som *er* argument, sidan me kan ha vedde<1,2,3,4> og wager<a,b,c>with<d>, kor det er umogleg å seie om d skal på plass 1,2,3 eller 4 (dvs. me kan ha vedde<a,b,c,d>, vedde<a,b,d,c>, vedde<a,d,b,c> og vedde<d,a,b,c> – men sannsynlegvis er altså a,b,c i same rekkjefølgje uansett...)

3.16 SKRIV Kan adjunkt lenkjast til nodar under morlenkja?

Krav (vi) i **?**, s. 5 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 *adjunkt*dotter av F_t, så vil døtre av D_t kunne lenkjast til D_s, så lenge D_t forblir ulenkja.

3.17 TOGROK kva var poenget med dette? :ROTETE:

«etter og uten er dei einaste prep som tek setn utan å vere arg»

3.18 ULEST Cyrus, FuSe-prosjektet :ROTETE:

Cyrus et al. (2004) «Abstract: We report on a recently initiated project which aims at building a multi-layered parallel treebank of English and German. Particular attention is devoted to a dedicated predicate-argument layer which is used for aligning translationally equivalent sentences of the two languages. We describe both our conceptual decisions and aspects of their technical realisation. We discuss some selected problems and conclude with a few remarks on how this project relates to similar projects in the field.»

3.19 TODO Konstruksjonar og komposisjonell inekvivalens

$\overline{\text{X}}$-teori føreset at det finst éi dotter i kvart ledd som kan reknast som predikatet for dette leddet. Ei utfordring for $\overline{\text{X}}$-baserte teoriar er då handsaming av *komplekse predikat*. Desse har fleire grammatiske element innanfor same ledd som alle bidrar med «a non-trivial part of the information of the complex predicate» (Alsina et al., 1997). I LFG er det ein føresetnad at me berre har éin PRED ytterst i kvar f-struktur; ulike mekanismar har blitt føreslått for å handsame dette fenomenet.

I omsette tekster kan me få eit analogt problem:

(10) It can't be done
Det lar seg ikke gjøre

Her vil ytre predikat i f-strukturen på norsk vere 'la<det_1,XCOMP>PRO', kor XCOMP[PRED 'gjøre<NULL,det_1>NULL'].

På engelsk får me ‘can<XCOMP,it$_2$>’, kor XCOMP[PRED ‘do<NULL,it$_2$>’].

Skal me lenkje orda *can* og *la*? På *heile konstruksjonen* finn me iallfall eit omsetjingsforhold:

It can't be done	Det lar seg ikke gjøre	
can't be done	lar seg ikke gjøre	
be done	gjøre	s?
_ can't be VPASS	_ lar seg ikke VPASS	??
_$_1$ can _$_2$ be VPASS$_3$	_$_1$ lar seg _$_2$ VPASS$_3$	??

(kan me få den siste generaliseringa frå trebanken?)

3.20 SKRIV definer sitering frå MRS-suiten :ROTETE:

3.21 SKRIV setning 7 i MRS-suiten :ROTETE:

Ein samanstilling bør i det minste gi følgjande:

abramsi brouns daenajleva sigaretze, rom cvimda	Abrams veddet en sigarett med Brown på at det regnet
abramsi brouns daenajleva sigaretze	Abrams veddet en sigarett med Brown
brouns daenajleva sigaretze	veddet en sigarett med Brown
daenajleva sigaretze	veddet (en) sigarett (på)
daenajleva	veddet
sigaretze	(en) sigarett (på)
rom cvimda	at det regnet
cvimda	(det) regnet
abramsi	Abrams
brouns	Brown

3.22 TOGROK og så finst jo større forskjellar, stilistiske osb... :ROTETE:

3.23 TOGROK prosessering, kognitive modellar? :ROTETE:

finne empiri frå korleis menneske samanstillar? (dvs., korleis skjer omsetjing)

- Maier (2009), http://linguistlist.org/issues/20/20-1786.html

«cross-linguistic structural phenomena in the language production of bilinguals in the specific context of translation.»

- http://www.linguistlist.org/pubs/diss/browse-diss-action.cfm?DissID=143

- books.google bialystok????lpb: «Translation has been called "interlanguage paraphrase"», «a metalinguistic skill». «Paraphrasing consists in finding the meaning of two compared sequences and showing its equivalence, and this identification constitutes a judgment on the sequences»[s.˜151]

- books.google house????iic: «The process of translation, particularly if successful, necessitates a complex text and discourse processing. The process of interpretation performed by the translator on the source text might lead to a TL text which is more redundant than the SL text. This argument may be stated as "the explicitation hypothesis", [...] especially marked in the work of "non professional" translators» [s.˜19–20]

- Hutchinson: «What is a grammatical sentence?» (vanskeleg å unngå *talaren* i akseptabilitetvurderingar); kva er ei frasesamanstilling, sånn ute i naturen?

3.24 TOGROK Retningslinjer for samanstilling :ROTETE:

Ved korpusbygging er det vanleg at retningslinjer for samanstilling blir utvikla *etter kvart som ein finn problem...* (det er vanskeleg å seie noko *a priori* om kva for vanskar ein kan finne).

Kapittel 4

Korleis fungerer implementasjonen min

Programmet `lfgalign` [1] tek inn LFG-analysane av to setningar som me veit er omsetjingar av kvarandre. LFG-analysane må vere disambiguerte og i Prolog-formatet frå XLE[2]. Programmet les inn dei to filene og oppretter ein intern representasjon av LFG-analysen.

4.1 gjer ikkje dette lenger :ROTETE:

`lfgalign` kan i tillegg ta inn ein representasjon av moglege LPT-korrespondansar.

LPT-korrespondansane utgjer utgangspunktet for samanstillinga.

Der det ikkje finst informasjon om eit ord i LPT-basen, føreset me at alle ord kan lenkjast til dette. Pronomen/pro-element og substantiv/pronomen kan alltid lenkjast.

`permute` finn alle moglege permutasjonar av 1-1 LPT-korrespondansar.

Dette blir sendt gjennom `merge` som finn moglege mange-til-mange-korrespondansar av PRED-element. Moglege samanføyingar er: PRED til (X)COMP-dotter, ADJUNCT, ...

Så blir den utvida mengda med PRED-korrespondansar sendt vidare til eit filter som sjekker om skrankane er oppfylte.

4.2 fullstendig bottom-up

Eitt alternativ er å byrje med alle moglege permutasjonar av LPT-korrespondansar, og så sile ut dei som ikkje svarer til krava. Men dette er problematisk, sidan avskjeringa skjer så seint at utrekningane for lengre setningar blir ganske umogleg.

[1] Tilgjengeleg frå http://github.com/unhammer/lfgalign under GNU General Public License.

[2] Dokumentert på http://www2.parc.com/isl/groups/nltt/xle/doc/xle.html

Me må i alle tilfelle vere klar for ei setning der alle ord er ukjende (ingenting er LPT), slik at kvart kjeldeord kan lenkjast til kvart målord. Viss båe 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.

4.3 min metode

(iii) og (iv) i **?**, s. 5 krev LPT-korrespondanse mellom (L av Pr av) kvart kjelde-argument og målargument/-adjunkt. Eg krev i tillegg at f-strukturane deira kan samanstillast.

4.4 merge

filene

```
((tab_s (open-and-import "dev/TEST_argadj_s.pl"))
 (tab_t (open-and-import "dev/TEST_argadj_t.pl")))
```

viser at me kan trenge samanføying av pred på ulike nivå.

"sigaretten" og "sigaretze" er ikkje på same nivå i dei respektive f-strukturane, me har

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

og

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

Kapittel 5

Resultat av å automatisk samanstille norske og georgiske setningar

- om kjeldematerialet
- manglar med implementasjonen
- samanlikning av lenkjing basert på f-struktur og lenkjing basert på N-gram

5.1 TOGROK korleis gjenfinne there is/es gibt? :ROTE-TE:

1. N-gram kjem like ofte som heile konstruksjonen, då kan dette gjenfinnast
    - dvs., *there is NP/es gibt NP*-samanstilling kjem like ofte som *there is* eller *es gibt* førekjem. Eit TigerXML-type søk etter *there is NP/es gibt NP* burde jo vere mogleg, sjekk om dette er delmengd av *there is/es gibt*. * Avslutning

Referansar

Alsina, A., Bresnan, J. & Sells, P. (red.). (1997). *Complex predicates*. Stanford, CA, USA: Center for the Study of Language and Information. Paperback.

Aronson, H. (1990). *Georgian. A Reading Grammar. Corrected Edition*. Columbus, OH: Slavica Publishers.

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., 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

Cyrus, L., Feddes, H. & Schumacher, F. (2004). Annotating predicate-argument structure for a parallel treebank. *LISBON*, *2004*, 39. Tilgjengeleg frå http://arxiv.org/abs/cs/0407002

Giegerich, H. (2006). Attribution in English and the distinction between phrases and compounds'. *Englisch in Zeit und Raum-English in Time and Space: Forschungsbericht für Klaus Faiss. Trier: Wissenschaftlicher Verlag Trier*. Tilgjengeleg frå http://www.englang.ed.ac.uk/people/attributioninenglish.pdf

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

Kruijff-Korbayova, I., Chvatalova, K. & Postolache, O. (2006). Annotation guidelines for Czech-English word alignment. , 1256–1261. Tilgjengeleg frå http://www.mt-archive.info/LREC-2006-Kruijff.pdf

Maier, R.M. (2009). *Structural Interference from the Source Language: A psycholinguistic investigation of syntactic processes in non-professional translation*. Upublisert akademisk avhandling, University of Edinburgh. Tilgjengeleg frå http://linguistlist.org/issues/20/20-1786.html

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

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.)