Syntaktisk informert frasesamanstilling

Kevin Brubeck Unhammer

26/03, 2010

Kapittel 1

Introduksjon (+ samandrag/abstract)

Kapittel 2

Bakgrunn og relaterte metodar

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

Kapittel 3

Den ideelle frasesamanstillinga

3.1 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.

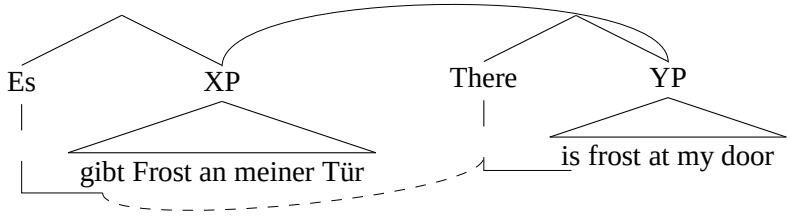
3.2 Kva er formålet med ei frasesamanstilling?

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 $p(\textit{es gibt}, \textit{there is})$ 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]

[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).

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. (**?**, s. 56–62), slik 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 (2009, 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.

3.3 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

[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.

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».

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

[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.

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.

3.4 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,
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.

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

[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).

[6] I del 3.5 kjem eg tilbake til spørsmålet om me vil inkludere visse f-strukturar utan PRED-element i kandidatane for samanstilling.

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.)

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:

(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 3.8 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.

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

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

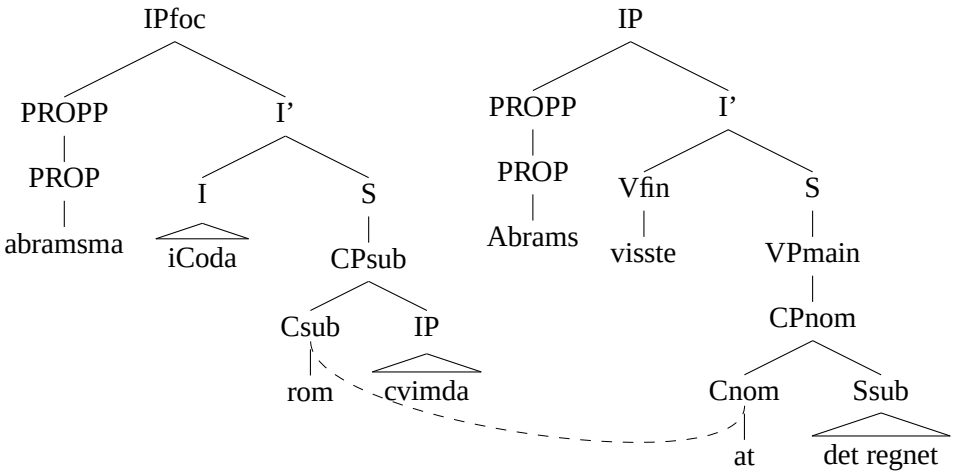
[7] Det funksjonelle domenet til ein f-struktur er gitt ved ϕ^{-1}, inversen av c-til-f-strukturavbildinga, og tilsvarer dei nodane i c-strukturen som projiserer denne f-strukturen, t.d. ein VP-node med dominerande IP og CP (Bresnan, 2001, s. 126). Sidan dette er inversen av ein funksjon, kan me ha diskontinuerlege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

3.5 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
   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)

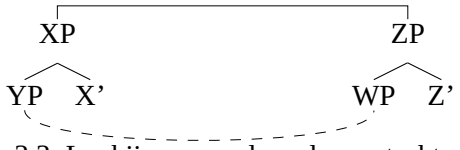
3.6 Lenkjing av underordna c-strukturnodar

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 un-

[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).

derordna 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?



Figur 3.3: Lenkjing av underordna c-strukturnodar

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 (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:

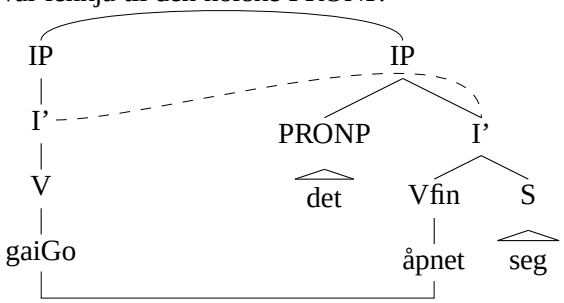
a. c-strukturnodar som ligg under øvste node i to funksjonelle domena kan berre samanstillast med nodar som ligg innanfor desse domena,
b. c-strukturnodar kan berre samanstillast om deira funksjonelle domene er lenkja på f-strukturnivå,
c. om ein c-strukturnode X' som ikkje er toppnode i det funksjonelle 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 georgisk og norsk (bokmål)

3.7 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.

3.8 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

[9] Det georgiske verbalsubstantivet (*masdar*) er i følgje **?**, 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] **?**, s. 61 gir ein gjennomgang av slike *klasseskifte*, og andre typar omsetjingsskifte.

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]:

(5) 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 (6) nedanfor har eg representert dei omsetjingsmessig korresponderande frasane i f-strukturane med dei norske omsetjingane for å illustrere dette:

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

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

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

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

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

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

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

[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).

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

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*.

3.8.1 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[13] for høgfrekvente innhaldsverb fekk analysar kor argument korresponderte med adjunkt (**?**).

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

(7) a. sie_j gefallen ihnen_i

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

$\leftrightarrow$

b. de_i liker dem_j

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

Argumentstrukturane i (7) har omvendt intern følgje, og som vist ved dette

[13] 25 om ein inkluderer analysar kor minst eitt av argumenta ikkje hadde korrekt analyse (t.d. eit PRO der grammatikken burde funne eit substantiv).

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.).

3.9 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» (**?**). 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:

(8) 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?)

Kapittel 4

Korleis fungerer implementasjonen min

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

Kapittel 6

Avslutning

Referansar

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

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

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

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

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

XPar. (2009). *Project description.* Tilgjengeleg frå http://ling.uib.no/lamore/xpar/index.php?page=description