Mot en trebank for talespråk

Victoria Rosén

Innledning

Norsk Talespråkskorpus – Oslodelen (NoTa-korpuset i det følgende) er en viktig ny ressurs for forskning om norsk talespråk. Det er mange trekk ved korpuset som gjør det særlig verdifullt for lingvistisk forskning: den ortografiske transkripsjonen, den morfologiske taggingen, koblingen til videoopptak av intervjuene, søkegrensesnittet osv.

En ting som mangler, og som man håper å få til etter hvert for korpuset, er syntaktisk annotasjon. Mange forskningsspørsmål kan bare besvares hvis man har et syntaktisk annotert korpus. For eksempel er alle spørsmål relatert til hypoteser om syntaktiske forskjeller mellom skriftspråk og talespråk vanskelige å undersøke uten syntaktisk annotasjon.

Mange av talerne ved "Oslomålet – et seminar med forskning fra NoTa-korpuset" holdt innlegg om syntaktiske fenomener som de hadde undersøkt ved bruk av korpuset. Flere av dem var opptatt av fenomener som de mente var vanligere i talespråk enn i skriftspråk. Eskil Hanssen skrev i sitt foredragssammendrag at ekstraponering i spørresetninger er "lite vanlig i skrift, men svært vanlig i tale", og illustrerte med eksempelet i (1).

(1) Fisk, fikk dere kjøpt det på kaia?

Jan Terje Faarlund var opptatt av syntaktisk endring, og undersøkte flere syntaktiske variabler hos to aldersgrupper (under 25 og over 70). Fenomener som er uvanlige i skriftspråk, kan være vanlige i talespråk bl.a. hvis det er tale om syntaktiske endringer, som normalt forekommer først i talespråk før de blir akseptert som en del av skriftnormen. Én av variablene som Faarlund undersøkte, var plasseringen av subjekt og setningsadverbial i leddsetninger, med de to mulighetene for leddstilling som vi ser i (2) og (3).

(2) Hvis PC-en ikke virker, gråter jeg.
(3) Hvis ikke PC-en virker, gråter jeg.

Johannessen, Janne Bondi og Kristin Hagen (red.). 2008: Språk i Oslo. Ny forskning omkring talespråk. Oslo: Novus forlag. ISBN: 978-82-7099-471-7.

Marit Julien var interessert i leddstillingen i innføyde setninger. Hun påpekte at i de germanske språkene som har leddstillingen i (4) i innføyde setninger, forekommer det likevel i visse innføyde setninger at man får den leddstillingen som er vanlig i hovedsetninger, slik som i (5).

(4) Men det som er er at han ikke kan lage sanger.
(5) Men det som er er at han kan ikke lage sanger.

Selv om det går an å finne ut noe om konstruksjoner som disse i (1)–(5) ved hjelp av et korpus tagget med morfosyntaktiske kategorier, sier det seg selv at det vil være mye lettere med en mer fullstendig annotasjon av syntaktiske strukturer.

I denne artikkelen vil jeg først gi en kort oversikt over trebanker. Dernest vil jeg si litt om syntaktisk annotasjon av talespråk. Hoveddelen av artikkelen vil dreie seg om LFG Parsebanker, et verktøy for syntaktisk annotasjon av korpora som er under utvikling ved AKSIS og Universitetet i Bergen.

Trebanker

Et korpus med syntaktisk annotasjon utover morfosyntaktiske tagger kalles en trebank. Termen *trebank* reflekterer at den vanligste formen for syntaktiske strukturer er syntaktiske trær. Men trebanker kan også ha andre former for syntaktisk og også semantisk annotasjon, for eksempel annotasjon av grammatiske funksjoner, dependensstukturer og predikat-argument-strukturer (Nivre, De Smedt og Volk 2005).

Det finnes ulike måter å lage trebanker på. Den eldste måten er å annotere de syntaktiske strukturene manuelt. Det er særlig to store utfordringer knyttet til manuell annotasjon. For det første er det svært arbeidsintensivt og dermed også dyrt. For det andre krever det et stort arbeid når det gjelder å spesifisere hvordan annotasjonen skal gjøres. Det er nemlig et stort problem med slik annotasjon at ulike mennesker annoterer på ulike måter. Det kan også være vanskelig for én og samme annotator å annotere like fenomener på samme måte over tid.

En annen måte å konstruere en trebank på, er å bruke automatisk analyse. Hvis korpuset parses automatisk, er det essensielt at det er en manuell valideringsprosess for å sikre kvaliteten. En grunn til dette er at naturlige språk har så mye leksikalsk og syntaktisk flertydighet at automatisk analyse ofte vil resultere i mange analyser. Det må da menneskelig medvirkning til for å finne den intenderte analysen. Det er heller ikke sikkert at den intenderte analysen finnes blant analysene som parseren leverer; det kan

hende at det er feil eller mangler i grammatikken som parseren bruker. Også i dette tilfellet er det viktig at det er et menneske som avgjør om den intenderte analysen foreligger. Den største fordelen med automatisk analyse er at analysene er konsistente med grammatikken. Det er mye lettere å sikre at like konstruksjoner får lik annotasjon med automatisk parsing enn med manuell annotering. En annen fordel er at komplekse strukturer kan bygges opp mer effektivt av en grammatikk enn av en annotator.

Trebanker er blitt en viktig ressurs for naturlig språkprosessering. De brukes blant annet som såkalte gullstandarder for evaluering av systemer som parsere og taggere, og som treningsmateriale for maskinlæringssystemer. Men trebanker er også en viktig ressurs for "vanlige" språkforskere. En nyttig oversikt over trebanker, hvordan de lages og hva de brukes til, finnes i Abeillé (2003).

Annotasjon av talespråk

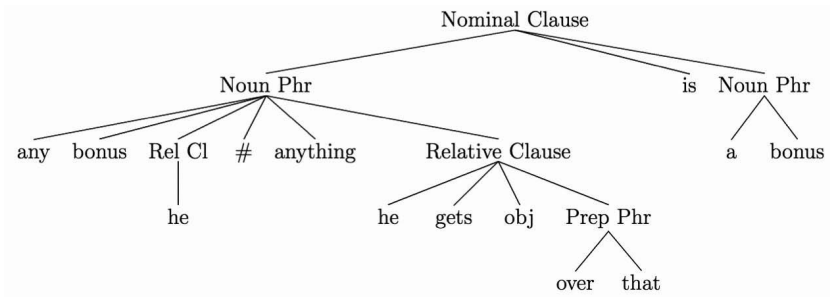
Annotasjon av talespråk innebærer noen flere utfordringer enn annotasjon av skriftspråk. For et talespråkskorpus er transkripsjon et nødvendig første skritt mot videre annotasjon. NoTa-korpuset er transkribert i standardisert ortografi, noe som gjør det lett å bruke en morfologisk tagger på materialet. Transkripsjonsmåten har nok tildels vært styrt av hva som er nyttig input til taggeren. For eksempel er det en hovedregel for transkripsjonen at det bare skal brukes former som forekommer i *Bokmålsordboka* (se Hagen 2005).

Transkripsjonen alene gjør det ikke mulig å bruke en parser på korpuset. Skriftspråk deles opp i klare syntaktiske enheter ved at setninger begynner med store forbokstaver og avsluttes med spesielle tegn: punktum, spørsmålstegn eller utropstegn. Selv om det er mange tvilstilfeller (egennavn begynner også med store forbokstaver, og mange forkortelser slutter med punktum), er det stort sett mulig å konstruere automatiske redskaper som deler skrevne tekster opp i setninger. Setninger markeres ikke som sådanne i NoTa-korpusets transkripsjon. Noen spørsmål markeres med spørsmålstegn, men ellers brukes ikke tegnsetting, og heller ikke store forbokstaver i begynnelsen av setningen. Den oppdelingen i enheter som er foretatt i NoTa-transkripsjonen, er hovedsakelig en oppdeling i samtaleturer, som ofte ikke tilsvarer syntaktiske enheter. For at korpuset skal kunne parses automatisk, er det imidlertid nødvendig med en oppdeling i syntaktiske enheter. Naturlige språk inneholder nemlig så mye leksikalsk og syntaktisk flertydighet at det vil være umulig å få brukbare resultater fra automatisk parsing hvis materialet ikke er delt opp i syntaktiske enheter, se diskusjon om dette under avsnittet *Interaktivt annotasjonsgrensesnitt* nedenfor.

En annen viktig utfordring for annotasjon av skriftspråk er at talespråk er mer spontant enn skriftspråk. Når vi skriver, har vi tid til å tenke oss om, rette opp ulike typer performansefeil, og produsere et ferdig produkt som er i overensstemmelse med vår egen grammatiske kompetanse. Når vi snakker, blir alle slags performansefeil stående. Vi kan naturligvis korrigere oss selv underveis, og det gjør vi regelmessig. Et viktig spørsmål for annotasjon av et talespråkskorpus er da om alt skal annoteres og i så fall hvordan.

Det er ulike syn blant talespråksforskere når det gjelder hva som bør gis en syntaktisk representasjon i et talespråkskorpus. Johannessen og Jørgensen (2006) diskuterer ulike muligheter, med to ytterpunkter: enten kan man annotere alt som blir sagt som en del av den syntaktiske annotasjonen, eller man kan se bort fra det som betraktes som spesielle trekk ved talespråk, for eksempel interjeksjoner og korrigeringer.

Geoffrey Sampson bruker den første strategien. I artikkelen "Thoughts on two decades of drawing trees" (2003) gir han et eksempel på et syntaktisk tre for en ytring som inneholder en "speech repair", altså en ytring der taleren korrigerer seg selv underveis. Treet gjengis i figur 1.



Figur 1: Et analysetre fra CHRISTINE Corpus (Sampson 2003: 36).

Denne representasjonen har formen til et syntaktisk tre og spenner over hele ytringen, inkludert pausetegnet #, selv om ytringen ikke har en koherent syntaktisk form. Et syntaktisk tre kan også konverteres til syntaktiske regler. En slik konversjon for dette treet ville gi bl.a. de syntaktiske reglene i (6).

(6) Noun Phr → *a bonus*
Noun Phr → *any bonus* Rel Cl # *anything* Relative Clause
Relative Clause → *he gets* obj Prep Phr
Rel Cl → *he*

Reglene i (6) utgjør åpenbart ikke generaliseringer over formen som nomenfraser og relativsetninger kan ta i engelsk. Det vil ikke si at deler av

denne ytringen ikke har en klar syntaktisk form. Alt etter #-symbolet er en velformet setning: *anything he gets over that is a bonus*. Det virker mer hensiktsmessig å gi en syntaktisk analyse bare til den delen av ytringen som har en klar syntaktisk struktur.

En helt annen tilnærmingsmåte omtales av Wallis (2003) i forbindelse med ICE-GB, den britiske delen av *International Corpus of English*. Her ble selvkorrigeringer kommentert ut av teksten slik som i (7).

(7) so it's [no] effortless for Canonbury

Annotatoren kommenterer ut ordet *no* slik at parseren kan hoppe over dette ordet. Denne strategien virker mer motivert hvis man vil kunne fortolke trærne slik det er vanlig å gjøre det, nemlig som representasjoner som uttrykker syntaktiske generaliseringer. Det innebærer ikke at man ikke kan bruke en annen form for annotasjon for elementer som ikke passer inn i den.

Johannessen og Jørgensen (2006) foreslår en metode for automatisk parsing av talespråk. Metoden er ikke implementert, men skisseres i form av prinsipper for parsing. Det er mulig at prinsippene som er utformet vil kunne resultere i en brukbar algoritme som gir gode analyser. Men forfatterne diskuterer ikke problemet med flertydighet. I bygging av en trebank er det viktig at man ikke bare har en analyse, men at man har en korrekt analyse, og det må derfor være et element av menneskelig editering for å sikre kvaliteten.

LFG Parsebanker

LFG Parsebanker er et avansert verktøy for syntaktisk analyse av korpora. Den er under utvikling i forbindelse med TREPIL-prosjektet, et pilotprosjekt for å utvikle metoder og ressurser for bygging av en trebank for norsk (Rosén, De Smedt et al. 2005, Rosén, Meurer et al. 2005, Rosén et al. 2006). Grammatikken som brukes, er NorGram, en komputasjonell grammatikk for bokmål utviklet ved Universitetet i Bergen innenfor The Parallel Grammar Project (Dyvik 2003, Butt et al. 2002). NorGram ble først utviklet innenfor et eget prosjekt, men er senere videreutviklet i forbindelse med LOGON-prosjektet for maskinoversettelse fra norsk til engelsk (Oepen, Dyvik et al. 2004). Selv om LFG Parsebanker er utviklet spesifikt for bruk med NorGram, kan den også brukes for andre språk. Forutsetningen er at man har en LFG-grammatikk (Bresnan 2001) som er implementert på XLE-plattformen (Maxwell og Kaplan 1993).

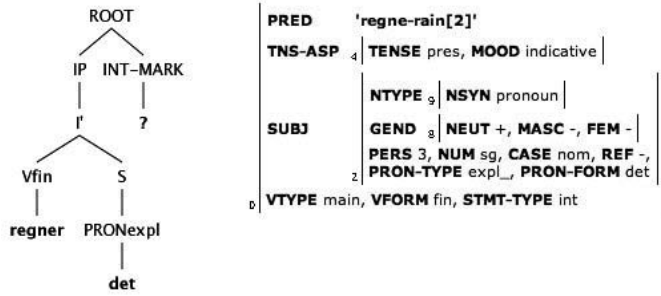
Første trinn i bygging av en trebank med LFG Parsebanker er at korpuset parses automatisk og analysene lagres i en database. Ved hjelp av en effektiv

disambigueringsprosedyre velger en menneskelig annotator den beste analyse for videre lagring. Annotatorvalgene tas også vare på. Etter revisjon av grammatikken og leksikon kan korpuset reparses, og de samme annotatorvalgene kan automatisk gjenbrukes for disambiguering.

I det følgende presenteres noen av egenskapene til LFG Parsebanker som gjør den spesielt velegnet til parsing av et talespråkskorpus.

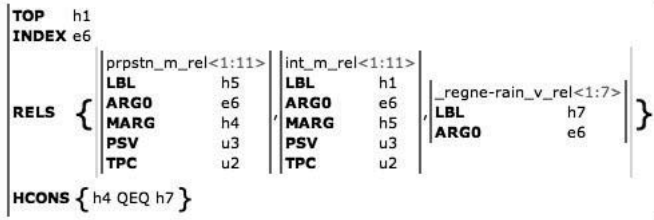
Rike grammatiske annotasjoner

Syntaktisk analyse i LFG er på to nivåer: c-struktur og f-struktur. En c-struktur er et frasestrukturtre. En f-struktur er en matrise med attributter og verdier. I figur 2 vises c- og f-strukturene som den norske grammatikken genererer for setningen *Regner det?* Dette eksempelet illustrerer hvor rik annotasjonen er; i tillegg til konstituentstrukturen til setningen, får man detaljert informasjon om grammatiske trekk og syntaktiske funksjoner.



Figur 2: c- og f-strukturer for setningen Regner det?

Den norske grammatikken leverer i tillegg en semantisk analyse i form av en mrs-struktur. Denne semantiske analysen er basert på Minimal Recursion Semantics (Copestake et al. 2005). En mrs-struktur er en flat struktur som består av et sett av elementære predikasjoner, der hver predikasjon har en relasjon (REL), en etikett (LBL), og et sett av argumentroller (ARG0, ARG1 osv.). Kvantorrekkevidde er underspesifisert, slik at det kan være én mrs-struktur for en setning som er flertydig med hensyn til kvantorrekkevidde. I figur 3 vises mrs-strukturen for setningen *Regner det?*



Figur 3: mrs-struktur for setningen Regner det?

Vi kommer senere tilbake til hvordan slike detaljerte strukturer kan være nyttige i forbindelse med søking etter bestemte konstruksjoner.

Effektiv disambiguering

Som nevnt ovenfor er effektiv disambiguering essensiell for automatisk syntaktisk analyse. Disambiguering gjøres i LFG Parsebanker ved hjelp av diskriminanter. Denne teknikken ble først foreslått av Carter (1997), og er senere brukt innenfor HPSG (Oepen, Flickinger et al. 2004). Metoden går ut på at disambiguering skjer på grunnlag av elementære lingvistiske egenskaper ved analysene. Hovedideen er at hvilken som helst lokal egenskap ved en analyse som ikke deles av alle analyser, kan anvendes til å skille mellom analyser.

I TREPIL-prosjektet har vi definert tre typer diskriminanter for LFG-grammatikker: morfologiske diskriminanter, c-struktur-diskriminanter og f-struktur-diskriminanter. Morfologiske diskriminanter er ord med taggene de får i den morfologiske analysen. Disse diskriminantene er spesielt velegnet for disambiguering av leksikalsk flertydighet.

(8) Hun ser mannen med bøkene.

Denne setningen får fire analyser. Det er to flertydigheter: én leksikalsk og én syntaktisk. Den leksikalske flertydigheten går ut på at *bøkene* kan være bestemt flertall av enten *bok* eller *bøk*. Denne flertydigheten resulterer i flere diskriminanter, bl.a. de morfologiske diskriminantene i tabell 1. Gjennom å velge én av disse diskriminantene kan annotatoren løse den leksikalske flertydigheten.

bøk+Pl+Noun+Masc+Def
bok+Pl+Noun+Masc+Def

Tabell 1: Morfologiske diskriminanter for (8).

Den andre flertydigheten, som angår tilknytningen av PP-en, kan løses gjennom å velge enten en c-struktur- eller en f-struktur-diskriminant. Denne enkle setningen har fire analyser, og kan altså disambigueres gjennom to enkle valg.

Mange setninger har langt flere analyser. For eksempel får setningen i (9) fra NoTa-korpuset hele 54 analyser av NorGram. Med LFG Parsebanker kan den likevel disambigueres med bare fire enkle valg.

(9) F11 gråt og hulket og mor gråt og det var helt kaos

Dette tilsynelatende enkle eksempelet (der et personnavn er blitt anonymisert i transkripsjonen gjennom erstatning med *F11*) viser at disambiguering er et spørsmål som ikke kan ignoreres og som bør adresseres helt fra begynnelsen av i ethvert system for syntaktisk annotasjon.

Interaktivt annotasjonsgrensesnitt

Som et effektivt redskap for semiautomatisk annotasjon av et korpus kan LFG Parsebanker like godt brukes for annotasjon av talespråk som skriftspråk. Den viktigste forskjellen er at det for talespråk er mye mindre klart hva som bør være input til parseren. Jørgensen (2007) har arbeidet med maskinlæringsteknikker for å automatisk finne frem til makrosyntagmer i NoTa-materialet, som han mener bør utgjøre de syntaktiske enhetene som er gjenstand for syntaktisk analyse. Men uansett hvor god en slik prosess er til å finne frem til de syntaktiske enhetene i et talespråkskorpus, vil det helt sikkert alltid være behov for manuelle justeringer av grensene mellom disse enhetene.

For å takle dette problemet har vi implementert et nytt grensesnitt i LFG Parsebanker der det er mulig for annotatoren å manipulere grensene mellom ytringer. Ytringen i (10) fra NoTa-korpuset består av to setninger.

(10) jeg husker jeg fikk en sjampo for e farget hår jeg hadde ikke farget hår engang

Teknikkene som Jørgensen beskriver, kan sannsynligvis ikke oppdage at denne ytringen inneholder to setninger. Når en slik streng parses av grammatikken, vil den enten ikke få noen analyse, eller den vil få en fragmentanalyse. Uansett vil annotatoren i disambigueringsprosessen oppdage at den ønskede analysen ikke er tilstede, og også grunnen til at analysen mangler, nemlig at ytringen består av to setninger. Annotatoren kan da resegmentere ytringen gjennom å klikke på mellomrommet som deler setningene fra hverandre, slik at input til parseren nå blir setningene i (11) og (12).

(11) jeg husker jeg fikk en sjampo for e farget hår
(12) jeg hadde ikke farget hår engang

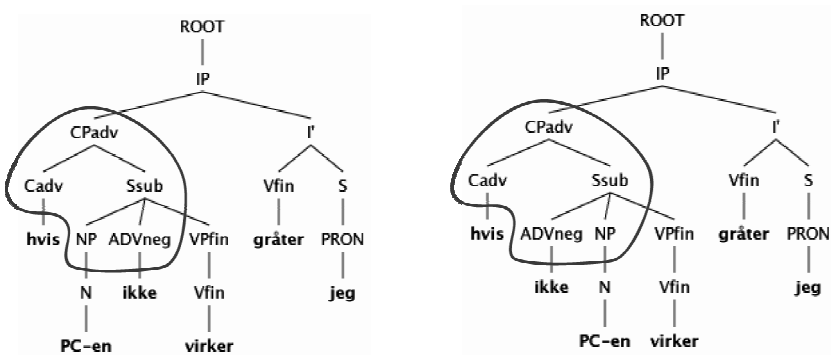
Setningen i (11) inneholder nok et problem for parseren: det som er transkribert som et vanlig grafisk ord *e*, og som brukes for "nøling – uansett lengde på een" (Hagen 2005: 36), har ingen opplagt plass i den syntaktiske analysen. Hvis et "ord" kan forekomme hvor som helst i en ytring, uten at det er mulig å begrense dets syntaktiske distribusjon, er det ikke rimelig at det får plass i den syntaktiske analysen. Annotatoren kan markere at parseren skal ignorere dette ordet gjennom å sette det i krøllparenteser før setningen blir sendt til reparsing, slik som i (13). Flere eksempler fra NoTa-korpuset på ytringer der denne teknikken kan brukes, vises i (14) og (15).

(13) jeg husker jeg fikk en sjampo for {e} farget hår
(14) det {det får d-} kan du få uansett
(15) det var {var var} både cowboyer og indianere der

Sofistikert søkeverktøy

For en sluttbruker av en trebank er det ikke bare viktig at den grammatiske annotasjonen er rik, men også at det går an å søke på bestemte fenomener og konstruksjoner. LFG Parsebanker har implementert en utvidet versjon av TIGERSearch (Lezius 2002). Dette er et søkeverktøy som opprinnelig ble utviklet for avanserte søk blant trestrukturer; i TREPIL-prosjektet er det utvidet slik at søk også kan foretas blant f-strukturer.

La oss si at en forsker er interessert i plassering av setningsadverbialer i leddsetninger, altså setninger av den typen vi så i (2) og (3) ovenfor. Disse setningene får treanalysene som vi ser i figur 4.



Figur 4: c-strukturer for setningene i (2) og (3).

Hvis man bruker et korpus uten syntaktisk annotasjon, er man nødt til å søke etter spesielle ordsekvenser (eller sekvenser av morfosyntaktiske etiketter) for å finne slike konstruksjoner. Med LFG Parsebanker kan vi abstrahere bort fra alle ord og søke på den delen av treet som skiller disse analysene fra hverandre, nemlig den delen som er innringet i figur 4. Dette kan gjøres gjennom å utforme krav til dominans- og presedensrelasjoner slik som vist i (16) og (17) for henholdsvis (2) og (3).

(16) #1:[cat="CPadv"] > [cat="Cadv"] &
#1:[cat="CPadv"] > [cat="Ssub"] &
#2:[cat="Ssub"] > #3:[cat="NP"] &
#2:[cat="Ssub"] > #4:[cat="ADVneg"] &
#4:[cat="NP"] .* #3:[cat="ADVneg"]

(17) #1:[cat="CPadv"] > [cat="Cadv"] &
#1:[cat="CPadv"] > [cat="Ssub"] &
#2:[cat="Ssub"] > #3:[cat="NP"] &
#2:[cat="Ssub"] > #4:[cat="ADVneg"] &
#4:[cat="ADVneg"] .* #3:[cat="NP"]

Det regulære uttrykket *[cat="CPadv"] > [cat="Cadv"]* betyr at en node med etiketten *Cadv* må være direkte dominert i treet av en node med etiketten *CPadv*. Det regulære uttrykket *[cat="NP"] .* [cat="ADVneg"]* betyr at en node med etiketten *NP* må komme før (må være til venstre for) en node med etiketten *Cadv*. Indeksene #1, #2 osv. sikrer at føringene gjelder én og samme node. Disse føringene sammen vil resultere i at man får selektert alle setninger i trebanken som har disse mønstrene.

Lignende søk kan gjøres for ulike kombinasjoner av leksikalske og grammatiske fenomener, og man kan søke på trekk som forekommer i f-strukturene.

Konklusjon

Syntaktisk annotasjon av korpora er krevende, men åpner for interessante lingvistiske undersøkelser. Annotasjon av talespråk er enda mer krevende enn annotasjon av skriftspråk. Jeg har diskutert noen krav og strategier for effektiv annotasjon. Automatisk parsing er en god metode for å sikre konsistent annotasjon av høy kvalitet hvis parsingen kobles til effektiv menneskelig disambiguering. LFG Parsebanker kan, brukt sammen med NorGram, allerede nå levere detaljerte syntaktiske og semantiske

annotasjoner for mye av materialet i NoTa-korpuset. Siden automatisk parsing forutsetter at input er enheter med syntaktisk struktur, vil det være en fordel med preprosessering av NoTa-transkripsjonen i sannsynlige syntaktiske enheter. Men det er viktig at den endelige avgjørelsen av hva som skal parses, tas av et menneske. Derfor har vi også implementert et grensesnitt der annotatoren kan resegmentere ytringer i forbindelse med disambiguering.

Referanser

Abeillé, Anne (red.). 2003: *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers.

Bresnan, Joan. 2001: *Lexical-Functional Syntax*. Malden, MA: Blackwell.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi og Christian Rohrer. 2002: The Parallel Grammar project. *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*. Taipei, Taiwan, 1–7.

Carter, David. 1997: The TreeBanker: A tool for supervised training of parsed corpora. *Proceedings of the Fourteenth National Conference on Artificial Intelligence*. Providence, Rhode Island, 598–603.

Copestake, Ann, Dan Flickinger, Carl Pollard og Ivan A. Sag. 2005: Minimal Recursion Semantics. An introduction. *Research on Language and Computation* 3, 281–332.

Dyvik, Helge. 2003: ParGram: Developing Parallel Grammars. Feature article, *Elsnews* 12.2, 12–14.

Hagen, Kristin. 2005: *Transkripsjonsveiledning for NoTa-Oslo.* http://www.tekstlab.uio.no/nota/oslo/index.html

Johannessen, Janne Bondi og Fredrik Jørgensen. 2006: Annotating and parsing spoken language. Henrichsen, Peter Juel og Peter Rossen Skadhauge (red.): *Treebanking for Discourse and Speech: Proceedings of the NODALIDA 2005 Special Session on Treebanks for Spoken Language and Discourse.* Copenhagen: Samfundslitteratur, 83–104.

Jørgensen, Fredrik. 2007: Clause Boundary Detection in Transcribed Spoken Language. Nivre, Joakim, Heiki-Jaan Kaalep, Kadri Muischnek og Mare Koit (red.): *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*. Tartu: University of Tartu, 235–239.

Lezius, Wolfgang. 2002: TIGERSearch – ein Suchwerkzeug für Baumbanken. Busemann, Stephan (red.): *Proceedings der 6. Konferenz zur Ver-arbeitung natürlicher Sprache (KONVENS 2002)*. Saarbrücken.

Maxwell, John og Ronald M. Kaplan. 1993: The interface between phrasal and functional constraints. *Computational Linguistics* 19, 571–589.

Nivre, Joakim, Koenraad De Smedt og Martin Volk. 2005: Treebanking in Northern Europe: A White Paper. Henrik Holmboe (red.): *Nordisk Sprogteknologi 2004*. Museum Tusculanums Forlag, 97–112.

Oepen, Stephan, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård og Victoria Rosén. 2004: Som å kapp-ete med trollet? Towards MRS-based Norwegian–English Machine Translation. *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Baltimore, MD.

Oepen, Stephan, Dan Flickinger, Kristina Toutanova og Christopher D. Manning. 2004: LinGO Redwoods, a rich and dynamic treebank for HPSG. *Research on Language and Computation* 2, 575–596.

Rosén, Victoria, Koenraad De Smedt, Helge Dyvik og Paul Meurer. 2005: TREPIL: Developing methods and tools for multilevel treebank construction. Civit, Montserrat, Sandra Kübler og Ma. Antònia Martí (red.): *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005),* 161–172.

Rosén, Victoria, Paul Meurer og Koenraad De Smedt. 2005: Constructing a parsed corpus with a large LFG grammar. *Proceedings of LFG'05.* CSLI Publications, 371–387.

Rosén, Victoria, Koenraad De Smedt og Paul Meurer. 2006: Towards a toolkit linking treebanking to grammar development. *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, 55–66.

Sampson, Geoffrey. 2003: Thoughts on two decades of drawing trees. Abeillé, Anne (red.): *Treebanks: Building and Using Parsed Corpora.* Dordrecht: Kluwer Academic Publishers, 23–41.

TREPIL – Treebank Pilot Project. Aksis, Unifob, Universitetet i Bergen. http://gandalf.aksis.uib.no/trepil/

Wallis, Sean. 2003: Completing parsed corpora. From correction to evolution. Abeillé, Anne (red.): *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers, 61–71.