Referat frå møte om transkripsjonsstandardar i ScanDiaSyn

Dato: 10.-11. februar 2006
Stad: Lunds universitet, SOL-centrum, Lingvisthuset
Deltakarar: Victoria Absalonsen (Færøyane)
Philipp Conzett (Tromsø)
Janne Bondi Johannessen (Oslo)
Lars-Olof Delsing (Lund)
Karen Margrethe Pedersen (København)
Christer Platzack (Lund)
Henrik Rosenkvist (Lund)
Ásta Svavarsdóttir (Island)
Øystein Alexander Vangsnes (Tromsø)
Sten Vikner (Århus)
Jan-Ola Östman (Helsingfors)

Program:

Fredag 10.2.:
Ulike presentasjonar kring transkripsjon:

1. Øystein Vangsnes: Innleiing
2. Janne Bondi Johannessen: Transkripsjon i et talespråkskorpus
3. Sven Strömquist: Webb-baserat system för humanistisk forskningsdata
4. Janne Bondi Johannessen: "Offline"-demo av programmet Transcriber
5. Henrik Rosenkvist: Transkriptioner i CHAT/CLAN (erfarenheter, problem)

Laurdag 11.2.:
Rapportar og diskusjon:

6. Situasjonsrapportar frå dei ulike gruppene
7. Diskusjon

1. Øystein Vangsnes: Innleiing

Nokre av gruppene er godt i gang med transkribering av dialektmateriale. Difor er det på tide å verta samde om felles **transkripsjonsstandardar/-reglar**.

Meir generelt viste Øystein til ScanDiaSyn-heimesida og etterlyste meir aktivitet av og mellom prosjektdeltakarane. Han framheva særskilt to element på heimesida: bibliografien og diskusjonsforumet. Det er ynskjeleg at ein tek i bruk internett til å samarbeida i nettverket. Bloggen opplyser om kva som er nytt på heimesida og i prosjektet og er med det ei effektiv hjelp til å halda seg oppdatert.

2. Janne Bondi Johannessen: Transkripsjon i et talespråkskorpus

Føredraget var delt inn i fem hovuddelar:

- Presentasjon av NoTa-prosjektet
- Ortografisk transkripsjon
- Prinsipp for transkripsjon
- Utfordringar
- Framtida

Av særskild interesse for ScanDiaSyn er måten NoTa-materialet er transkribert på. Tekstlaboratoriet har valt å nytta **ortografisk transkripsjon** med visse tilpassingar til talemålet. Denne typen transkripsjon svarar stort sett til det ein i SweDiaSyn kallar "glossning". Ortografisk transkripsjon er ikkje ei ljodtru attgjeving av det informantane faktisk seier, men ei slags omsetjing til tilsvarande skriftspråkleg normerte ord. Som referanseverk har ein i NoTa nytta Bokmålsordboka. Hovudprinsippa for transkripsjonen i NoTa er:

- Nytta ordboksformer
- Aldri endra ordstillinga for å tekkjast standardspråket
- Nytta ordboksformer også der dei er nytta "gale" i høve til skriftspråksnormen

Stundom byr dei ortografiske transkripsjonsreglane på **utfordringar**. Dette er særleg tilfelle når ortografisk tanskripsjon innfører sjølvmotseiingar i grammatikken til informantane, eller når informantane nyttar ord som ikkje finst i ordboka. Her har ein valt pragmatiske løysingar. Hovudmålet har alltid vore å vera så trufast som mogleg mot dei faktisk observerte ytringane. (Fullstendig transkripsjonsrettleiing kan lastast ned frå http://www.hf.uio.no/tekstlab/prosjekter/NoTa/NoTa-transkripsjonsveil7.pdf)

3. Sven Strömquist: Webb-baserat system för humanistisk forskningsdata

Strömquist har jobba ein del med **samling og tilgjengeleggjering av forskingsdata** innanfor humaniora. Her er det to europeiske prosjekt som har stått sentralt: ECHO ("European Culture Heritage Online": http://echo.mpiwg-berlin.mpg.de/home) og oppfølgjingsprosjektet DAM-LR ("Distributed Access Management for Language Resources": http://www.mpi.nl/DAM-LR/). Målet med desse er at eksisterande forskingsmateriale vert beskrive med eit metadata-sett som gjer det meir søkbart. Forskarar kan også få rettleiing og tekniske hjelpemiddel til å oppretta korpus.

4. Janne Bondi Johannessen: "Offline"-demo av programmet Transcriber

Janne presenterte korleis talematerialet i NoTa er transkribert med hjelp av programmet Transcriber. Her er ortografisk transkripsjon (sjå 2. ovanfor) knytt til ljodfiler.

5. Henrik Rosenkvist: Transkriptioner i CHAT/CLAN (erfarenheter, problem)

Henrik fortalde om røynsler og problem frå SweDiaSyn-prosjektet:
- Berekning av arbeidsinnsatsen til transkribentar: Her er det fleire faktorar som spelar inn.
    Kor mykje røynsle har transkribenten?
    Kor vanskeleg er dialekten (avstand frå rikssvensk)
    Type av transkripsjon
    Kvalitet på innspelinga

- Kvaliteten på innspelingane i SweDia: Ofte var ikkje intervjuarane syntaksorienterte, eller dei hadde ikkje nok øving. I ein del av innspelingane oppmodar dessutan feltarbeidaren informanten til ikkje å snakka dialekt sidan han/ho har problem med å forstå dialekten. I tillegg kjem tekniske problem.

- Tilgang på kvalifiserte transkribentar: Det er eit krav om kjennskap til dei einskilde dialektane, CHAT-kompetanse og eit høveleg utdanningsnivå.

- Standardisering av transkripsjonar: Di meir detaljerte reglar, di betre samsvar mellom transkribentane, men også lengre tid per transkripsjon.

- Tidsbruk: Mest tid er nytta på transkribering (assistentar), møte, gjennomgangar, kontroll av CHAT-filer, tagging, oppretting av database, oppfølgjing (spørreskjemaundersøkingar, utfyllande innspelingar), forsking osb.

- Bemanning: Til no har det berre vore ein tilsett i 55%-stilling. Ideelt sett skulle ein ynskja seg ei forskargruppe.

6. Situasjonsrapportar frå dei ulike gruppene

- **Sverige** ved Lars-Olof: Innanfor SweDiaSyn-prosjektet er det gjort mest ortografisk transkripsjon ("glossning"), men i somme tilfelle finst det fonologisk transkripsjon (t.d. frå Gotland). 15-20 målepunkt er tekne med i SweDiaSyn, og frå kvart målepunkt ligg det føre om lag 1,5 timar med innspeling.

Øystein vil gjerne ha eit oversyn over dei målepunkta som er transkriberte, til å leggja ut på ScanDiaSyn-heimesida.

- **Island** ved Ásta: Gruppa held på med å transkribera eksisterande talespråksmateriale. I tillegg har dei utarbeidd eit spørjeskjema med utvalde problem til mindre undersøkingar, t.d. "den nye passiven".

- **Noreg** (Tromsø) ved Øystein: Det er svært uheldig at den norske gruppa ikkje har fått pengar frå

Forskingsrådet. Han jobbar elles med å få lagt ut rapportane frå stormøtet på Leikanger. Ei lesegruppe i verbflytting startar opp 16. februar.

- Noreg (**Tekstlaboratoriet/**Oslo) ved Janne: Tekstlaboratoriet har brukt ein del ressursar på ScanLex, og har saman med systerprosjektet Netordbogen kome eit godt stykke med norsk, svensk og dansk. Det er mykje arbeid som står att. Tekstlaboratoriet har fått litt meir moralsk støtte frå instituttet, og dei vonar at det i framtida skal vera lettare å søkja om pengar frå universitet sentralt. Særleg NoTa har vorte lagt meir merke til. Det er å håpa at denne positive merksemda òg kan koma arbeidet med ScanDiaSyn-databasen til gode.

- **Finland** ved Jan-Ola: Gruppa jobbar no med å få med finske dialektar også, ikkje minst sidan Sjef Barbiers ynskjer å inkludera dei i Edisyn. Det skal sendast inn ein felles søknad for både finske og svenske dialektar i Finland til Finlands Akademi i september. Elles er det arbeid i gang med å utvikla eit pedagogisk verkty om dialektar.

- **Danmark** ved Sten Vikner: Gruppa har søkt om pengar til DanDiaSyn, men fekk avslag. Neste år skal København (Karen Margrethe Pedersen og Inge Lise Pedersen) stå ansvarleg for søknaden. Det må også diskuterast om DanDiaSyn kan slåast i hop og/eller samarbeida med Frans Gregersens Sprogforandringscenter. I Danmark finst det allereie eit stort tekstkorpus med dansk dialektmateriale, korpus CorDiale, som omfattar 170 tekstar (på i alt 1 mill. løpeord). Etter at dette korpuset vart etablert, har det vorte transkribert ytterlegare 32 tekstar (på i alt 300 000 løpeord) som ikkje vert lagde inn i CorDiale av di den it-ansvarlege ikkje har tid. Desse tekstane vil kunna inngå i ScanDiaSyns database.

7. Diskusjon omkring transkripsjon

Først vart det gjeve nokre allmenne opplysningar om aktivitetar i ScanDiaSyn. Jan-Ola informerte om neste stormøte som skal arrangerast i Solf i Österbotten, 8.-11. juni 2006. Det vart bestemt at hovudemne for møtet skal vera “Dialects and syntactic theory/theories” i tråd med den opphavlege prosjektskissa. Ulike forskarar skal gje eit oversyn over ulike teoretiske tilnærmingar innanfor syntaks. Desse personane kan vera aktuelle:

- Fried/Östman: konstruksjonsgrammatikk
- Vikner: OT-syntaks
- Platzack: minimalistisk syntaks
- Henrik Jørgensen: Diderichsen-tradisjonen

Det skal også vera ein workshop for yngre forskarar. Dersom det vert gruppearbeid, bør det setjast av meir tid til diskusjon enn det som var tilfelle på Leikanger.

I forkant av stormøtet (3.-7. juni) skal NORMS ha ein dialektdugnad om nordøsterbotniske dialektar. Talet på deltakarar vil liggja på om lag 10.

NORMS skal også lysa ut inntil 5 postdoktorstillingar i nær framtid.

Transkripsjon:

Med utgangspunkt i diskusjonen som har pågått på debattforumet på heimesida til ScanDiaSyn sidan i haust, skisserte Øystein opp dei mest nærliggjande typane av transkripsjon:

(a) ljodnær transkripsjon (SweDiaSyn: "transkription")
    + meir informasjon
    - saktare = dyrare
(b) ortografisk transkripsjon (SweDiaSyn: "glossning")
    + kan taggast automatisk
    + lettare søkbar
    + raskare = billegare
    - mindre informasjon
(c) kommentert/annotert versjon

For Øystein veg ikkje argumentet at ortografisk transkripsjon er billegare så tungt at det utelukkar ljodnær transkripsjon. Materialet skal også kunna brukast og vidareutviklast om nokre år, også av andre enn syntaktikarar. Han vil ikkje ha eit kompromiss/ei blanding, for då vert det mest tilfeldig kva fenomen ein ynskjer å inkludera. Ortografisk transkripsjon kan vera greitt for Oslo-mål (jf. NoTa), men er i praksis vanskeleg å nytta for andre dialektar. Her treng ein i tillegg ljodnær transkripsjon.

Karen Margrethe ynskte ein lettare tillempa ortografisk versjon (b), der einskilde sentrale avvik frå rikksspråket er tillatne, t.d. merking av framforstilt artikkel *æ*, som taggaren kunne lærast opp til. Ho argumenterte for denne løysinga med at det ikkje ville vera mogleg å søkje automatisk på eit trekk som ikkje er merkt av i rikksspråksversjonen. Til dømes kan ein ikkje søkja på framforstilt artikkel dersom han berre er merkt av i kommentarversjonen (c). Karen Margrethe fekk her støtte frå Lars-Olof, som føreslo at kvart land kunne fastleggja eit knippe av utvalde avvik. Øystein var oppteken av å halda dei ulike transkripsjonsnivåa åtskilde, dvs. at det skal finnast ein rikksspråksversjon som er så "rein" som råd. Han la også til at dette til ein viss grad var eit teknisk og praktisk spørsmål. Karen Margrethe meinte at ein kunne ha nøgd seg med ein rein rikksspråksversjon i Danmark dersom ein her (som i Noreg) hadde hatt fleire skriftspråk: eit vestjysk og eit bornholmsk forutan det eksisterande sjællandskbaserte. Ho reiste også spørsmålet om ein ville kunna stola på ljodtranskripsjonen. Til dette svara Øystein at transkripsjonen skulle gjerast av transkribentar som både var vel kjende med dialekten og hadde ein viss fagleg bakgrunn.

Deltakarane vart samde om at ein i ScanDiaSyn i alle fall treng (b). For nokre dialektar vil (a) vera best eigna. Om det finst fenomen som ikkje går fram av (b), vil ein kunna merkja dei av i (c). Øystein reiste igjen spørsmål om kva for kriterium spesielle fenomen i (c) skal veljast ut etter. Ifølgje Lars-Olof er det stort sett semje om kva for fenomen som er mest aktuelle. Han gjer framlegg om at kvar gruppe bestemmer seg for eit knippe fenomen som skal undersøkjast.

I nokre tilfelle vil det vera mest nærliggjande å starta med den ljodnære transkripsjonen (a) og fyrst deretter leggja til (b). Jan-Ola sa at dei i Finland kjem til å arbeida på den måten.
Diskusjonen munna ut i ei semje om å halda dei tre transkripsjonsnivåa nemnde ovanfor skilde så vidt det er føremålstenleg og teknisk mogleg. Janne fekk i oppgåve å undersøkja visse tekniske spørsmål knytte til eksisterande transkripsjonsprogram (m.a. om det er mogleg å modifisera Transcriber slik at ein får fleire nivå). Janne og Henrik fekk i oppgåve å utarbeida eit grunnlagsdokument (policy-dokument), og med utgangspunkt i dette kan meir språkspesifikke retningsliner utformast av dei einskilde gruppene. Kvar gruppe står fritt til å velja kva for transkripsjonsprogram (Transcriber, CHAT/CLAN, ELAN, Praat osb.) som vert nytta, men det er viktig å få klarlagt kva for tekniske begrensingar som eventuelt ligg i dei. Det mest ideelle ville naturleg nok vera at ein fann fram til eitt program som vart nytta av alle.

Tromsø, 16. februar 2006

Philipp Conzett