Om NoTa-korpuset og artiklene i denne boka

Janne Bondi Johannessen og Kristin Hagen

Innledning

I november 2006 arrangerte vi et seminar om Oslo-språk og talemålsforskning[1] Foredragene på seminaret var interessante og allsidige, og viste nyskapende forskning på en rekke områder. Alle foredragene hadde én ting til felles: de hadde brukt *NoTa-korpuset* (Norsk Talespråkskorpus – Oslodelen), et nylig ferdigstilt, avansert forskningsmateriale med talespråk fra Oslo. Den forskningen som ble presentert på seminaret, synes vi var så spennende at den burde komme allmennheten til gode. Dermed ble ideen om denne boka til.

Alle artikkelforfatterne har blitt bedt om å popularisere stoffet. Mange av artiklene framstår derfor som lettfattelige og lesbare for et publikum av ikke-eksperter med interesse for språk. Avhengig av hvilken bakgrunn leseren har, er det ulike ting som vil appellere. Har man teknisk bakgrunn, kan artiklene om språkteknologi være mer lettfattelige enn de om ordstilling, for eksempel.

Artiklene i boka er skrevet av foredragsholderne fra seminaret (de fleste valgte å være representert her), samt noen av ordstyrerne. Den direkte bakgrunnen for seminaret var ferdigstillelsen av NoTa-korpuset, som er beskrevet nedenfor. Vi ønsket at så mange ulike områder av språkforskningen som mulig skulle ta i bruk det flotte forskningsmaterialet som NoTa-korpuset representerer. Derfor inviterte vi et knippe av språkforskere fra hele Skandinavia, som tilfredsstilte en del forutsetninger: De måtte være kjent som toppforskere innenfor sine felt, ha forsket på minst ett skandinavisk språk, være vant til å bruke empiriske data i form av korpus og ha evne til å kaste seg over et nytt materiale. Til sammen representerte de mange ulike forskningsfelt.

I tillegg inviterte vi noen i kategorien "yngre forskere", nemlig postdok-stipendiater, doktorgradsstipendiater og masterstudenter som hadde vært med i utviklingsarbeidet rundt korpuset. Vi synes artiklene av de yngre forskerne viser høy kvalitet og navngir dem her: Gunnar Hrafn Hrafnbjargarson,

[1] Seminaret het *Oslomålet - et seminar med forskning fra NoTa-korpuset* og fant sted på Universitetet i Oslo 23. og 24. november 2006.

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

Fredrik Jørgensen, Anders Nøklestad, Inger Margrethe Hvenekilde Seim og Åshild Søfteland.

Det er flere som må takkes. Først takker vi Norges forskningsråd, som gjennom en spesialbevilgning for utstyr til humanistisk forskning gjorde det mulig å bygge opp NoTa-korpuset ved Tekstlaboratoriet, UiO. En takk også til Humanistisk fakultet, som anbefalte søknaden overfor Forskningsrådet, og Institutt for lingvistiske og nordiske studier, som bevilget penger til seminaret og til trykkestøtte av boka.

Dernest takker vi artikkelforfatterne, samt fagkonsulentene for de enkelte artiklene. Fagkonsulenter (i dette tilfellet utvalgte foredragsholdere og ordstyrere fra seminaret) gjør en viktig jobb, men frister ofte en anonym rolle, siden de alltid holdes skjult overfor artikkelforfatterne. Derfor navngir vi dem som har medvirket i tillegg til redaktørene: Kristin Melum Eide, Jan Terje Faarlund, Gunnar Hrafn Hrafnbjargarson, Marit Julien, Gjert Kristoffersen, Svein Lie, Helge Lødrup, Victoria Rosén, Hanne Gram Simonsen, Jan Svennevig og Tor A. Åfarli. Arne Martinus Lindstad har lest korrektur og foretatt nødvendige formateringsendringer.

Vi vil også nevne medarbeiderne på NoTa-Oslo-prosjektet: prosjektleder Janne Bondi Johannessen og instituttleder Hanne Gram Simonsen, daglig leder Kristin Hagen, vitenskapelige assistenter Anne Marit Bødal og Fredrik Jørgensen, programmerere Joel J. Priestley og Anders Nøklestad, og transkribører Hilde Cathrine Haug, Ingunn Indrebø Ims, Signe Laake, Inger Margrethe Hvenekilde Seim og Åshild Søfteland.

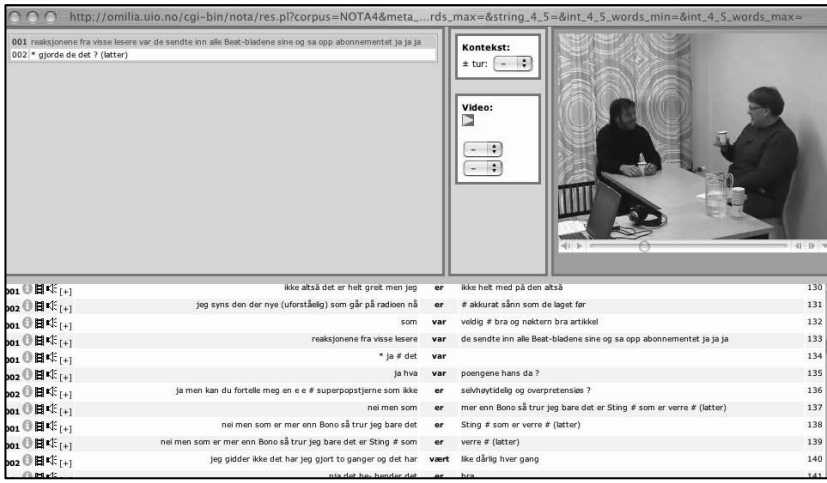
Til slutt vil vi takke NoTa-Oslos informanter som sporty stilte opp til intervju og samtaler. Uten informantene hadde det ikke blitt noe talespråkskorpus!

Litt om NoTa

Norsk Talespråkskorpus – Oslodelen, eller NoTa-Oslo, består av talespråk fra 166 personer i Oslo-området, og omfatter om lag 900 000 ord. Informantene er jevnt fordelt på variablene kjønn, alder, utdannelsesbakgrunn og geografisk tilhørighet. Hver person (informant) forekommer i opptak med lyd og video, og tar del i to opptakssituasjoner: I en kort sekvens blir informanten intervjuet av en prosjektassistent, og i en lengre sekvens på en halvtimes tid deltar to informanter sammen i en samtale, hvor de kan snakke fritt om en rekke temaer. La oss likevel nevne at de også har fått presentert en

liste over emner de ikke får snakke om, på grunn av hensynet til personvern[2]. De får f.eks. ikke snakke om sensitive temaer som sykdom, fengselsopphold eller politikk. Noen av informantene kjenner hverandre, eller er i familie, andre er ukjente for hverandre. Innsamlingen av talemålsmateriale i NoTa-Oslo er inspirert av flere liknende prosjekter, men kanskje særlig av *Corpus Gesproken Nederlands* (CGN) og *Talemålsundersøkelsen i Oslo (TAUS)*[3].

Vi har transkribert alt lydmaterialet til skrift, såkalt ortografisk transkripsjon, og alt er lagt inn i et korpus med database som kan brukes via en webside. Her er det muligheter for å søke i materialet ut fra en rekke kriterier som kan brukes alene eller i kombinasjon. Man kan søke etter bestemte ord, deler av ord, sekvenser av ord, samt spesifisering av grammatiske kategorier som ordklasse eller bøyingstrekk. Dette kan man så kombinere med utvalg av informanter basert på kriterier som alder, bosted, utdannelse, kjønn eller annet. Slik er det duket for mange typer studier, slik også artiklene i denne boka viser.



Figur 1: Resultatbilde av et søk i korpuset. Alle forekomster av søkeordet (her: verbet være *i alle former) er plassert under hverandre med konteksten rundt. Et videoopptak av et valgt treff avspilles øverst.*

[2] Prosjektet er forskriftsmessig meldt til Personvernombudet for forskning ved NSD, Norsk samfunnsvitenskapelig datatjeneste AS, og har forholdt seg til retningslinjer fra NSD angående personvern.

[3] CGN: se http://lands.let.kun.nl/cgn/ehome.htm,
TAUS: se http://www.tekstlab.uio.no/nota/taus/index.htm

Materialet skal brukes til forsknings- og utviklingsformål. Det er nesten ingen grenser for bruksområder. Noen forskere er mest interessert i hvordan språket anvendes av ulike befolkningsgrupper. Andre er mer interessert i å beskrive regelmessigheter for enkelte språklige forhold, eller finne ut hva som er bruksbetingelsene eller hva som er de grammatiske omgivelsene for enkelte uttrykk. Atter andre undersøker semantiske forhold, dvs. uttrykks betydning, eller lydlige forhold. Videre er det forskere som er mer interessert i abstrakte regelmessigheter som kan duke for ikke bare generaliseringer, men til og med prediksjoner om ulike konstruksjoner.

De ulike forskerne kan i tillegg være interessert i enkeltord, bøyingsformer, sammensatte ord eller sekvenser av flere ord i løse og fastere fraser. Mange foretrekker å se språklige fenomener i et komparativt, kontrastivt eller typologisk perspektiv. Språkteknologer trenger materiale for å analysere eller utvikle verktøy som til sist kan brukes i ulike produkter. NoTa-materialet er nyttig for alle slike og mange flere formål.

NoTa-Oslo-prosjektet har sin egen hjemmeside der en kan lese mer om informantene, transkripsjonene, web-grensesnittet og prosjektet. På denne siden kan en også klikke seg videre til selve søkesiden. (Se bak i artikkelen for referanse.)

Kort om de enkelte artiklene i boka

Artiklene i boka gir et rikt bilde av talespråksforskning og Oslo-språksforskning. Nedenfor gir vi en liten oversikt over dem. Vi har valgt å plassere artiklene i enkelte hovedkategorier, men – som alle som har prøvd å kategorisere noe som helst, vet – er det også her slik at enkelte artikler like gjerne kunne passet i én kategori som i en annen.

Ord: Leksikografi, sosiolingvistikk, grammatikk, fonetikk

Ruth Vatvedt Fjeld ser i artikkelen *Talespråksforskningens betydning for leksikografien* på flere sider ved talespråket som er ignorert i eksisterende ordbøker, blant annet på grunn av manglende materiale. Fjeld tar opp typiske talespråksord som mangler i ordbøkene, både grammatiske og mer typiske slangord, samt tabuaktige ord relatert til "å gjøre ens fornødne". Hun viser også hvordan betydningsendringer kommer tydelig fram i et talespråkskorpus (f.eks. at *dass* oftest brukes i overført betydning), mens ordbøkene hittil ikke har viet dette oppmerksomhet. Til slutt peker hun på at enkle frekvensopptellinger viser klart motsatte mønstre mellom skriftspråk og talespråk når det gjelder bøyingsformer.

Toril Opsahl, Unn Røyneland og Bente Ailin Svendsen undersøker i

sin artikkel "*Syns du jallanorsk er lættis, eller?" – om taggen [lang=X] i NoTa-Oslo-korpuset* de ordene i korpuset som er unormerte i forhold til ordbøkene. De ser spesielt på hvilke befolkningsgrupper som bruker flest unormerte ord, og finner at det er klare forskjeller mellom de ulike gruppene. Unge gutter fra østre deler av Oslo er mest kreative i dette nyordsperspektivet.

Kjell Ivar Vannebo har i sin artikkel *NoTa-informantene og tellemåten* sett på hvordan gammel og ny tellemåte (eksempelvis *treogtredve* mot *trettitre*) fordeler seg i NoTa, nå som det er over femti år siden den nye tellemåten ble innført. Vannebo påpeker at det er eneste gang en normering for talemål har funnet sted for norsk, og konkluderer med at reformen må sies å ha vært vellykket, selv om andre har hevdet det motsatte!

Øystein Alexander Vangsnes tar opp bruken av spørreord alene og foran substantiver i artikkelen *Omkring adnominalt* åssen/hvordan *i Oslo-målet*. Han diskuterer hvilke betydninger som er mulige når folk sier for eksempel *åssen bil*, og han ser også på denne bruken fra en sosiolingvistisk synsvinkel. For eksempel viser han med klare tall at typen *åssen bil* er mindre prestisjefylt enn det tilsvarende *hvordan bil*.

Janne Bondi Johannessen diskuterer ordene *han* og *hun* i artikkelen *Psykologiske demonstrativer*. Brukt foran substantiver er disse sjeldne i skriftspråk, og sjeldne i eldre oslospråk. Det viser seg nå at disse demonstrativene har vært på frammarsj i Oslo de siste årene. Johannessen forklarer hvordan psykologisk avstand er en viktig bruksbetingelse for akkurat disse demonstrativene. Resultatet er at grammatikken er blitt mer kompleks enn før.

Svein Lie tar opp ordene *veldig* og *sånn* i artikkelen *Veldig sånn festejente*. Begge ordene har fått nye betydninger. Lie påpeker hvordan ordet *veldig* brukt med substantiv – som i artikkelens tittel – er mulig bare hvis man oppfatter et bestemt trekk eller en egenskap ved substantivet, som så kan graderes. Ordet *sånn* er et ekstremt vanlig talespråksord, og har mange viktige bruksbetingelser. En av dem er faktisk å signalisere høflighet.

Gjert Kristoffersen og Hanne Gram Simonsen ser på uttalen av konsonantgruppen *sl* i artikkelen *Osjlo! En undersøkelse av* sl-*sekvensen i NoTa-korpuset*. Kristoffersen og Simonsen viser at Oslos befolkning helt klart foretrekker den ene uttalen av denne konsonantgruppen. Dette gjelder for uttalen av *sl* i *Oslo*, men også for andre ord med *sl* i.

Samtaleanalyse

Elisabet Engdahl ser i artikkelen *Frågor i NoTa* nærmere på ulike spørsmålsformuleringer. Hun ser både på *ja/nei*-spørsmål, på ekkospørsmål og på spørsmål med deklarativ ordstilling, og tar opp noen bruksbetingelser

for de forskjellige spørsmålstypene. Engdahl er også opptatt av korpuset som forskningsverktøy, og diskuterer hvordan transkribørene har annotert spørsmålene, samt hvordan korpuset fungerer. Heldigvis er konklusjonen positiv!

Inger Margrethe Hvenekilde Seim er opptatt av innholdet i samtaler. Tittelen på artikkelen, *Innhold og struktur i en samtale mellom to ungdommer i et flerkulturelt miljø i Oslo*, er beskrivende for hva hun tar opp. Seim er interessert i hvilke emner de to guttene med innvandrerbakgrunn er innom i sin timelange samtale. Samtidig ser hun på samtalens forløp og hvordan emnene tas opp. Seim påpeker som spesielt interessant at guttene eksplisitt tar opp emnet etnisitet, både sin egen gruppetilhørighet, hverandres, og andres. Akkurat dette emnet kommer guttene inn på flere ganger i løpet av samtalen.

Jan Svennevig tar opp et nytt svaruttrykk i artikkelen "*Ikke sant" som respons i samtale*. Svennevig sammenligner flere små, eldre samtalekorpus og NoTa-korpuset, og viser at *ikke sant* brukt som respons har økt fra null i 1994 til en tredel av alle forekomster i 2005. Dette er mest vanlig blant yngre mennesker. Svennevig viser flere betydninger for uttrykket brukt som respons, og felles for dem er at det brukes som bekreftelse på noe som er sagt i samtalen. Interessant nok kan dette virke ganske irriterende på samtalepartneren, og Svennevig forklarer hvorfor.

Syntaks: Ordstilling og konstruksjoner

Lars-Olof Delsings artikkel med den bergmanske tittelen *Viskningar och rop – eller hur vi undrar och förundras* redegjør for ordstillingen og ordene som er brukt i utrops- og undringskonstruksjoner. *Så du har vackra rosor!* kan man si i finlandssvensk, men i norsk må vi nevne de vakre rosene mye tidligere i utropet. Delsing tar opp mange uttrykk og avdekker en enorm variasjon mellom de skandinaviske språkene.

Gunnar Hrafn Hrafnbjargarson drøfter i sin artikkel *Substantiverte adjektiv: Det er verste jeg har hørt* en type konstruksjon som er forholdsvis ukjent, nemlig nakne superlativer, dvs. superlativer som er svakt bøyd, men likevel mangler artikkel, slik som eksemplet i tittelen. Han viser at superlativene ikke er utbrytinger, slik de tilsynelatende kan se ut, men faktisk at de er en type ikke-referensiell, substantivert predikativ.

Marit Julien tar opp ordstillingen i underordnede setninger i artikkelen *Så vanleg at det kan ikkje avfeiast – om V2 i innføyde setningar*. At det av og til finnes hovedsetningsordstilling i *at*-setninger, har lenge vært kjent, og mange forskere har prøvd å finne ut under hva slags betingelser dette er mulig. Julien viser med eksempler fra NoTa-korpuset at det ikke spiller noen direkte rolle hva slags verb som står foran *at*-setningen, slik mange tidligere

har ment. Det som teller, er utelukkende semantikk: *at*-setningen må være hevdet av taleren.

Mari Nygård, Kristin M. Eide og **Tor A. Åfarli** tar i artikkelen *Ellipsens syntaktiske struktur* for seg et spennende språktrekk som er vanlig i talespråk, nemlig lydløse ord. Uuttalte, lydløse eller stumme ord kalles gjerne ellipse, men det er ikke noen allmenn forståelse blant språkforskerne om hva fenomenet egentlig innebærer. Noen forskere mener at det som ikke er uttalt, heller ikke finnes. Nygård, Eide og Åfarli tar et klart standpunkt for at ellipse er like eksisterende i språket som de ordene vi faktisk hører. De argumenterer for dette synet fordi de tilsynelatende stumme ordene bare kan forekomme i klart definerte kontekster, og ikke er tilfeldig spredt rundt omkring.

Språkteknologi

Peter Juel Henrichsen beskriver en forholdsvis overraskende måte å lage norsk lydskrift av norsk ortografisk transkripsjon på, nemlig ved å gå via dansk! I artikkelen *NoTa – nu med lydskrift* beskriver han sitt system NoTaFon. Ved å utnytte at NoTa-korpuset er grammatisk tagget med ordklasser og bøying, samt noen enkle ordlister over de vanligste ordene som er mest forskjellige på dansk og norsk, får han automatisk laget den norske teksten om til dansk. I dansk kan han så bruke et system som finnes fra før for å gå fra ortografi til lydskrift. Ved å sette opp noen enkle regler for hvordan dansk og norsk lyd skiller seg fra hverandre (bløte og harde konsonanter, for eksempel), kan hans automatiske system så lage norsk lydskrift fra dansk lydskrift!

Fredrik Jørgensen diskuterer i sin artikkel *Automatisk gjenkjenning av ytringsgrenser i talespråk* en utfordring som må løses for at man siden skal kunne analysere talespråkets grammatikk automatisk. Mens skriftspråk vanligvis er tydelig avgrenset ved tegnsetting, er transkribert talespråk ikke inndelt på noen enhetlig måte, så her må det nytenkning til. Jørgensen viser hvordan man kan bruke automatiske metoder til å finne ytringsgrenser i NoTa-korpuset, med noen foreløpige tall over hvor vellykket det ble.

Victoria Rosén tar også opp utfordringene med syntaktisk analyse av talespråk i artikkelen *Mot en trebank for talespråk*. Roséns utgangspunkt er at hun har et analysesystem for skriftspråk som hun ønsker å bruke for talespråk. Hun har derfor valgt å følge råd gitt i faglitteraturen om å overse de typiske talespråkssidene i det transkriberte talespråket, og foretatt enkelte endringer i sitt system. Dermed kan talespråket analyseres med en avansert skriftspråksanalysator.

Åshild Søfteland og **Anders Nøklestad** beskriver i artikkelen *Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger* hvordan de har gått fram for at hvert ord i korpuset skal få en

ordklassemarkering. De fant ut at en kombinasjon av flere metoder fungerte bra, og diskuterer i artikkelen hvilke utfordringer de støtte på.

Øst er øst og vest er vest

Janne Bondi Johannessen tar i den avsluttende artikkelen *Oslospråket i tall* for seg noen generaliseringer som er gjort med bakgrunn i enkle søk i NoTa-korpuset. Det viser seg at det fortsatt er stor forskjell på østkant og vestkant. Det er også forskjeller på språket til menn og kvinner, og det er forskjell mellom skriftspråk og talespråk. Johannessen viser at på dette store feltet kan det fortsatt gjøres mye når en har et korpus som NoTa-Oslo til rådighet.

Mer informasjon – nettsteder om NoTa

- Om NoTa-korpuset, inkludert informanter, utvalgskriterier, opptakssituasjon, utstyr, transkripsjon og øvrig prosjektinformasjon:
  http://www.tekstlab.uio.no/nota/oslo/index.html
- Om Oslomålet – et seminar med forskning fra NoTa-korpuset:
  http://www.tekstlab.uio.no/nota/oslo/seminar.html
- Om å få tilgang til å bruke NoTa-korpuset:
  http://omilia.uio.no/swamp
- Om andre korpus ved Tekstlaboratoriet:
  http://www.hf.uio.no/tekstlab/korpus.html