

Tekstlaboratoriet

Det historisk-filosofiske fakultet

Boks 1102 Blindern

0317 Oslo

Innhold

- Gjest i dette nummer
- Anmeldelse:
nynorsk-oversetter
- Seminar i Trondheim
- Konferanse på Lysebu
- Nye tilbud
- Tagger-prosjekt og
nasjonalt samarbeid
- Kort om Tekstlaboratoriet



Tekstlaboratoriet
Det historisk-filosofiske
fakultet

ΒΥΛΛΕΤΙΝ
Nr. 1/vår 1996 • 4. årgang

UNIVERSITETET I OSLO

Nyhets-
βυλλετιν

fra Tekstlaboratoriet, HF, UIO

Nr.1/vår 1996 • 4. årgang

Innhold

- Gjest i dette nummer
 - Anmeldelse: nynorsk-oversetter
 - Seminar i Trondheim
 - Konferanse på Lysebu
 - Nye tilbud
 - Tagger-prosjekt og nasjonalt samarbeid
 - Kort om Tekstlaboratoriet
-

• Gjest i dette nummer

Navn:	Jens Braarvig	
Stilling:	Førsteamanuensis i religionshistorie	
Sivilstatus:	Gift, datter på 10 år	
Hobby:	Ingen	

Jens Braarvig er førsteamanuensis i religionshistorie og bestyrer ved Institutt for kulturstudier. Hans interessefelt er alle religioner, og i studiet av disse har han hatt stor nytte av elektroniske hjelpemidler.

- Hva vil det si å være religionshistoriker?

- Vel, jeg er i bunn og grunn filolog. Jeg studerer religionshistorie med filologiske metoder - selvsagt i tillegg til de andre religionshistoriske metodene. Religionshistorie innebærer studiet av alle religioner, og gjerne da med et sammenlignende perspektiv: På engelsk heter faget Comparative Religion. Ved siden av lingvistikk er det det eneste universelle faget her ved fakultetet, ihvertfall i praksis om ikke i prinsippet.

Vi prøver å bevare de filologiske metodene, til tross for at studentene ikke lenger kan bruke originalkilder til sine studier. Vi arbeider med å styrke filologien i vårt fag og i idé- og kulturfagene. Uten filologien står mange av fagene i fare for å bli ren sengelektyre. For å drive med religionshistorie trengs egentlig kunnskap om både de klassiske filologiene, dvs. gresk og latin, og andre relevante historiske filologier, som klassisk arabisk, sanskrit, kinesisk og gamleegyptisk.

- Hva er ditt spesielle interessefelt?

- Jeg er opptatt av mytologi som universelt fenomen. Mine spesialiteter er arkaisk gresk religion og indisk religion, med særlig vekt på buddhismen. Filologiene er et absolutt nødvendig redskap for meg. Gresk og latin er jo nødvendig, selvsagt, og i tillegg er den trilinguale filologien med sanskrit, tibetansk og kinesisk helt nødvendig for buddhologien. Jeg må nevne at ingen av buddhologistudentene kan alle tre språkene, dessverre. De fleste kan ett, eller i høyden et par.

Jeg driver med en slags kulturlingvistikk: jeg ser på hvordan buddhismen faller ned og viser seg i de tre språkene, og hvordan språkene selv påvirker kulturen. Det buddhistiske tankeparadigmet er konspert i sanskrit og pali, og transformeres når det kommer i kontakt med den kinesiske kulturen og det kinesiske språket, som allerede har en tung, kulturell semantikk. Tibetansk, derimot, har hatt en enklere kultur, og buddhismen har dermed satt et sterkere preg på tibetansk språk og kultur enn på kinesisk. Forholdet mellom språk og religion er veldig spennende.

- Hvordan kan man holde to så ulike spesialiteter ved like?

- Jeg har et prosjekt som går ut på å sammenligne begge tradisjonene. Jeg ønsker å se på overgangen fra mytisk tenkning til abstrakt, filosofisk tenkning. Jeg vil altså sammenligne den greske og den indiske kulturen. Det er jo interessant at det abstrakte språkets fødsel skjedde samtidig begge steder, rundt 500 f.Kr. Min tese er at begge har en tilsvarende bevegelse. Selvsagt er begrepsapparatet helt forskjellig i gresk og indisk filosofi, men problemene og de strukturelle formene kan være parallelle.

Antikken og den indiske kulturen er de to språklige tradisjoner i verdenshistorien som i størst grad har skapt et metafysisk språk med et abstrakt begrepsapparat. Å påvise dette er et filologisk, og dessuten et historisk og filosofisk anliggende.

La meg ta noen eksempler. Det latinske ordet *spiritus* (ånd) er blitt et metafysisk begrep, som kommer fra det konkrete *spirare* (puste). Tilsvarende kommer det greske *psyche* (sjel) fra *psycho* (blåse), og sanskrit *atman* (selvet, den evige sjel) fra samme rot som tysk *Atem* (pust). Dette er faktisk uavhengige utviklinger, som har foregått på forskjellige tidspunkter.

Innenfor denne delen av min forskning har jeg nylig skrevet en artikkel om ordet *gud*, som antagelig kommer fra sanskrit *huta* (uthelt, gytt), perfektum partisipp av roten *hu-*. Opprinnelig hadde dette forbindelse med drikkofferet eller libasjonsofferet, men fikk senere en abstrakt/metafysisk betydning helt løsrevet fra sitt opprinnelige substrat. Filologien med sine etymologier kan slik altså gi oss innsikt i det prehistoriske mennesket og dets tenkesett.

- Hvordan kan elektronikken komme en så metafysisk forskning til gode?

- Jeg har faktisk hatt uvurderlig nytte av datateknologien bl.a. til min doktorgradsavhandling *Akṣayamatinirdeśasūtra*, som er en kritisk utgave og oversettelse (til engelsk) av en buddhistisk læretekst om et etisk-filosofisk system, hvor en viktig del er de buddhistiske evighetsforestillingerne. Jeg begynte med datateknologi allerede i 1982, og har vært igjennom mange generasjoner med tekstbehandlingsprogrammer. Med de mange fontene, som jeg dels måtte lage selv, og de mange manuskriptene jeg måtte skrive inn på de tre språkene (tibetansk, kinesisk og sanskrit), var det mye arbeid. Men jeg fikk mye igjen for det, ved at jeg enkelt kunne sammenligne tekstene, og skrive kommentarer der det passet. Jeg laget også en index verborum.

Når det gjelder tekstsøking, er jo Thesaurus Linguae Graecae uvurderlig for den greske litteraturen. Og nå har jeg nettopp fått den store, buddhistiske Pali-kanonen, som man kan søke i, den er riktignok transkribert. De gammeldagse indeksene er jo dermed ikke så nyttige lenger. Og det er jo en stor fordel at man selv kan søke på hva det skulle være. Men det skal sies: Det er ikke bare fordeler med den nye

teknologien. De som laget indeksene, lærte jo språket veldig godt, og det gjør man ikke i dag. Det var en levende tradisjon, i dag er tekstene døde inni maskinen.

Det er en stor utfordring for oss å bruke det, og når det gjelder begrepene utvikling, er de elektroniske tekstene selvsagt veldig nyttige. Fremdeles er det jo svært mange tekster som ikke ligger der. Min egen avhandling skal snart utgis elektronisk. Og mye orientalistisk materiale begynner å komme på elektronisk format nå. Jeg er svært optimistisk på vegne av religionsforskningen i fremtiden nettopp på grunn av datateknologien. Men samtidig er det viktig at teknologien ikke blir en sovepute som lokker studentene til å ta lett på den filologiske siden av forskningstradisjonen.

Braarvig smiler, og vi aner at den entusiasmen han har overfor fagets filologiske sider, er den beste forsikring mot den lettvinthet han frykter hos studentene sine.

जयशङ्कराष्टो ताम्रापन्न

[illegible]

Eksempel på tekst som Braarvig arbeider med. Denne er fra Nepal.

Anmeldelse: nynorsk-oversetter

Tekstlaboratoriet har, sammen med Institutt for nordistikk og litteraturvitenskap, gått til innkjøp av oversettelsesprogrammet Nyno for Windows. Nyno oversetter tekster fra bokmål til nynorsk, og er utviklet i Bø i Telemark. Det er tilgjengelig for alle ved HF-fakultetet, men for at man skal kunne bruke det på sin maskin, må man få hjelp av Tekstlab-ingeniøren første gang til å installere det. Man trenger en PC med Windows. De som ikke selv har slikt utstyr, kan komme og bruke programmet på Tekstlaben, hvor vi har gjestemaskiner.

For den uinnvidde kan det kanskje virke såre enkelt å konstruere et program som oversetter fra bokmål til nynorsk. Det er tross alt ikke mye som skiller de to skriftspråksvariantene fra hverandre, så det eneste man skulle tro programmet trenge å gjøre, var å slå opp i en ordbok og finne det tilsvarende ordet der. Men så lett er det langt ifra, som vi snart skal se.

Menyvalgene

La oss, før vi ser på det problematiske, ta for oss noen av valgmulighetene programmet gir. Når man åpner programmet, kommer man inn i et tekstvindu med et fullt menysystem. Den teksten som skal oversettes, kan man hente fram fra en fil, eller man kan skrive den inn der og da. Som alle vet, er det svært mange varianter av nynorsk, og programmet gir oss full valuta for pengene i denne henseende. Drar vi ned menyen Bøying, kan vi velge å presisere substantiv-, verb- og adjektivbøying. Velger vi det første, må vi igjen foreta et valg, nå mellom Valfrie hovedformer og Sideformer. Velger vi igjen det første, kan vi bl.a. krysse av for om vi ønsker at intetkjønnsord som ender på -um skal få "fleirtal på -um, -uma eller -a, -a. Eks. faktum - fleire faktum eller fakta". Under verbbøying kan vi blant de valgfrie hovedformene bl.a. velge infinitivsending (*å reise* eller *å reisa*), preteritumsending (*lærte* eller *lærde*) og perfektumsending (*har vore*

eller *vori*). Blant sideformer av svake verb kan vi f.eks. velge om det skal være samsvarsbøying ved perfektum partisipp (*han er dømd/dømt, dei er dømde/dømt*). Blant andre sideformer kan vi velge ending for presens partisipp (*hoppende* eller *hoppande*), og om sterke verb skal ha ending i presens (*finn* eller *finner*). For adjektiver kan vi bl.a. velge intetkjønnsending (*ope* eller *opi*).

I tillegg til bøyingsformer er den andre store forskjellen mellom nynorsk og bokmål orddanning og ordforråd. Igjen er Nyno imponerende. Vi har en egen meny Ordval, og her har vi muligheter for å bestemme innenfor områdene Konservativt språk, Lånord, Genus, Andre og Sideformer. Under valget Konservativt språk, kan vi bestemme oss for konservativt eller radikalt ordval (*skole* eller *skule*), utlyd i adjektiver (*bakket* eller *bakkut* eller *bakkete*), diftong eller ikke (*lydhør* eller *lydhøyr*), eller forstaving (*forsmak* eller *føresmak*). Under Lånord kan vi bl.a. velge skrivemåte, f.eks. utenlandsk eller "språkrådsinspirert" ortografi (*clutch* eller *kløtsj*), og hvordan greske diftonger skal skrives (*neurolog* eller *nevrolog*). Under Genus kan vi velge f.eks. mellom hankjønn og intetkjønn (*ein* eller *eit baksmell*). Punktet Andre gir oss valg som fugeformativ -e (*fuglunge* eller *fugleunge*), fugeformativ -s (*nedbørmengd* eller *nedbørsmengd*), form på sammensetninger med tallord (*femdel* eller *femtedel*) osv. Sideformer gir oss muligheten til å velge ord som er nærmere bokmål (*bilete* eller *bilde*).

Alternativ-menyen gir muligheten til å være interaktiv eller ikke. Man kan velge om man skal godkjenne den oversatte setningen før den blir sendt til fil, og man kan velge å tas med på råd om bruken av genitiv. Programmet gir fire forskjellige måter å oversette genitiv på. F.eks. kan *mannens dagbok* oversettes med preposisjon (*dagboka til mannen*), garpegenitiv (*mannen si dagbok*), s-genitiv (*mannens dagbok*) og samanskriving (*mannedagboka*). Videre kan man konsulteres ved omskriving, alternative eller kjente ord. Dessuten kan man krysse av for hvordan man ønsker at genitiv skal oversettes i forbindelse med navn (personnavn såvel som geografiske og institusjonsnavn), og hvordan passiv skal oversettes (med verbet *bli* eller *verte*). Oversettelsen av

genitiv er nok den delen av programmet som viser at Nynorsk virkelig har tatt mål av seg til å være noe mer enn en ord-for-ord-oversetter, og at det er idiomatisk nynorsk som tilstrebes. Og oversettelsesmulighetene her er ikke trivielle. Ser vi på oversettelsen til preposisjonsfrase, har det skjedd en endring av ordstillingen, og i tillegg er hovedleddet (*dagbok*) satt i bestemt form, og preposisjon er føyet til. Veien fra genitiv til sammensetning er enda vanskeligere. I sammensetningen *mannedagboka* er ikke bare ordet i genitiv føyet til hovedleddet, men det er også lagt til en fuge-e. Alle som har arbeidet med sammensetning, vet at det er svært komplekse regler for når sammensetninger kan få fuge-e, fuge-s eller ingenting. Riktignok skal vi se nedenfor at Nynorsk ikke alltid gjør rett her, men forsøket er verdt å ta av seg hatten for.

Problematiske oversettelser

La oss se på noen setninger som inneholder problematiske ting. I setning (1a) er det flere vanskeligheter som programmet må løse.

- (1) a. Dette fellesskapet på skolen kan man beskrive som solen min.

For det første har ordet *fellesskap* intetkjønn på bokmål, og hankjønn på nynorsk. Determinativen *dette* skal kongruere i genus med substantivet, det betyr at *dette* må endres til *denne* pga substantivets endrede genus. For det andre har vi i bokmål brukt et ord med tysk prefiks, *beskrive*, og det er ikke god tone i nynorsk. For det tredje har vi brukt ordet *man*, som heller ikke skal brukes i nynorsk. For det fjerde har vi tatt med atter et substantiv med ulikt genus i bokmål og nynorsk, nemlig *sol*, som kan være hankjønn på bokmål og hunkjønn på nynorsk. Vi har tatt med et etterstilt possessiv-pronomen, og slike skal alltid kongruere. Her får vi altså et tilsvarende problem som med den foranstilte determinativen *dette*.

Når vi ber Nynorsk om å oversette setningen i (1a), blir vi først bedt om interaktivt å bestemme hvilket verb som skal velges istedenfor "fyordet" *beskrive*. Her kan vi velge mellom *skildre*, *teikne*, *framstille*, *karakterisere* og *nemne*. Etter at vi har valgt, blir resultatet som i (1b).

(Man kan jo sette spørsmålsteget ved hvorfor nynorsk bør foretrekke ord med gresk snarere enn tysk etymologi, men dette er ikke først og fremst Nynorsk ansvar.) I resten av omtalen vil (a)-eksemplene være bokmålstekster og (b)-eksemplene være Nynorsk oversettelse av dem.

- (1) b. Denne fellesskapen på skolen kan ein skildre som sola mi.

Vi må vel si at resultatet ville tilfredsstilt selv den strengeste norsklærer. Valg av genus og av ord gikk greit, og kongruens mellom determinativ/possessivpronomen og hovedledd likeså.

I setning (2a) er det to nye problemer. Det første er særdeles interessant fordi det så tydelig illustrerer at man ved oversettelse fra bokmål til nynorsk ikke bare kan plukke ut ord fra ordboka. De to ordene *lærer* i bokmålsteksten er jo henholdsvis substantiv og verb, og må, i motsetning til på bokmål, få forskjellig form. Det andre problemet er ordet *like*, som kan være både adjektiv, verb og adverb, og som igjen trenger ulik form på nynorsk avhengig av ordklasse. Resultatet er gitt i (2b):

- (2) a. En lærer lærer bort like mye som elevene kan.
b. Ein lærar lærer bort like/like/like mykje som elevane kan.

Det første problemet har Nynorsk løst helt tilfredsstillende. Den første forekomsten av *lærer* er analysert som substantiv, og har dermed fått formen *lærer*, den andre er analysert som verb i presens, *lærer*. Det andre problemet har ikke Nynorsk løst. Det har ikke klart å velge mellom ulike analyser av *like*, og har rett og slett presentert alle løsningene.

I eksempel (3a) ser vi på spørreordet *hvor*. På nynorsk finnes det tre oversettelser av *hvor*, avhengig av om det er et spørreord for geografisk sted (*kvar*), et spørreord for grad (*kor*), eller et relativord (*der*).

- (3) a. Hvor bor du? Det er byen hvor jeg er best kjent. Hvor stor er du?

- b. Kvar bur du? Det er byen der eg er best kjend. Der/Kor/Kvar stor er du?

Som vi ser i (3b) har Nyno klart oversettelse til det geografiske spørreordet og relativkonjunksjonen, men gradspørreordet har det ikke klart å gjenkjenne. På bakgrunn av at programmet tydeligvis har bestemt de to førstnevnte ordene ut fra det påfølgende ordet (verb eller nomen), er det underlig at ikke gradspørreordet kunne bestemmes ut fra et påfølgende adjektiv eller adverb. Vi får anta at dette rettes i en eventuelt ny utgave.

Samsvarsbøying i frie predikativer er et annet problem enn samsvarsbøyingen vi så på tidligere. Her står ikke ordet som skal samsvarsbøyes, rett ved siden av ordet det skal bøyes med (stjerne markerer ugrammatisk resultat):

- (4) a. Kråkene fløy skremt bort.
b. Kråkene flaug skremde bort.
- (5) a. Skremt fløy kråkene bort.
b. *Skremt flaug kråkene bort.
- (6) a. Bort fløy kråkene skremt.
b. Bort flaug kråkene skremde.

Nyno takler den frie predikativten når ordstilling gir subjektet foran predikativten, (4b) og (6b). Men når ordstillingen endres slik at predikativten står foran subjektet, (5b), må Nyno melde pass. Vanlige subjektspredikativer følger mønsteret for frie predikativer:

- (7) a. Barna ble funnet ved godt mot.
b. Barna vart funnne (sic!) ved godt mot.
- (8) a. Funnet ved godt mot ble barna.
b. *Funnet ved godt mot vart barna.

- (9) a. Ved godt mot ble barna funnet.
b. Ved godt mot vart barna funnne.

Objektspredikativer er litt mindre vellykket i Nyno:

- (10) a. Han kaller dem verdenskjent.
b. Han kallar dei verdskjende/verdskjend
- (11) a. Verdenskjent kaller han dem.
b. *Verdskjent kallar han dei.

Nyno greier ellers predikativkongruens bra, selv fra overledd og inn i en relativsetning:

- (12) a. Kvinnen så to måner som var tatt rett ut av boken.
b. Kvinna såg to månar som var tekne rett ut av boka.

I noen grad kan man forstå at samsvarsbøyingen ved predikativer i Nyno er mindre vellykket. For å få det rett måtte man krevet en omfattende syntaktisk analyse. Men når programmet ikke klarer samsvarsbøying ved koordinering, må det betraktes som en glipp. Det skal ikke store analysen til for å finne at to nomenfraser er koordinert med hverandre, slik at to entallsfraser gir flertallsbøying:

- (13) a. En gutt og en jente var stukket av vepsen.
b. *Ein gut og ein jente var stukken av kvefsen (Skulle vært stukne.)

Passiv er forholdsvis godt behandlet i Nyno, men (16) viser igjen at når ordstillingen ikke er subjekt-verb, blir oversettelsen gal.

- (14) a. De to kuene skal ha blitt slått helseløse.
b. Dei to kyrne skal ha vorti slått helselause.

- (15) a. Kuene melkes tomme.
b. Kyrne blir mjølka tomme.

- (16) a. Tomme melkes kyrne.
b. *Tomme blir mjølka kyrne.

Nyno følger regelen om morfologisk passiv i nynorsk: Slik passiv er tillatt hvis den forekommer i forbindelse med et modalt hjelpeverb, og ellers ikke. I motsetning til i (15) ovenfor blir den morfologiske passiven beholdt i (17), og det er fint:

- (17) a. Kyrne skal melkes.
b. Kyrne skal mjølkast.

Men atter en gang er det slik at når ordstillingen endres, blir resultatet som galest:

- (18) a. Melkes skal de.
b. *Blir Mjølka (sic!) skal dei.

Ordstillingen i setninger med hjelpeverb, hovedverb og adverb, og med foranstilt preposisjonsfrase, er kanskje problematisk i seg selv, men i Nyno ser vanskelighetsgraden ut til å variere med hvilken preposisjon som tilfeldigvis innleder preposisjonsfrasen. Den feilaktige oversettelsen i (20b) har ikke bare feil ordstilling, men også feil partisippform.

- (29) a. På fredag selges nok kuene.
b. På fredag blir nok kyrne selde.
- (20) a. Om vinteren selges nok kuene.
b. * Om vinteren blir seld nok kyrne

Det ble nevnt ovenfor at sammensetninger er forsøkt behandlet i Nyno på en fullstendig måte med fuge-s og fuge-e. Målsetningen er dessverre ikke nådd:

- | | |
|-------------------------------------|-------------------------------|
| (21) a. den lærerens område | b. *den lærarsområdet |
| (22) a. den vognas hjul | b. *den vognehjulet |
| (23) a. den guttens hår | b. *den gutshåret |
| (24) a. det husets vindu | b. *det hussvindaug |
| (25) a. den mannens bein | b. *den mannebeinet |
| (26) a. den rørleggerens bedrift | b. *den røyrleggjarsbedrifta |
| (27) a. det gardinopphengets kroker | b. *det gardinopphengekrokane |

I tillegg ser vi at det mangler samsvarsbøying mellom determinativ og sammensetninger som er dannet av programmet selv. Dette er jo i motsetning til det vi så ovenfor i forbindelse med eksempel (1), og burde være en smal sak å rette på i en eventuell ny utgave.

La oss ta med en siste forskjell mellom bokmål og nynorsk. I bokmål brukes pronomenene *den* og *det* for å referere til ting. På nynorsk er derimot pronomenene kjønnsdelt. Det klarer ikke Nyno å takle:

- (28) a. Per sto på stolen, men den knakk.
b. Per stod på stolen, men den knakk. (Burde vært: *han knakk*)

Dette punktet er likevel mindre kritikkverdig enn de ovenstående, fordi det er vanskelig å tenke seg hvordan et dataprogram skal kunne velge mellom det påpekende pronomenet *den* og det vanlig refererende *han/hun*. Likevel burde alternativene gis her som andre steder hvor det er tvil om riktig valg, slik at eventuelle brukere lett kan finne de stedene det er snakk om ved korrekturgjennomgang senere.

Ordlista til Nyno er nokså begrenset. Det er derfor uunngåelig at en del ord markeres som "ukjent" av programmet. Dette gjelder selv vanlige ord som *kurvstol*, *slukke* og *syd*. Heldigvis er det mulig å legge inn nye ord i lista, slik at programmet kjenner ordene neste gang. Dette gjøres gjennom menyen, og man får også enkelt lagt inn ordklasse og bøyingstype, ved at programmet selv viser hvilke alternativer som er aktuelle.

Avslutningvis vil jeg si at Nyno er et interessant forsøk på å få til en automatisk oversetter fra bokmål til nynorsk. Brukergrensesnittet er vellykket, med forståelige menyer, og imponerende mange valg m.h.t. nynorskvariant. Selve oversettelsesresultatet gir et mer blandet inntrykk. Programmet er sterkest når det gjelder setninger med vanlig subjekt-verb-objekt-ordstilling, og ganske svakt når andre andre ord enn subjektet står først i setningen. Omskriving fra genitivsfrase til preposisjonsfrase løses fint, men oversettelse til sammensetting er ikke så vellykket. At Nyno ikke er mer enn en maskin, kommer tydelig fram i oversettelsen av *for min del til for delen min*. Kritikken til tross, for mange formål vil Nyno sikkert være til stor hjelp i oversettelsesarbeidet. Man må bare være klar over at resultatet må korrekturleses av en nynorskkyndig person.

Seminar i Trondheim

I april var Tekstlaboratoriets forsker invitert til et korpusseminar arrangert av Nordisk institutt i Trondheim for å holde et foredrag om bruk av elektroniske tekstkorpora. Det var også foredrag av Eiríkur Rögnvaldsson fra Island, Jan-Ragnar Hagland og Torbjørn Nordgård.

Mange interessante emner ble berørt. Nordgård snakket om typer av korpora og vanskeligheter med å definere representativitet. Rögnvaldsson viste eksempler fra den nye CD-ROM'en som er under utvikling på Island, hvor gamle sagaer legges inn med mulighet for konkordanser og annen interessant informasjon. Det er interessant, og forsåvidt talende, at det mest frekvente substantivet både i sagaene og i moderne islandsk er *maður*. Til gjengjeld er det nest mest frekvente ordet i sagaene *konungur*, og dette finnes ikke blant de 100 mest frekvente moderne islandske substantivene. Enda mer interessant er det at islandsk ser ut til å ha endret kasusfordelingen. Mens det i sagaene var 41 prosent av substantivene som hadde akkusativ og 22 prosent dativ, er forholdet nå 31 og 30 prosent. Hva denne utjevningen skyldes, er ikke godt å si, men det ble foreslått at det hadde sammenheng med at

kasusystemet på islandsk har stabilisert seg.

Det er sikkert ikke alle av Nyhetsbulletinens lesere som er klar over at Universitetet i Oslo faktisk er i en særstilling hva Tekstlab-tilbud angår. Det finnes ikke noe tilsvarende ved de andre universitetene. Vi får håpe at seminaret i Trondheim er starten på noe mer også for deres del.

Konferanse på Lysebu

Kollegiet har nå åpnet for et nytt satsingsområde ved Universitetet, nemlig Kommunikasjon: Teknologi og kultur. I den forbindelse ble det arrangert en konferanse på Lysebu, ledet av Kristin Braa (informatikk) og Dag Wiese Schartum (rettsinformatikk). I slike sammenhenger har det lett for å bli mange samfunnsfaglige innfallsvinkler av typen "Hvordan påvirker teknologien oss?" og "Hvordan kan vi påvirke teknologien?".

Det var derfor gledelig at det blant arbeidsgruppene på konferansen var en som omhandlet kommunikasjonsteknologi og språk. På denne gruppen var det fire personer: Kjell Lars Berge (INL), Svennik Høyer (IMK), Nicolay B. Johansen (kriminologi) og Tekstlaboratoriets forsker. Berge er i gang med et forskningsprosjekt som omhandler den nye teksttypen Skriftbasert dialog i det virtuelle rommet. Nærmere bestemt skal han (i samarbeid med Svein Lie, også INL) studere møter på nettet, hvor altså møtedeltagerne diskuterer i et skriftlig medium, men hvor "samtalen" likevel foregår samtidig. Berge og Lie er interessert i å se på i hvilken grad denne sjangeren har andre trekk enn mer tradisjonelle skriftlige og muntlige tekster. Høyer er interessert i å studere journalistiske endringer som følger av endringer i teknologien. Han nevnte som et eksempel det faktum at mens avisartikler tidligere var formet som en omvendt pyramide, med mest generell informasjon øverst, og så mer og mer detaljert informasjon nedover i artikkelen, er artiklene nå mer utjevnet, og informasjonen er spredt jevnt ut. Dette følger direkte av at artikler i blytsatsens tid vanskelig kunne kuttet (noe

som ofte er nødvendig i avisene) andre steder enn nederst, mens elektronisk sats gjør at kutting kan skje hvor som helst i artikkelen.

Tekstlabens forsker redegjorde for to forhold. For det første ble rettighetsproblemet med norske forfattere og forlag (særlig de respektives foreninger) tatt opp. Det er som kjent et problem at det er vanskelig å få tilgang på elektroniske tekster av moderne forfattere, særlig tekster av en viss lengde og til mer varig bruk. Når språk- og litteraturforskere ønsker å bruke nyere metoder i forskningen, slik som elektronisk tekstsøking, blir de hindret fordi utbudet av tekster er lite. Dette er jo et alvorlig problem for et lite språk som det norske. Her burde man tvert imot hatt ekstra store muligheter til forskning på eget språk. Alternativet kan være at man går over til å forske i engelske tekster, hvor utvalget er enormt.

Det andre punktet som ble tatt opp av Tekstlabens utsendte, gjaldt et nytt forskningsprosjekt som forhåpentligvis kan bli satt ut i livet, med egen forskningsassistent, men ledet av undertegnede og Kjell Johan Sæbø (GI). Prosjektet går ut på å avdekke regler for pronominal referanse. Det er jo en interessant forskjell mellom nynorsk og bokmål m.h.t. pronomener. Mens bokmål skiller mellom animate og ikke-animate objekter (*han/hun/det* vs. *den/det*), har nynorsk (ved språkintern referanse) bare et genuskille (og bruker altså bare *han/hun/det*). Når en datamaskin skal oversette, er det nødvendig at bruksbetingelsene for referanse avdekkes og beskrives konkret.

Nye tilbud

Program: Oversettelsesprogrammet Nyno for Windows er kjøpt inn av Tekstlaboratoriet og INL i fellesskap. Se egen omtale.

Tekster: Tekstlaben har kontaktet alle de norske forfatterne og forlagene som har stilt tekster til disposisjon for parallell-

korpusprosjektet ved IBA. De fleste aksepterte at også Tekstlaben får tilby disse tekstene. Vi kan derfor gi tilgang til utdrag fra 28 moderne romaner og noveller (av f.eks. Herbjørg Wassmo og Knut Faldbakken), og en del sakprosa (f.eks. fra en bok om samer og en om Norges historie). Tekstene ligger på samme område på Tekstlab-disken som de andre norske tekstene, nemlig under Norsk. Her ligger også en Lesmeg-fil med opplysninger om tekstene. Eget skjema må underskrives av den enkelte forsker for å få tilgang til tekstene.

Pekere fra Tekstlabens Web-side: Fra Tekstlabens hjemmeside kan man finne fram til en ordbok i særklasse, nemlig Grunnmanuskriptet, som er grunnlagsmateriale for Norsk Ordbok (Avdeling for leksikografi). Den er tilrettelagt for nettsøking av Christian-Emil Ore ved Dokumentasjonsprosjektet, og det er virkelig verd å ta en titt, selv om man ikke skulle ha direkte faglig nytte av det. Her kan man søke på ord fra hele landet, eller begrense seg til et fylke eller to, og man kan f.eks. velge om ordet skal være forledd eller etterledd i en sammensetning. Prøver man å slå opp ordet *kvende* (kvinne), får man tilslag på ord som *bengalkvende*, *bokkvende* (=forfatterinne) og *brotskvende*. Slår man opp gjente, får man blant annet *floggjente* (jente som flyr) og *fornegjente* (jente som kommer med sending).

Tagger-prosjekt og nasjonalt samarbeid

Mange av Tekstlaboratoriets brukere kunne nok ønsket seg at de norske tekstene var grammatisk merket. Skal man for eksempel studere verb i presens, kan man ikke bare be maskinen hente dem ut, for maskinen vet ikke hvilke det gjelder. Ber man om alle ord som slutter på -er, får man ikke bare *spiser* og *sover*, men også *snekker*, *fotballspiller*, *computer*, *vakker* og *sikker*, samtidig som man ikke får *drar*, *glir*, *spår*, *gryr* og *før*.

Det er derfor gledelig at Tekstlaboratoriets forsker Janne Bondi Johannessen sammen med professor Jan Tore Lønning ved ILF har fått

støtte fra Norges forskningsråd til et prosjekt som nettopp har som mål å lage en tagger for norsk (bokmål og nynorsk). Om to og et halvt år er det å håpe at vi har taggerne klare, slik at fakultetets forskere skal kunne komme med vanlige tekster og få dem grammatisk tagget av oss. Det dreier seg om et stort prosjekt på til sammen syv årsverk, inkludert assistenter og programmerer. To av årsverkene finansieres av Dokumentasjonsprosjektet, som vi har et nært samarbeid med.

Tagger-prosjektet er en del av et nasjonalt samarbeidsprosjekt, NIFST (Nasjonal infrastruktur for språkteknologi). Her inngår også Universitetet i Trondheim med et datalingvistisk leksikon-prosjekt og Universitetet i Bergen og Humanistisk datasenter, Bergen, med et korpus-prosjekt. Vi ser det som viktig at det finnes språkteknologiske ressurser for norsk, slik at man også kan benytte moderne forskningsmetoder for vårt språk.



Kort om Tekstlaboratoriet

Tekstlaboratoriet disponerer en lang rekke tekster som er elektronisk tilgjengelige. Tilbudet på norsk og engelsk er stort, her har vi en god del romaner, lovtekster og avistekster, i tillegg til noen andre typer tekster og noen ferdigsamlede korpora. Vi har også en del elektroniske tekster og ordbøker for andre språk, som albansk, bulgarsk, dansk, estisk, finsk, fransk, gresk, gælisk, italiensk, japansk, kinesisk, latin, litauisk, malayisk, nederlandsk, portugisisk, russisk, serbo-kroatisk, spansk, svensk, tibetansk, tsjekkisk, tyrkisk, tysk og walisisk. Tekstene er av forskjellig type, noen er skjønnlitterære, andre sakprosa, og noen inneholder grammatisk merking, mens andre ikke gjør det. Noen av tekstene er flerspråklige (parallellkorpus), slik at de inneholder originale og oversatte tekster mellom språk. Utvalget utvides stadig. Særlig ser Tekstlaben det som viktig å øke mengden av norske tekster, men vi prøver også å innhente tekster fra andre språk i den utstrekning vi har grunn til å tro at det er behov for dem. Dessuten kan vi være behjelpelige med å få tak i spesifikke tekster, hjelpe til med å scanne dem inn osv.

I tillegg til tekstene har Tekstlaben en del programvare som er nyttig for å søke i tekstene, og en del programvare som kan være interessant og nyttig for språk- og litteraturforskere mer generelt. Dette gjelder f.eks. WordNet, som inneholder et korpus av engelske leksemer relatert semantisk til hverandre som synonymer, antonymer, hyponymer, antonymer, polysemer osv. Vi har også avanserte søkeprogrammer for PC og Macintosh, som Tact, Word Cruncher, Conc, Pandora o.a. Tekstlaben holder jevnlig kurs i disse programmene, men kan også demonstrere dem individuelt når det er interesse for det.

Tekstlaboratoriets tilbud er tilgjengelig over nettet for brukerne, dvs. vitenskapelig ansatte og høyeregradsstudenter. For at man skal kunne benytte seg av det, må man ha en konto på Hedda. Det har man rett til hvis man er ansatt eller hovedfagsstudent. I tillegg må man fylle ut et skjema hvor man lover å bruke materialet til ikke-kommersielle formål.

Tekstene er lagret på en disk knyttet til fakultetets filtjener Hedda. Se etter filer med "Lesmeg" på disken. Der finnes det nyttig og kortfattet informasjon om de andre filene.

- Slik nås disken fra Macintosh: Hent "Velger" fra eplemenyen. Velg sone HF og symbolet for AppleShare. Velg så MacHedda, og klikk OK. Skriv inn brukernavn og passord, og klikk OK. Velg MacTekstlab, og klikk OK. Nå kommer tekstlabdisken (MacTekstlab) opp på skrivebordet ditt, og da er det bare å åpne den.

- Slik nås disken fra PC: Med et passende tekstbehandlingsprogram legger du følgende linje til fila C:\nfs\drives.bat:

NET USE G: hedda:/hf/hedda/tekstlab/pc

Lagre og start maskinen på nytt. Nå er G: definert som tekstlabdisk.

- Slik definerer du ECI-tekstene for PC: Med et passende tekstbehandlingsprogram legger du følgende linje til fila C:\nfs\drives.bat:

NET USE I: hedda:/hf/hedda/cd/eci-mc1/mnt /cd

(NB! Mellomrom etter mnt)

Lagre og start maskinen på nytt. Nå er I: definert som ECI-område.

Spørsmål om oppkobling mot tekstlabdisken kan rettes til Tekstlaben eller IT-ansvarlige ved instituttene.

Tekstlaboratoriet, HF, UiO, P.b. 1102 Blindern, 0317 Oslo

Faks: 22 85 69 19

Janne Bondi Johannessen (forsker) Helge Hauglin (ingeniør)

Rom B 152, P.A. Munchs hus Rom B 150, P.A. Munchs hus

Telefon: 22 85 68 14 Telefon: 22 85 78 35

E-post: jannebj@hedda.uio.no E-post: helgeha@hedda.uio.no

tekstlab@hedda.uio.no

URL: <http://www.hf.uio.no/tekstlab/>

Ansvarlig for Nyhetsbulletinen: Janne Bondi Johannessen