

Nyhets-

ΒΥΛΛΕΤΙΝ

fra Tekstlaben, HF, UIO

3. årgang, nummer 1, april 1995

☆☆

Innhold

1. Kurs
2. Nye elektroniske tekster
3. Tekstlaboratoriet på World Wide Web
4. Hva kan Tekstlaboratoriet tilby forskerne og omvendt?
5. Visitas, tagging og forskningsprosjekt
6. Symposium og gjesteforelesninger med korpusvinkling
7. Kort om Tekstlaboratoriet

☆☆

1. Kurs

1.1 Pandora

I januar holdt Tekstlaboratoriet kurs i Macintosh-programmet Pandora, som er et søkeprogram for Macintosh. Programmet kan brukes for å søke i Thesaurus Linguae Graecae (alle greske tekster fra 800 f.Kr. til 500 e.Kr., som vi har elektronisk tilgjengelig. Tidligere har det vært holdt kurs i PC-programmet View&Find, denne gangen var det altså Macintosh-brukernes tur. Det var ca. 10 deltagere, fra Klassisk-romansk institutt, Religionshistorie og Menighetsfakultetet. Kursdeltagerne var begeistret over hva programmet kunne bety for deres videre forskning. Noen kunne ønsket at "det kritiske apparatet" (fotnoter, tolkninger osv.) også hadde vært tilgjengelig, og det burde man kanskje vente i en senere utgave av TLG.

Pandora-programmet er svært enkelt å bruke. Man kan velge seg ut en eller flere forfattere eller verker, og for eksempel søke etter hele ord, deler av ord eller forskjellige typer grupper av ord. Resultatet kan man få lagret i fil, og eventuelt skrive ut. Selvsagt brukes det greske alfabetet.

1.2 Innføringskurs

Mange av fakultets ansatte har nå gjennomført Tekstlaboratoriets tilbud. Men det er sikkert noen som har sett det litt an. Videre er det nye hovedfags- og doktorgradsstudenter som kan ha nytte av Tekstlabens tilbud. Vi vil derfor holde nye

innføringskurs (for henholdsvis PC- og Macintosh-brukere) i begynnelsen av høstsemesteret. Mer informasjon kommer senere.

☆☆

2. Nye elektroniske tekster

Tekstlaben er stadig interessert i å utvide tilbudet av tekster, slik at vi har et bredt og godt tilbud innenfor forskjellige språk, og et ekstra godt tilbud i norsk. Alle tekstene er tilgjengelige fra øverste nivå på Tekstlab-disken, under språkets navn. (Det gjelder bare Macintosh-brukere. PC-brukere må gå til det området tekstene ligger på, for å kunne finne dem.) Nye ting som vi har tilgjengelig, er:

Masse språk: Vi har en mengde tekster på forskjellige språk som vi har kjøpt inn fra European Corpus Initiative (ECI). Språkene er: **albansk, bulgarsk, dansk, engelsk, estisk, fransk, gælisk, italiensk, japansk, kinesisk, latin, litauisk, malayisk, nederlandsk, norsk, nygresk, portugisisk, russisk, serbisk, spansk, svensk, tsjekkisk, tyrkisk, tysk, uzbekisk.** Tekstene er svært uensartet, og spenner fra skjønnlitteratur og sakprosa (fra bruksanvisninger via lovtekster til aviser, oversettelser og ordbøker). Nærmere beskrivelse av tekstene fås i filen Corpdesc.txt på ECI-området. Tekstene er TEI-merket, som betyr at de er utsyrt med visse koder som gir opplysninger om tekstene. Kodene er ikke til stor sjenanse, siden de er samlet i bolker i begynnelsen av visse tekstutsnitt. Tekstene kan nås via ECI-filen, for Macintosh-brukere er de også tilgjengelige via mappen som har språkets navn. (PC-brukere må først definere CD-platen som ECI-tekstene ligger på. Dette er beskrevet i punkt 7.)

Norsk: Vi har nå fått talemålsmaterialet TAUS, basert på intervjuer med Oslo-folk. Det er morfologisk tagget (merket), altså er det noe informasjon om ordklasser og morfologiske kategorier som kjønn og tall, grad, bøyingsklasse osv. Den taggede teksten, som ligger på fila "Taus alt tekst", er skrevet i lydskrift (med latinske bokstaver) som skal gjenspeile talemålet. Skal man søke etter spesielle ord, kunne det bydd på problemer, siden man ikke alltid kan forutsi hvordan et ord uttales, og hvordan det er transkribert. Det finnes derfor en ordliste til hjelp (på fila "Ordliste m/leksemer"), hvor leksemet står i sin ortografiske form sammen med de(n) ordformen(e) som er funnet i teksten. Frekvens er også oppgitt. Et eksempel:

#annenetasje aanetasje<SB,m,ub,sg> (2)

#annenetasje aantasje<SB,m,ub,sg> (3)

Nygresk: CTI-korpuset: tekster fra avisen To Vima og noe annen sakprosa. Disse har vi hatt en stund, men nå er de blitt gjort tilgjengelige.

Parallele tekster: Som nevnt over, er ECI-korpuset svært innholdsrikt, og inneholder 13 forskjellige parallelle tekster i forskjellige språk. For eksempel finnes det en tekst i den tekniske sjangeren med oversettelser mellom 9 EU-språk av annonseringen av EUs Esprit-program. Et annet eksempel er oversettelser av lovtekster mellom engelsk, fransk og spansk, fra flere årganger av ILOs offisielle bulletin. Det finnes også skjønnlitterære tekster. Fra før har vi EØS-avtalen på norsk og engelsk. For dem som er interessert i oversettelse, skulle det derfor være noe å hente her. (Vi vil også minne om at det finnes et norsk-engelsk parallellkorpus på IBA og et norsk-tysk korpus under utvikling på GI.)

Tibetansk: Vi har lagt inn tre tibetanske tekster av Buddha (transkribert til latinske bokstaver) på Tekstlab-disken. Vi takker Jens Braarvig for tilgang til dem. Skulle det være noen som ønsker flere tekster, eventuelt med annet skriftsystem, kan de henvende seg til oss eller til ham.

☆☆

3. Tekstlaboratoriet på World Wide Web

Tekstlaben har nå sin egen hjemmeside på World Wide Web. Her ligger det pekere til forskjellige tjenester rundt i verden som vi tror språkforskere vil ha glede av. I forrige utgave av nyhetsbulletinen nevnte vi at man kan finne tekstbaser som f.eks. Oxford Text Archive (forskjellige språk og tekstsjangere, men mest engelsk) og Project Gutenberg (engelsk skjønnlitteratur). Dessuten har man bl.a. direkte tilgang til bokmåls- og nynorskordboka, Websters engelsk-ordbok og andre ordbøker.

Det ligger pekere til andre nyttige ting også. For dem som er interessert i grammatikkformalismer, har vi lagt inn pekere til noen "grammatikk-laboratorier" (PSG Laboratory 1.1.1, osv.). Dette er prøveprogrammer hvor man kan lage sine egne grammatikker i hhv. frasestrukturgrammatikk, DCG, PATR og kategorialgrammatikk. Programmene i sin fulle form kan man få kjøpt gjennom Lingvistisk institutt, Universitetet i Gøteborg.

Det er veldig enkelt å bruke World Wide Web, siden alt foregår med museklikking, og

det er helt selvforklarende. IT-ansvarlige ved instituttene kan hjelpe en til å komme i gang. Vår hjemmeside har denne adressen:

<http://www.hf.uio.no/tekstlab/velkommen.html>

Velkommen!

☆☆

4. Hva kan Tekstlaboratoriet tilby forskerne og omvendt?

For å begynne med "og omvendt": Tekstlaben har nå fungert en god stund, og vi vet at mange har benyttet seg av tilbudet i forskningen sin. Tekstlaboratoriet er veldig interessert i å høre fra dere, så vi kan få en idé om i hvilket omfang og til hvilket formål Tekstlab-tilbudet har vært brukt. Vi blir svært glade for tilbakemeldinger, muntlige eller skriftlige.

Vi kan tilby forskerne ved fakultetet en innføring i ulike søkeprogrammer for søking i elektroniske tekster. Innføringen skjer fortrinnsvis gjennom kurs, som vi avholder med ujevne mellomrom, og gjerne etter forespørsel. Videre har vi en god del elektroniske tekster på forskjellige språk. For norsk prøver vi stadig å utvide utvalget, mens vi er noe mer selektive for andre språk. Vi er imidlertid åpne for ønsker og forslag.

Ønsker noen å bygge opp sine egne elektroniske tekster, kan de få bruke Tekstlabens scanner (optiske leser). Er ikke originalen i for dårlig stand, kan man spare mye tid på å scanne i stedet for å skrive inn på nytt.

I tillegg til elektroniske tekster og søkeprogrammer har Tekstlaboratoriet også andre nyttige ting for språkforskere og studenter. Her kan vi nevne CELEX-databasen, som inneholder svært mye informasjon om ordtilfanget i nederlandsk, tysk og engelsk. Videre har vi et lingvistikk-læreprogram, "HyperLinguistics - An Educational Hypertext for English Linguistics", som kan brukes med stor nytte og glede av studenter (og forskere) med interesse for lingvistikk eller engelsk. Henvend dere til Tekstlaben for disse to tilbudene.

☆☆

5. Visitas og forskningsprosjekt

Tidligere i vinter hadde Tekstlaboratoriet den glede å få gi en aldri så liten innføring om vår virksomhet til rektoratet, i forbindelse med rektoratvisitasen ved Institutt for lingvistiske fag. Både Lucy Smith og Arild Underdal viste stor interesse for det vi hadde å vise fram. Et kritisk spørsmål fra rektors side om hvorvidt Tekstlabens tilbud

egentlig ble brukt noe særlig, ble heldigvis besvart med et unisont "Ja!" fra resten av følget.

Rektor fortalte også en liten historie om sin mann, som hadde arbeidet med samespørsmål. Han hadde tilgang til en elektronisk tekstbase, og søkte da etter ordet "same". Resultatet ble ikke så nyttig som han hadde håpet, siden ordet ikke bare kan være substantiv, men også adjektiv eller pronomen. Han fikk da selvfølgelig litt for mange tilslag av feil kategori.

Det er velkjent at mange ord med ulik betydning og grammatisk status skrives likt. Homonymi kan faktisk være et så stort problem at det kan frarøve en lysten til å gå i gang med søking i elektroniske tekster. En måte å unngå det problemet på, er å tagge (merke) alle ordene i tekstene med grammatiske merkelapper. Men å gjøre det manuelt er mer tidkrevende enn at det er realistisk å gjøre det i stor skala. Automatisk tagging er derfor nødvendig. Når en ordbok konsulteres av en automatisk tagger, er man likevel ikke så mye nærmere målet, siden taggeren ikke uten videre kan gjette hvilket av de ulike homonymene som er den som skal velges for den aktuelle merkelappen.

Tekstlaboratoriet har her konkrete planer om å gjennomføre et forskningsprosjekt som består i å lage en tagger og lære den hvordan den skal velge den (oftest ene) taggen som er den rette når det er flere å velge mellom. Dette prosjektet vil være en del av et større nasjonalt prosjekt hvor forskere fra andre universiteter skal utvikle andre ting som kan være av stor nytte for dem som kan tenke seg å bruke elektroniske hjelpemidler i sin forskning. Vi kommer tilbake med fyldigere omtale senere.

☆☆

6. Symposium og gjesteforelesninger med korpusvinkling

For dem som er interessert i oversettelse, kontrastive perspektiver og parallelle korpora, er det mye å glede seg over i april. Da arrangerer nemlig Institutt for britiske og amerikanske studier både symposium og gjesteforelesninger innenfor emnet.

Symposiet heter Languages in contrast/Språk i kontrast, og holdes fredag 21. april (kl. 14.00-17.30) og lørdag 22. april (kl. 9.00-15.00) på Blindern, på s.r. 3, brakke 1, før kaffepause og r. 454 P.A. Munchs hus etter. En rekke forskere fra inn- og utland forteller om sine prosjekter. Gjesteforelesningene er ved Antoinette Renouf, University of Liverpool, og holdes 26. april (kl. 9.15-10.00 og 10.15-11.00) på D81 Niels Treschows hus.

☆☆

7. Kort om Tekstlaboratoriet

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. I tillegg må man fylle ut et skjema hvor man lover å bruke tekstene til ikke-kommersielle formål.

Tekstlaben har to ansatte: Janne Bondi Johannessen, lingvist og forsker, og Helge Hauglin, informatiker og ingeniør ved Tekstlaben.

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
Rom B 152, P.A. Munchs hus
Telefon: 22 85 68 14
E-post: jannebj@hedda.uio.no

Helge Hauglin
Rom B 150, P.A. Munchs hus
Telefon: 22 85 78 35
E-post: helgeha@hedda.uio.no