

Βυλλετιν

fra

Tekstlaboratoriet

2. årgang, nummer 2

November 1994

Innhold

1 Gjesteforelesninger i november	1
2 Hvilke tekster og programmer har Tekstlaboratoriet?	2
3 Nye tekster og programmer	3
4 Kurs og demonstrasjoner	5
5 Nytt utstyr	6
6 Annet nytt	6
7 Kort om Tekstlaboratoriet	7

1 Gjesteforelesninger i november

Tekstlaboratoriet har invitert Lou Burnard fra Oxford Text Archive og Willard McCarthy fra University of Toronto til å holde gjesteforelesninger i uke 46. Douglas Biber fra University of Northern Arizona er i Oslo samme uke etter invitasjon fra Lingvistisk Forum. I tillegg til å holde en forelesning for Lingvistisk Forum, vil Biber holde en forelesning arrangert av Tekstlaboratoriet.

Douglas Biber er professor i engelsk lingvistikk ved Northern Arizona University. Han er kjent spesielt for sitt arbeid med flerdimensjonal analyse av tekst. Hans mest kjente bok er: *Variation Across Speech and Writing* (Cambridge University Press, 1988).

Lou Burnard har arbeidet mye med TEI, dvs. Text Encoding Initiative, som er en standard for koding av tekster basert på SGML (Standard Generalized Markup Language). Sammen med Michael Sperberg-McQueen har Burnard redigert "Guidelines for Electronic Text Encoding and Interchange (TEI P3)".

Willard McCarthy har vært spesielt opptatt med å studere litterære tekster vha. PC-programmet TACT, og har skrevet artikkelen "Finding implicit patterns in Ovid's Metamorphosis". McCarthy har vært med på å utvikle TACT.

Program:

Mandag 14/11 kl. 9.15, seminarrom 454, P.A Munchs hus:

Willard McCarthy: *Thinking like a thinking machine: computational modeling in literary research, with an example from Ovid.*

Mandag 14/11 kl. 11.15, seminarrom 454, P.A Munchs hus:

Lou Burnard: *Encoding for Information Interchange: an overview of the TEI Proposals.*

Tirsdag 15/11 kl. 9.15, seminarrom 454, P.A Munchs hus:

Willard McCarthy: *From concordance to textual analysis: applications of TACT to literary and linguistic research.*

Tirsdag 15/11 kl. 11.15, seminarrom 454, P.A Munchs hus:

Lou Burnard: *Introducing SARA: An SGML-Aware Retrieval Application for use with large corpora*

Onsdag 16/11 kl. 9.15, seminarrom 454, P.A Munchs hus:

Douglas Biber: *Cross-linguistic patterns of register variations*

Onsdag 16/11 kl. 11.15, seminarrom 454, P.A Munchs hus:

Lou Burnard: *Introducing The British National Corpus*

The British National Corpus er et samarbeidsprosjekt mellom British Library, forleggere, og universiteter. En samarbeidspartner er Oxford University Computing Service, hvor Burnard arbeider.

Onsdag 16/11 kl. 14.15-16.00 i Lingvistisk Forum, rom 105, Vilhelm Bjerknes' hus:

Douglas Biber: *A Multidimensional approach to historical genre variation*

2 Hvilke tekster og programmer har Tekstlaboratoriet?

I tillegg til de nye tekstene og programmene som annonseres i denne bulletinen, har Tekstlaben en god del materiale fra før. På norsk har vi 60-70 bokmåls- og nynorskromaner og noe sakprosa og avismateriale. Vi har også en god del engelske korpora og et tysk korpus, pluss endel andre tekster, først og fremst tekster på språk med latinske alfabeter. Legg merke til at en god del av tekstene er det bruksbegrensninger på, dvs. at tekstene bare kan brukes til forskningsformål. For å bruke tekstene må man undertegne en bruksavtale med oss hvor man lover at tekstene bare skal brukes til ikke-kommersielle forskningsformål.

De viktigste programmene vi har er Conc for Macintosh og Tact og WordCruncher for PC.

På disken hvor tekstene og programmene er lagret, tekstlabdisken, er det oversiktsfiler

som forteller hva som finnes på disken og hvordan materialet er organisert. Oversiktsfilene har navnet "Lesmeg". Se avsnitt 7 for en beskrivelse av hvordan man kobler seg til tekstlabdisken fra sin egen maskin.

2.1 Annet materiale

På tekstbaser og tekstsentre utenfor UiO finnes det mange tekster. Hvis det er noen som har spesielle ønsker eller behov, kan vi lete på disse. De som er interessert i å lete selv på World Wide Web, kan starte på <http://www.hf.uio.no/tekstlab/pekere.html>. Spesielt kan vi anbefale Oxford Text Archive, Project Gutenberg og The Human-Languages Page.

3 Nye tekster og programmer

3.1 Norsk

Nynorsk frekvensordbok er kjøpt fra Norsk Tekstarkiv, og er lagt opp på tekstlabdisken. Fra Norsk Tekstarkiv har vi også kjøpt "Medmenneske" av Olav Duun, "Trost i taklampa" av Alf Prøysen og fire årganger av Syn og Segn. Se lesmeg-fila på norsk-området for mer informasjon. Syn og Segn er ikke lagt inn på tekstlabdisken ennå.

3.2 Antikke greske og latinske tekster

Tekstlaboratoriet abonnerer på tre CD-plater med gammel gresk og latinsk tekst. Fra Thesaurus Linguae Graecae kommer en CD-plate som inneholder alle greske tekster f.o.m Homer (800 f.K) t.o.m år 600 e.K, pluss historiografiske, leksikografiske og skolastiske tekster fra perioden 600-1453. Denne CD-plata er tilgjengelig i nettet. Fra Packard Humanities Institute har vi enbrukerlisenser på to CD-plater med latinske og greske tekster fra antikken, bl.a papyrer og innskrifter og bibeltekster. Disse to CD-platene har vi ikke lov til å bruke i nettet, så for å bruke disse må man ha en CD-spiller koblet til sin egen maskin.

Vi har kjøpt inn PC-programmet View & Find, som kan brukes til å arbeide med materialet på de tre CD-platene. Med programmet kan man søke i enkelttekster, søke i flere tekster samtidig, man kan lage konkordanser og ordlister, og man kan ta ut tekst for å inkludere i sine egne tekster.

Kurs i View & Find er holdt, og gjentas senere dette semesteret, se eget avsnitt om kurs. Hvis man vil bruke View & Find fra sin egen PC, må programmet installeres på den. Dette tar Helge Hauglin seg av.

Foreløpig har vi ikke fått tak i et tilsvarende søkeprogram for Macintosh, men vi jobber med saken.

3.3 CD-plate fra European Corpus Initiative

CD-plata "Multilingual Corpus 1" fra European Corpus Initiative er kjøpt inn og gjort tilgjengelig i nettet for HF-brukere. Fra Macintosh finnes den som området ECI-MC1 på toppnivå på MacTekstlab. For PC-brukere er den tilgjengelig for montering fra Hedda som `hedda:/hf/hedda/cd/eci-mc1/mnt`. Se fila `G:\info\cd-liste` for mer informasjon, eller ta kontakt med Helge Hauglin.

På CD-plata er det rundt 100 millioner ord på forskjellige språk. Det er først og fremst tekster på europeiske språk, pluss noen andre, bl.a japansk og kinesisk. Fila CORPDESC.TXT på området DOC på CD-plata inneholder en oversikt over materialet.

3.4 Program for morfologisk merking av tysk tekst: GERTWOL

Fra samme firma som vi har kjøpt et program for morfologisk merking av engelsk tekst, ENGCG - English Constraint Grammar Parser, har vi kjøpt et program for morfologisk merking av tysk tekst, GERTWOL - German Two-Level Morphology. GERTWOL vant den første morfolympiaden ved Universitetet i Erlangen-Nürnberg i mars i år som beste morfologiske analyseprogram for tysk. Forskjellen mellom GERTWOL og ENGCG er at GERTWOL ikke har regler for å luke ut flertydigheter. Resultatet av å merke setningen "Sie ist verschwunden" er:

```
"<*sie">
    "sie" * PRON PERS SG3 NOM FEM
    "sie" * PRON PERS SG3 AKK FEM
    "sie" * PRON PERS PL3 NOM
    "sie" * PRON PERS PL3 AKK
    "*sie" PRON PERS HÖFLICH SG/PL NOM
    "*sie" PRON PERS HÖFLICH SG/PL AKK

"<ist">
    "sein" V IND PRS SG3

"<verschunden">
    "ver|schwunden" A(PART) POS
    "ver|schwinden" V PART PERF
```

Programmet presenterer alle mulighetene for det personlige pronomenet, selv om det bare er én som er den riktige i sammenhengen. GERTWOL er et Unix-program. Programmet som sådan er ikke vanskelig å bruke, men Unix-omgivelsen programmet kjører i er nok fremmed for de fleste HF-brukere.

3.5 Statistikkprogram: Minitab for Macintosh

Vi har kjøpt inn en enbrukerversjon av et statistikkprogram for Macintosh, Minitab. Programmet er brukervennlig og menystyrt på vanlig Macintosh-måte. I tillegg er det et eget kommandovindu hvor man kan skrive inn kommandoer. Menyene gir tilgang til de viktigste mulighetene i programmet, mens kommandostyringen gi muligheten til å benytte alle funksjonene i programmet fullt ut.

Av det vi har prøvd av programmet, ser det ut til å være et program beregnet mest for statistikere/realister, idet det er mange avanserte beregninger og tester programmet kan gjøre. Styrken til programmet ligger her. Svakheten er grafikkmulighetene, det er begrensede muligheter til å skreddersy de grafiske presentasjonene av rådata og beregnede verdier.

Programmet er installert på en av maskinene som står på Tekstlaben. Hvis det er noen som er interessert i å prøve programmet, er det bare å stikke innom.

4 Kurs og demonstrasjoner

4.1 Innføringskurs

I første uke av juni gjentok vi innføringskurset i bruk av elektroniske verktøy i språkforskningen. Som tidligere hadde kurset en del for Macintosh-brukere og en del for PC-brukere, hver med en demonstrasjonsdel og en øvelsesdel. Programmene som ble behandlet var Tact og WordCruncher for PC og Conc for Macintosh. Det var 15 deltagere på Macintosh-kurset og 7 deltagere på PC-kurset, og det var tilsammen 4 hovedfagssstudenter på kursene.

Etter kurset ble det foretatt en evaluering, og det kom inn ti svar. Disse ga uttrykk for at kursene stort sett svarte til forventningene deltagerne hadde til et slikt kurs. Deltagerne mente kursene fungerer fint som innføringskurs, og gir forståelse av hvordan datamaskiner kan brukes i språkforskningen. Programmene som kursene tar for seg er av generell karakter, og kan ikke tilfredsstille alle behov. Interessen for å delta på flere kurs senere er derfor stor: åtte av de ti som hadde svart ville gjerne dette. Den typen kurs det er størst interesse for er kurs i mer avanserte og spesialiserte programmer, og kurs som tar for seg praktiske erfaringer med bruk av tekstsøkeprogrammer.

4.2 Kurs i søking i antikke greske og latinske tekster

Den 13. oktober holdt Tekstlaben et kurs i programmet View & Find. Kursinvitasjon var sendt til Klassisk og Romansk Institutt, Filosofisk Institutt, Institutt for Østeuropeiske og Orientaliske studier, Institutt for Kultur- og Samfunnsfag (historie, idéhistorie og religionshistorie) og Teologisk Fakultet. Det var ni deltagere på kurset, hovedvekten kom fra Klassisk og Romansk Institutt, og det var en fra Teologisk Fakultet. View & Find er et PC-program, og det var både PC-brukere og Macintosh-brukere på kurset. Kurset hadde først en demonstrasjonsdel og så en øvelsesdel, på totalt bortimot tre timer. Alt foregikk på øvingsrommet til Anvendt Lingvistik.

Reaksjonene fra deltagerne etter kurset var at kurs i denne typen programmer er de interesserte i, dvs. kurs i programmer spesiallaget for den type tekster de arbeider med. PC-brukerne var ivrige etter å komme i gang på egenhånd etterpå, mens Macintosh-brukerne var litt avventende. Når Tekstlaben får anskaffet et tilsvarende program for Macintosh, vil kurs i dette bli gitt.

Kurset i View & Find vil bli gjentatt den 18. november, kursinnbydelse er sendt alle språkinstituttene. Er det noen som ikke har fått invitasjonen, ta kontakt med Tekstlaben. Denne gangen vil demonstrasjonsdelen foregå i seminarrom 454 i P.A. Munchs hus, øvingsdelen på øvingsrommet til Anvendt Lingvistik. Det blir derfor anledning til å delta på bare demonstrasjonsdelen eller bare øvingsdelen, og det blir mer tid til øving. Som vanlig tilbys individuell oppfølging etter kurset.

4.3 Demonstrasjon av morfologisk merker for tysk

Det morfologiske merkeprogrammet for tysk som Tekstlaben har kjøpt inn og installert (GERTWOL), blir demonstrert på Tekstlaben (B 150) kl. 10.00-11.00 mandag 7. november. Hvis bulletinen ikke når fram før demonstrasjonen til interesserte som ikke har fått invitasjon, ta kontakt med Tekstlaben, så kan vi ordne en demonstrasjon en gang senere.

4.4 Kurstilbudet fremover

I neste semester vil innføringskursene i bruk av elektroniske verktøy i språkforskningen bli holdt på nytt. Dato er ikke fastsatt ennå, kursinnbydelse vil bli sendt språkinstituttene god tid i forveien.

Som meldt tidligere i denne bulletinen, vil kurset i søking i antikke greske og latinske tekster bli holdt igjen 18. november.

Når tekstlabforskeren, Janne Bondi Johannessen, er tilbake fra barselpermisjon over nyttår, vil vi få bedre tid til å se nærmere på endel tekstsamlinger og programmer som vi har kjøpt inn, og vi vil holde kurs i dette.

5 Nytt utstyr

5.1 CD-jukebox

En CD-jukebox er kjøpt inn og for litt siden installert på Hedda. Dette er en CD-spiller med et magasin hvor det er plass til inntil 6 CD-plater, en mekanisme for automatisk skifting av plater, og en avspillingsenhet. CD-plater som ligger i spilleren er tilgjengelige på samme måte som CD-plater som leses fra enplate-spillerne som Tekstlaben har fra før på Hedda. Det er bare en avspillingsenhet i CD-jukeboxen, så det går ikke an å lese fra flere plater på samme tid, dvs. det er den fysiske lesingen som bare kan skje fra en plate om gangen. Det går fint å bruke tekster som ligger på forskjellige CD-plater på samme tid, bare de leses inn og overføres til maskinen de brukes på til forskjellig tid. Hastigheten som platene blir lest med i CD-jukeboxen er den dobbelte av hastigheten på CD-spillerne vi har fra før.

Til enhver tid er det alltid en plate i avspillingsenheten. Når man ber om å få lese en tekst fra en CD-plate kan det hende at CD-plata ikke er inne i avspillingsenheten. Den plata som ligger inne må da skiftes ut med den plata man ønsker å lese. Dette går automatisk, og tar litt over 5 sekunder.

Det gjenstår litt testing for å finne ut hva som skjer når store datamengder forsøkes lest fra forskjellige CD-plater på samme tid, og hvordan PC'er og Macintoshen reagerer på dette.

De CD-platene som blir gjort tilgjengelige fra CD-jukeboxen er de CD-platene som vi regner med blir minst brukt, men som vi likevel vil ha tilgjengelige. Sannsynligheten for at flere prøver å lese fra forskjellige CD-plater på samme tid blir mindre, og sannsynligheten for eventuelle problemer som følge av dette blir også mindre.

5.2 Ny kraftig PC

Universitetet i Oslo har kjøpt inn en stor mengde PC'er som ble brukt under OL på Lillehammer, til bruk for studentformål. Tekstlaben har fått tildelt en slik PC. Det er en IBM PS/2 56, med Intel 486-prosessor, dvs. en ganske kraftig maskin. Det er nå to kraftige PC'er på Tekstlaben.

6 Annet nytt

6.1 Bok om Unix

Nå som Tekstlaben har to Unix-program for merking av tekst (ENGCG for engelsk tekst og GERTWOL for tysk tekst), er det blitt mer aktuelt for tekstlabens brukere å bruke Unix,

dvs. det operativsystemet som kjører på Hedda m.fl. Vi har kjøpt inn en inføringsbok i Unix som spesielt tar opp bruk av Unix til behandling av tekster: "A Course in the Unix Operating System" av Martin Wynne. Boka kan lånes fra Tekstlaben.

En av grunntankene bak Unix er at istedenfor å ha programmer med alle tenkelige muligheter og finesser, har man et godt utvalg av relativt enkle verktøyprogrammer som kan settes sammen til å gjøre avanserte oppgaver. Med noen enkle håndgrep kan man få til ganske mye.

6.2 Tekstlaben på World Wide Web

Tekstlaben har fått sin hjemmeside på nettet, på World Wide Web (WWW). WWW er et nettverk av tjenester, hvor man med noen få tastetrykk eller musklikk kan navigere seg rundt i en overveldende mengde informasjon. Det kan være vanskelig å finne fram, så vi har laget en side med pekere til tekstbaser, tekstsentre og lignende som kan være av interesse for språkfolk. Vi på Tekstlaben bruker denne siden som utgangspunkt for å navigere på nettet, vi håper at andre har glede og nytte av den også. Mosaic er det mest brukte programmet til å bevege seg på WWW. Mosaic går på Macintosh og under Windows på PC. Hvis du trenger hjelp til Mosaic, spør instituttets IT-ansvarlige. Tekstlabens hjemmeside er <http://www.hf.uio.no/tekstlab/velkommen.html>.

7 Kort om Tekstlaboratoriet

Tekstlaboratoriet er et tilbud til språkforskere og hovedfagsstudenter på Historisk-Filosofisk fakultet. Vi har som mål å hjelpe forskere og hovedfagsstudenter til å ta i bruk datateknologi i språkforskningen.

Tekstlaben har to ansatte: Janne Bondi Johannessen som er lingvist og forsker, og Helge Hauglin som er informatiker og ingeniør ved Tekstlaben. Forskeren fikk en sønn i våres, og er i barselpermisjon ut året.

Tekstene og programmene våre er lagret på en disk knyttet til fakultetets filtjener Hedda. Se etter filer med navn "Lesmeg" på disken.

- Slik nås disken fra en Macintosh: Gå inn i eplemenyen og hent fram "Velger". Velg sone HF, tjeneste AppleShare og filserver (sic) MacHedda, klikk på OK-feltet. Skriv inn brukernavn og passord, og klikk på OK-feltet. Velg så MacTekstlab og klikk på OK-feltet. Voilà, tekstlabdisken dukker opp på skrivebordet som MacTekstlab.
- Slik nås disken fra en PC: Med et passende tekstredigeringsprogram 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 tekstlabdisken.

Har du spørsmål til oppkoblingen mot tekstlabdisken, ta kontakt med Tekstlaben eller instituttets IT-ansvarlige.

7.1 Adresser

Tekstlaben holder til i rom B 150 og B 152 i 1. etg. i P.A Munchs hus.

Tekstlaboratoriet HF, UiO, boks 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: j.b.johannessen@ilf.uio.no

Helge Hauglin

Rom B 150, P.A. Munchs hus

Telefon: 22 85 78 35

E-post: helge.hauglin@ilf.uio.no
