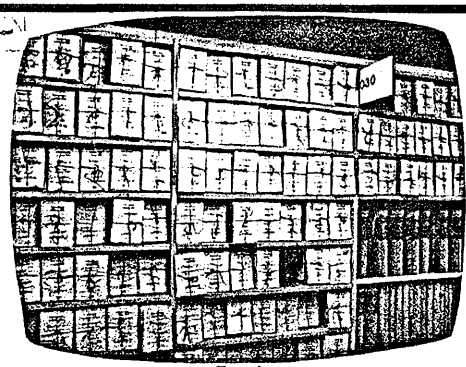
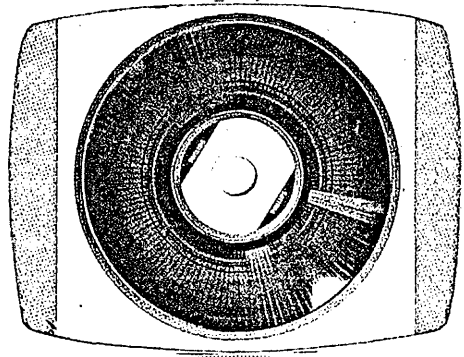
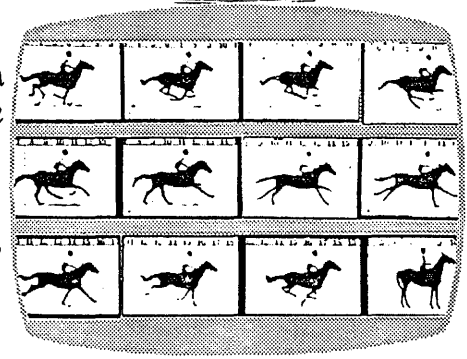
humanistiske data 3-83

VIDEOPLATE
for
lagring
av
tale, tekst og bilde





nesten ubegrensete muligheter for arkivering



NAVF

ARTIKLER
RAPPORTER
MELDINGER
SUMMARY

NAVFs EDB-senter for humanistisk forskning

The Norwegian Computing Centre for the Humanities

SENTERETS RAPPORTSERIE

RAPPORT nr. 1. *EDB i gjenstandsfagene.* Rapport fra en konferanse i Bergen, 18. og 19. april 1978. September 1978. 2. opptrykk november 1981. ISBN-82-7283-022-1 Pris kr. 40.

RAPPORT nr. 2. *Et norsk datamaskinelt tekstkorpus.* Rapport fra en konferanse i Bergen, 19. og 20. oktober 1978. Februar 1979. 2. opptrykk 1981. ISBN 82-7283-016-7 Pris kr. 20.

RAPPORT nr. 3. *Rapport fra den nasjonale konferanse om EDB i språk og litteraturforskning*, 4. og 5. januar 1979. Mars 1979. 2. opptrykk november 1981. ISBN 82-7283-024-8 Pris kr. 50.

RAPPORT nr. 4. *Oppbygging av EDB-katalog for folkemusea i Hordaland og kulturgeografisk registrering på Vestlandet.* April 1978. 3. opptrykk november 1981. ISBN 82-7283-000-0 Pris kr. 30.

RAPPORT nr. 5. *Rapport fra NKKM's EDB-komite.* August 1979. ISBN 82-7283-001-9 Pris kr. 15.

RAPPORT nr. 6. *Prøveprosjekt med EDB ved Norsk Folkemuseum.* Oktober 1979. ISBN 82-7283-002-7 Pris kr. 15.

RAPPORT nr. 7. *Ivar Fonnes: Norsk landbruksordbok.* Prosjektrapport om databehandling og tilrettelegging for trykking. September 1979. ISBN 82-7283-008-6 Pris kr. 25.

RAPPORT nr. 8. *SEFRAK. Rapport frå prøveprosjekt for databehandling av kulturminneregisteret.* Oktober 1979. ISBN 82-7283-003-5 Pris kr. 30.

RAPPORT nr. 9. *Jostein H. Hauge og Sigbjørn Århus: Dataregistrering i humanistiske fag med vekt på optisk lesing.* August 1978. 3. opptrykk januar 1981. ISBN 82-7283-004-3 Utsolgt.

RAPPORT nr. 10. *Roald Skarsten: Innføring i SPSS for humanister.* November 1977. 3. opptrykk november 1981. ISBN 82-7283-005-1 Pris kr. 30.

RAPPORT nr. 11. *Jostein H. Hauge og Knut Hofland: Rapport fra 4 konferanser i USA sommeren 1979.* The 17th Annual Meeting of Computational Linguistics. La Jolla Conference on Cognitive Science. The fourth International Conference on Computers in the Humanities. Data Bases in the Humanities and Social Science. November 1979. ISBN 82-7283-007- 8 Utsolgt.

RAPPORT nr. 12. *EDB og manuskriptregistraturer.* Oktober 1977. 2. opptrykk november 1979. ISBN 82-7283-009-4 Pris kr. 20.

RAPPORT nr. 13. *Datatjenester for og datasamarbeid mellom kunst og kulturhistoriske museer.* Februar 1980. 2. opptrykk november 1981. ISBN 82-7283-010-8 Pris kr. 50.

Forts. 3. omslagsside.

humanistiske data 3 -83

NAVFs EDB-senter for
humanistisk forskning

The Norwegian Computing
Centre for the Humanities

NAVF NORGES
ALMENVITENSKAPELIGE
FORSKNINGSRÅD

NAVFs EDB-senter for humanistisk forskning ble opprettet av Norges almenvitenskapelige forskningsråd i 1972. Senteret har som oppgave å arbeide på nasjonal basis for utbredelse av edb i forskningsarbeidet i de humanistiske fagene. Det er opprettet en samarbeidsavtale med Universitetet i Bergen som bl.a. gir Senteret adgang til edb-tjenester ved Universitetet.

Av sentrale oppgaver kan nevnes utvikling av programutrustning for humanistiske forskningsoppgaver, konsulenthjelp og informasjonstjenester.

Senteret utgir tidsskriftet Humanistiske Data (3 nr. pr. år) og en rapportserie (32 er utkommet pr. 1.11.83).

Senteret er sekretariat for International Computer Archive of Modern English (ICAME), og utgir bladet ICAME NEWS.

Senteret driver egne opplæringsprogram for vitenskapelig personale og medarbeidere i den kontor-tekniske gruppen innenfor de humanistiske fag. Det blir også holdt forskjellige kurs og seminar om edb og humanistisk forskning. Tidspunkt og emner blir kunngjort i Humanistiske Data og på institusjonene.

Interesserte kan kostnadsfritt bestille årsmelding og Humanistiske Data (kr. 50,- for institusjoner).

Humanistiske Data blir utgitt av NAVFs EDB-senter for humanistisk forskning. Redaksjonsgruppe: Jostein H. Hauge (ansv.), Rune Johansen, Kristin Natvig, Elin Solstrand.

Senterets adresse: Harald Hårfagresgt. 31, Boks 53, 5014 Bergen-Universitetet. Tlf. (05) 320040, linje 2956.

Artikler, rapporter, meldinger mottas. Redaksjonen avsluttet 1. november.

Humanistiske Data is published by The Norwegian Computing Centre for the Humanities. Editorial group: Jostein H. Hauge, Rune Johansen, Kristin Natvig, Elin Solstrand.

The journal can be ordered from the address mentioned above. Contributions are welcome.

Medarbeidere i dette nummer:
Tone Bratteteig, forskningsstipendiat, Universitetet i Oslo
Tove Fjeldvig, forskningsstipendiat, Universitetet i Oslo
Anne Golden, cand.philol., Universitetet i Oslo
Lars Otto Grundt, professor, Universitetet i Bergen
Stig Johansson, professor, Universitetet i Oslo
Elisabeth Johnsen, avd.leder, NAVF
Aagot Landfald, førstekonsulent, Norsk språkråd
Eirik Lien, konsulent, Universitetet i Trondheim
Bjarne Norevik, førstekonsulent, Norsk Termbank
Gunnar Thorvaldsen, daglig leder, Registreringssentral for historiske data
Fra Senteret: *Jostein H. Hauge, Rune Johansen, Ole Lauvskar, Kristin Natvig, Øystein Reigem, Elin Solstrand.*

Fotosats i kommunikasjon med Univac 1100/82
Sats: Universitetet i Bergen/NAVFs EDB-senter for humanistisk forskning.
Grafisk design og montasje: NAVFs EDB-senter for humanistisk forskning
Trykk: Nortrykk a/s
Forsideillustrasjon: Rune Johansen

Økonomiske data

Humanistiske Data har vokst - og det har også utgiftene våre til produksjon og distribusjon av bladet. Derfor ser vi oss dessverre nødt til å be om følgende for 1984:

Institusjoner betaler **kr 50** i årsabonnement (inkl. Senterets årsmelding). Fra og med nr. **1-84** vil institusjoner automatisk få tilsendt **1** eksemplar av hvert nummer - ønskes det flere eksemplarer, kryss av på kupongen nedenfor.

Til alle som **ikke** ønsker å motta HD lenger: vennligst kryss av på kupongen nedenfor.

Abonnementet kan innbetales til vår postgirokonto 3 38 45 67 eller bankkonto 3625.88.53657.

På forhånd takk for hjelpen!

☐ Ønsker i alt ..... eks. av HD
☐ Ønsker ikke å motta HD

Navn: _____

Adresse: _____ _____

_____

POSTKORT

Frimerke

TIL

NAVFs EDB-senter for humanistisk forskning
Postboks 53
N-5014 Bergen-Universitetet
NORGE

Innhold

Artikler:

Rapporter:

Videoplateteknikk

Elin Solstrand

Denne artikkelen gir en innføring i hva videoplateteknikk er og hvordan den kan anvendes. Vi vil også referere til konkrete prosjekter der denne lagringsteknikken er tatt i bruk i utlandet.

Videoplateteknikk er ny teknikk, selv om diskusjoner om den startet i fagpressen for minst sju år siden, lenge før teknologien ble introdusert kommersielt i USA i 1979. Teknikken er tatt i bruk i såvel hobby/hjemme-markedet som i mer profesjonelle anvendelser i arbeidslivet. Et eksempel på førstnevnte er de såkalte «compact-discs» for lagring av musikk, som forholdsvis nylig ble lansert i Norge. Vi vil imidlertid konsentrere oss om anvendelser utenfor underholdningsindustrien, selv om det også kan være interessant å skjele til hva som skjer der. Suksess på underholdningsmarkedet betyr sannsynligvis at en har klart å frembringe et anvendbart produkt til en overkommelig pris.

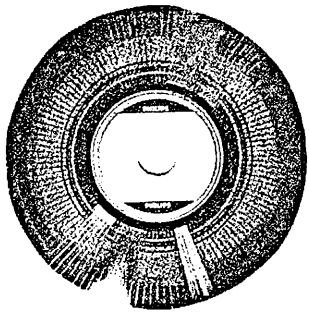
En vanlig innvending mot denne nye teknologien har nemlig vært at de produkter som bygger på den er for dyre. Dette er nå i ferd med å endres - ifølge tall fra de selskapene som frembyr sine produkter. De hevder at kostnadene for lagring av data på optiske videoplater er klart lavere enn på de andre lagringsmediene vi kjenner (magnetplate, diskett, magnetbånd etc.).

Kapasitet og kostnad

Hva er det som gjør denne nye teknologien så forlokkende? Først og fremst er det videoplatens kapasitet. En videoplate kan lagre langt mer data enn et magnetbånd, og i tillegg er informasjonen som oftest direkte tilgjengelig. (Se figur a). Det er vanskelig å oppgi tall for hvor mye som kan lagres på en videoplate - det er avhengig av benyttet teknikk og av type informasjon en ønsker å lagre. I tillegg går utviklingen så fort at de tallene en oppgir raskt foreldes. Det er også stor forskjell på den kapasitet en oppnår i store, «tunge», dyre konfigurasjoner (ofte ennå på prototyp-stadiet) og på det som oppnås på rimelige systemer. Således rapporteres det om et system fra RCA (kostnad $ 500 000 som kan lagre 12 Gbyte pr. side. Philips kan tilby en plate som lagrer 1 Gbyte pr. side. (1 Gbyte tilsvarer ca. 54 000 stillbilder i farger eller 500 000 A4-sider tekst.) Denne kan avspilles på en laser-spiller til omlag 6000 kroner. Programvare for søking kommer i tillegg. Produksjonen av en plate («master») koster kr. 15 000. Kopier av denne kan fås svært rimelig, noen hundre kr. pr. stk. LaserData, et amerikansk firma, markedsfører et system som kan lagre en million tekstsider på en enkeltsidig plate. Firmaet estimerer lagringskostnadene til å være mindre enn 2 cents pr. million tegn, mot $ 4 på magnetplate og $2 på

diskett. Systemet består av en kontroller/mikromaskin og en spiller (pris omlag $ 5000). Platen har en kapasitet på 4.8 Gbyte, og produksjon av en «master»-plate samt 10 kopier koster $ 15 000. Øvrige kopier koster omlag $100. Hvis det er 300 brukere pr. plate vil kostnadene pr. bruker bli omlag $ 150.

Figur a.



20.000 bøker

eller

54.000 bilder

eller

30 magnetbånd

Holdbarhet og kvalitet

Et annet viktig moment er den optiske videoplatens holdbarhet og kvalitet. De fleste produsenter garanterer platene i 10 år. Det er ventet at en snart vil kunne garantere en enda høyere levetid - 30 år. Magnetbånd garanteres i dag for 2 år.

En optisk plate vil ha samme kvalitet uansett spilletid, i motsetning til magnetbånd og diskett der det magnetiserbare belegget slites ved bruk. Videoplateindustrien har imidlertid slitt med en del barnesykdommer. Vanskelighetene ligger bl.a. i produksjon av feilfrie plater. Produksjonsprosessen krever et absolutt støvfritt miljø for å kunne brenne ut de ørsmå hullene (i platen) som inneholder informasjonen. Dette vil vi komme nærmere inn på senere. Suksessen for denne teknikken står og faller mye på om en klarer å utvikle tilfredsstillende prosedyrer for plateframstilling. Det er imidlertid få som betviler at det vil være mulig.

Minus-sider?

Hittil har de optiske platene «bare» hatt én vesentlig minus-side - de har ikke vært mulige å skrive over («write-once-read-only»). En kan selvsagt tenke seg bruk der dette er en fordel (f.eks. i arkiver), men det må vel generelt ses på som en begrensning ved teknikken idet bruken begrenses til statiske data som ikke må oppdateres for ofte. Oppdatering på videoplate må i dag skje ved at en produserer en ny plate, eller at en skriver videre på den gamle (men merker gamle data som

slettet). Denne siste muligheten har en i dag bare i de mer avanserte systemene. Den vanlige produksjonsteknikken for videoplater er nemlig å presse kopier fra en «master»-plate. Disse kopiene kan en ikke skrive videre på. I det aller siste er det imidlertid annonsert fra flere kanter, hovedsakelig fra amerikanske og japanske firma, at den «overskrivbare» optiske videoplaten kommer. For å få dette til, har en benyttet seg av en helt annen teknikk ved skriving på platen.

Bakgrunn

Det er nå på sin plass med en nærmere forklaring av teknikken som gjør det mulig å lage plater med ovennevnte egenskaper. Videoplaten ble utviklet med tanke på bruk i underholdningsindustrien. Den skulle først og fremst være bildets grammofonplate. På denne måten håpet industrien å sikre at copyright-lovene ikke ble neglisjert på samme måte som ved bruk av magnetbåndteknikk (både for lyd og bilde). I og med at både grammofonplaten og videoplaten ikke kan skrives over, vil en ha fullstendig oversikt over bruken.

Men av forskjellige grunner gikk det ikke som industrien håpet. Videoplatespillere er fremdeles dyrere enn videobåndspillere, de er mindre robuste, og sist men ikke minst, programtilbudet er for dårlig. Mange spår imidlertid videoplaten en lysende framtid i videospillbransjen. En kan tenke seg en kombinasjon av film, stillbilder og tale/musikk, pluss at det legges inn valgmuligheter slik at spilleren selv kan ha innflytelse på spillets utgang.

En oppdaget fort at videoplateteknikken også hadde andre anvendelser. I tillegg til lagring av bilder egnet den seg også til lagring av tekst og lyd. Listen over mulige anvendelsesområder økte: lagring av musikk, kontorautomatisering, elektronisk publisering, databaselagring, undervisningshjelpemiddel etc.

Analog eller digital lagring

De første videoplatene var analoge. Det ser imidlertid ut som om trenden går fra analoge til digitale plater.[1] Det snakkes også om utstyr som vil kombinere begge teknikker alt etter hvilken type informasjon som behandles. Det er uten tvil enkelte typer informasjon som egner seg best til digital gjengiving (tekst og tall) og enkelte som egner seg bedre for analog representasjon (bilde, film og lyd). Digital representasjon har imidlertid mange fordeler framfor analog representasjon. Den gir lav forvrengning, noe som gir god gjengiving av signalet. Den gir også gode muligheter for feilkorrigering, noe som gjør data mindre utsatt for støy både ved lagring og overføring. Viderebehandling av tekst og tall *krever* data på digital form, men også bilde og lyd kan behandles bedre digitalt. Data er kort sagt lettere å ha med å gjøre i digital form.

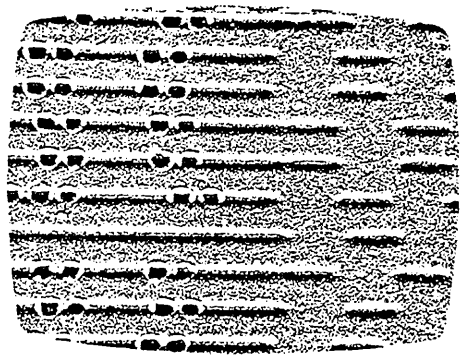
Både analoge og digitale data kan lagres med noenlunde samme teknikk. Den mest interessante og også mest vanlige teknikk som brukes

(ihvertfall av dem som har produkter som er beregnet for et «seriøst» marked), kalles ofte optisk laserlagring. I tillegg finnes det andre teknikker som minner mer om vanlig grammofonplateteknikk. I det følgende beskrives den mest interessante teknikken for våre formål.

Optisk laserlagring

Data blir her lagret som ørsmå hull i en tynn metallfolie (ofte tellurium). (Se figur b). Hvert hull er mindre enn 1 mikron (tusendels millimeter) i diameter. Hullene blir brent inn av en laser. For analog informasjon, vil hullene variere i lengde «i takt med» det analoge signalet. (Dvs. at signalet blir *frekvensmodulert* og hullene får lengder som varierer med bølgelengden til enhver tid. Se figur c.) Hvis vi har digital informasjon, vil hullene være omtrent sirkelformete. Vi skriver bitene ved å brenne et hull for hver ener. Via en laser kan vi også lese det som er skrevet på platen. Laserlyset vil reflekteres i varierende grad (derav navnet optisk laser-lagring) alt ettersom det treffer et hull eller ikke. Dette registreres av en detektor. (Se figur d.) Platen er dekket med et gjennomsiktig materiale, som oftest plast. Platen tåler dermed røff behandling, riper og fingermerker hindrer ikke laserlyset i å trenge igjennom. Disse platene kan man ikke skrive over. Hullene i metallfolien lar seg ikke lappe.

Figur b.

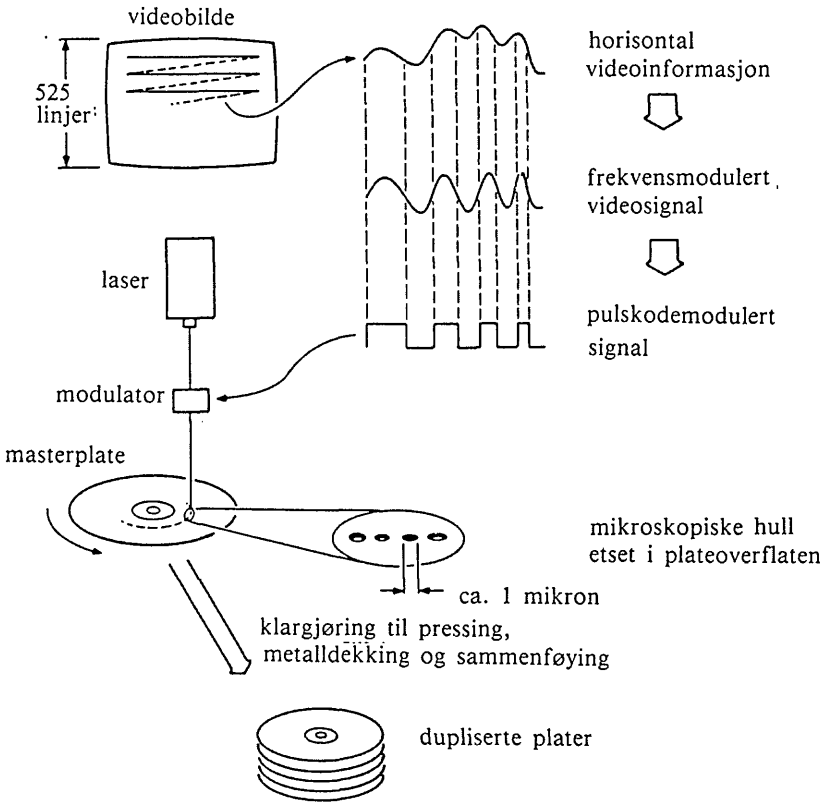


Følgende karakteristika er typiske for en digital optisk laserplate: Det kan som oftest skrives på begge sider, men platen må snus for at en skal kunne lese baksiden. Det franske firmaet Thompson CSF har imidlertid laget en gjennomsiktig plate der begge sidene kan avleses uten snuing.

Pregingen (brenning av hull) starter innerst på platen og følger en spiral utover (1-4 mikron bred). Antall spor (dvs. «omdreininger») på platen er rundt 30-40 000. Sporene er delt inn i adresserbare sektorer.

Platen er på størrelse med en vanlig grammofonplate, både når det gjelder diameter og tykkelse. Den roterer med en fast hastighet, oftest 900 eller 1800 omdreininger i minuttet. Gjennomsnittlig aksesstid er lav - et typisk eksempel er Philips' DOR-plate med 135 ms.

Figur c.

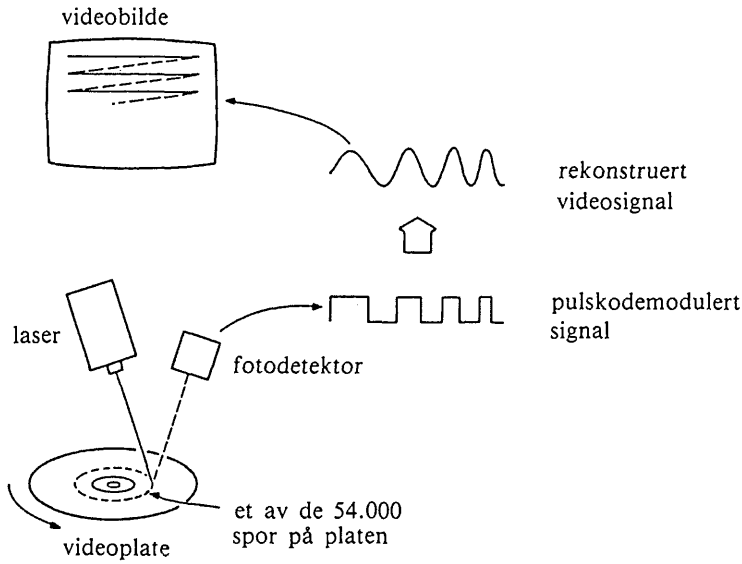


For å sikre at data blir skrevet korrekt, kan feilsjekking gjøres i skriveøyeblikket. Teknikken kalles DRAW (direct-read-after-write). Når en bit er skrevet, blir den øyeblikkelig lest igjen. Hvis en feil er oppstått, merkes biten som slettet på platen og en gjør et nytt forsøk.

Data kan leses med en fart på minst 2 Mbit/s, dvs. 250 Kbyte/s. Ved overføring av film trengs imidlertid en hastighet på 80 Mbit/s. Men utviklingen går fort - det rapporteres allerede nå om et eksperimentelt system fra RCA som kan klare 50 Mbit/s. Dessuten ville bare omlag 10 sekunder film fått plass på dagens digitale plate. Film må således lagres analogt i dag. En analog videoplate kan lagre en vanlig spillefilm.

Følgende firma lager det vi har valgt å kalle optiske laserplater: Philips, Thompson/CSF, RCA, DiscoVision Association (IBM/MCA), Sony, Matsushita, Hitachi og Toshiba. Listen er ikke uttømmende.

Figur d.



Mangfoldiggjøring

Optiske laserplater mangfoldiggjøres enten ved laserskriving eller mekanisk pressing. Hvis en skal masseprodusere platene, lønner det seg å lage en «master»-plate som kan brukes til å lage en presseplate. På denne måten kan en tenke seg at leksika, telefonkataloger, spillefilmer etc. kan produseres i stort antall.

«Master»-platen produseres ved at en laser lager et mønster på en roterende glassplate dekket med en fotosensitiv oppløsning (emulsjon). (Se figur c.) Etter fremkalling dekkes platen med sølv, og en foretar feilsjekking. Hvis platens kvalitet er god nok, blir den dekket med nikkel og aluminium for å kunne skille en negativ-kopi fra «master»-platen. Denne negativ-kopien blir så brukt til å lage en positiv-kopi. Til slutt fås en plate som kan brukes i pressingen.

Laserskriving, dvs. at hvert eksemplar skrives for seg, er ideelt til mindre opplag. Teknikken vil kunne brukes til anvendelser som masselagring, dataarkivering, programarkivering og dokumentlagring. Det antydes at et antall på 100 eksemplarer vil være skillet mellom lønnsomheten for de to reproduseringsmetodene.

Overskrivbare plater i 1985?

Som før nevnt er det «overskrivbare» plater under utvikling. En aktuell teknikk er den såkalt magneto-optiske. Platene er da magnetiserbare. 3M beskriver sin teknikk slik (Electronics, 14 July 1983): Skriving gjøres med en diodelaser[2] som varmer opp et 1 mikrometer-diameter punkt til 150 grader C. Ved denne temperaturen klarer et ytre magnetisk felt å endre magnetiseringsretningen – en enerbit blir lagret i én retning, en null i den andre. For å lese data blir polarisert lys fra en annen laser fokusert på hver bit. I det reflekterte lyset blir polariseringen endret avhengig av magnetiseringen. Utfordringen ligger i å finne et materiale som gir størst mulig forskjell i polariseringen samtidig som det reflekterer godt, slik at feilprosenten ved lesing blir lav. Ekspertene snakker om en nedre grense for en «carrier-to-noise ratio» på 45 decibel for at teknikken skal være kommersielt anvendbar. 3M hevder at de nå har et materiale som gir et så godt mål at produktet kan settes i produksjon. Platene kan lagre 1.5 Gbyte pr. side, omtrent det samme som deres vanlige optiske plate. Det er ikke oppgitt hva prisen for denne nye platen vil bli.

Philips arbeider også med overskrivbare laserplater. Deres produkt kan imidlertid bare lagre en brøkdel av hva de permanente platene kan ta. Det er snakk om et maksimum på 200 Mbyte når platen er ferdig utviklet. De skal imidlertid være billige. Hvis den prototypen de har nå ble introdusert kommersielt, ville den kunne selges for omlag 2-3000 kroner (Mini-Micro Systems, august 1983.) Platene er aktuelle som et alternativ til diskett, de vil nemlig kunne lagre minst 30 ganger så mye, og er prismessig konkurransedyktige. Teknikken som brukes, ligner den vi har beskrevet tidligere, altså at oppvarming gjør det mulig å skrive på platen. Dette har negativ effekt på dataoverføringshastigheten. Etter at en bit er skrevet er det nødvendig at den får kjølne før den neste skrives. Dette problemet kan en løse på to måter, enten ved at lese/skrive-hodet forbedres eller ved at en bruker to skriveenheter, men i første omgang er det siste uaktuelt pga. at det blir for dyrt.

Mange tror at de optiske platene vi har i dag (de som ikke kan skrives over) alltid vil være attraktive for arkiveringsformål. Det er ting som tyder på at det ikke kan oppnås tilsvarende lagringstetthet med overskrivbare plater, og i tillegg er det jo nettopp et ønske om å bevare som får oss til å arkivere dokumenter. Data lagret på optisk videoplate kan ikke ødelegges med magneter eller usynlige programmer.

Andre selskaper annonserer overskrivbare plater med høyere lagringstetthet enn Philips' plate (3M 1.5 Gbyte, Matsushita 0.7 Gbyte og Sony 1.0 Gbyte), men disse er mye dyrere.

Andre teknikker

Det finnes systemer som bruker en slags stift til å avlese den informasjonen som er lagret på videoplaten («grooved capacitance»). På grunn

av at stiften faktisk er i kontakt med videoplaten, er disse platene utsatt for stor slitasje, samtidig som en mister mange av fordelene ved optisk laserlagring, bl.a. direkte aksess til data. Det er derfor lite trolig at systemer som bruker denne type teknikk vil være egnet til bruk utenfor underholdningsindustrien. Dens største fortrinn er at den gir billige produkter.

Teknikken er imidlertid i stadig utvikling. Det rapporteres fra JVC at de har utviklet en «stift» som ikke er i kontakt med platen, men som flyter på en luftpute og oppfatter elektriske signaler fra platen.

Anvendelsesområder

Videoplateteknologien har fått både potensielle brukere, fagpresse og produsenter til å begeistres. Brukerne øyner sjansen til å få løst sine informasjonsproblemer, og produsentene ser for seg et milliardmarked. Fram til i dag har det imidlertid vært «mye skrik og lite ull». Men nå ser det ut til å løsne. De fleste teknologiske problemene er overvunnet, og flere firma har nå den nødvendige programvare. Hvor kan teknikken så best gjøre nytte for seg?

På en optisk videoplate kan en tenke seg at både film, tale, tekst, stillbilder, grafikk og andre digitale data kan lagres i et eneste massivt elektronisk arkiv. Dette gir store muligheter både innen undervisning, publisering og dokumentlagring. For å kunne utnytte informasjonen best mulig, er det imidlertid nødvendig at videospilleren er koblet til en kontroller/mikromaskin med intelligent programvare for søking. I en primitiv applikasjon uten slik programvare kan en tenke seg at de lagrede data gjenfinnes ved at en manuelt taster inn nummeret til det bildet/dokumentet en ønsker å hente fram. I mer sofistikerte anvendelser må en ha bedre hjelpemidler. En kan tenke seg et internt databasesystem skreddersydd til videospiller og kontroller. Dette systemet ville ha de samme karakteristika som et vanlig edb-basert gjenfinningssystem, og språket ville i prinsippet også være identisk.

Indeks-informasjon (tesaurus) kan en tenke seg lagret sammen med den andre informasjonen på platen. Hver informasjonsenhet på platen må således merkes, fordi den kan inneholde alt fra tekst- og bildeinformasjon til tesaurusinformasjon og programmer for styring av videospilleren. På output-siden vil videospilleren være koplet til en TV-skjerm, hi-fi utstyr, skriver, telelinje eller datamaskin, alt etter hva som er lagret og hvordan en ønsker å utnytte det.

Optisk laserlagring vil bety svært mye for alle som er avhengig av å lagre og å ha rask tilgang til store mengder informasjon. *Arkivinstitusjoner* av alle slag, og også større organisasjoner, har i dag store problemer med å oppbevare data, enten de finnes i form av arkivmapper eller magnetbånd. Det koster å oppbevare informasjon på kilometervis av hyller, og i tillegg er den vanskelig tilgjengelig. I USA brukte det offentlige 2 millioner magnetbånd bare i 1975. Forsvaret og romfarts-

organisasjonen (NASA) har spesielt store problemer. Den kolossale mengde data som satelittoverføres fra rommet, sprenger alle rammer for nåværende teknologi. De to nevnte organisasjoner har gitt store midler til bl.a. MIT for å få utviklet systemer der videoplateteknologi inngår.

Videoplaten kan bli nyttig for *edb-sentre* som trenger et trygt sted å oppbevare data og programmer. Det er også behov for jevnlig å ta sikkerhetskopi av hele systemet («full-save»). Dette gjøres nå vanligvis på magnetbånd. Ulempene med magnetbånd er selvsagt at data tar stor plass, har kort holdbarhetstid, og at de er vanskelig aksesserbare.

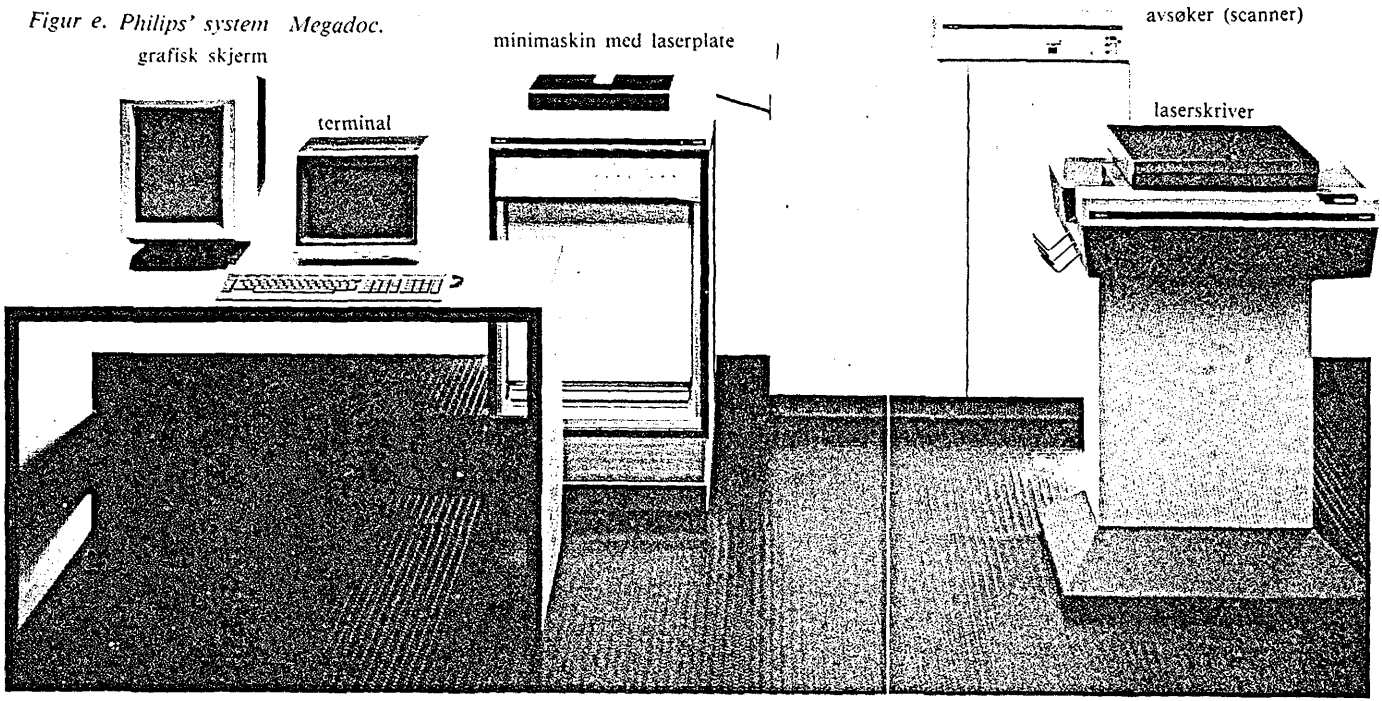
En eneste videoplate kan som tidligere nevnt lagre minst 500 000 A4-sider på hver side (f.eks. Philips). LaserData selger en plate som kan lagre det doble. (Fremtidige systemer vil ganske sikkert doble dette igjen.) Det medfører at innholdet av 60 tettpakkete magnetbånd kan lagres på en dobbeltsidig optisk laserplate. Som en illustrasjon sier LaserData at det vil ta over et år å sende innholdet av en plateside over en 24-timers telefonlinje. For alle som har plassproblemer, må dette gi eventyrlige perspektiver.

Termen *kontorautomatisering* kan gis et helt annet innhold i disse dager. De nye systemene fra f.eks. Philips eller Toshiba består av en «scanner», en laserskriver, en datamaskin og en videoplatespiller. (Se figur e.) Input-delen virker som en kopi-maskin. Dokumentet du ønsker å arkivere leses («scannes») i løpet av noen sekunder. Deretter lagres informasjonen på en laserplate. Hvis du ønsker å gjenfinne dokumentet, kan du få det fram på en skjerm eller du kan få en papirkopi ved hjelp av laserskriveren. Systemet kan altså lagre et hvilket som helst dokument, tekst eller bilde i svart-hvitt (ingen gråtoner). (Merk at dokumenter som scannes blir liggende som *bilder*, selv om de inneholder tekst. Det er ingen optisk lesing av de enkelte tegnene (OCR). Vanlig tekst som skal behandles eller søkes i, må legges inn via tastatur.) Disse systemene koster fra 300 000 kr og oppover, avhengig av hvor mye som kan lagres, og faktorer som aksess-tid o.l. Det hevdes at denne prisen absolutt kan konkurrere med tilsvarende systemer, f.eks. de som bruker mikrofilm.

Termen *elektronisk publisering* høres stadig oftere. Med det menes at publikasjoner av alle slag ikke produseres på papir, men på et elektronisk medium. En eneste videoplate vil uten problemer kunne lagre en boksamling på 20 000 bind. Fordelene ved å oppbevare bøkene på denne måten er at det tar lite plass og at det er billig. Ulempen blir at det er lite hyggelig og noe trøttende å lese fra en TV-skjerm istedenfor en bok. Mange tviler derfor på at den trykte boka helt kan erstattes av et elektronisk medium. Det er derimot enighet om at oppslagsbøker med fordel kan lagres elektronisk. Likeledes egner videoplaten seg godt til *oppbevaring* av boklig informasjon.

Videoplateteknikken gir også nye muligheter for det en ofte kaller «on-demand» publisering, dvs. at publikasjoner ikke trykkes før de er

Figur e. Philips' system Megadoc.



bestilt. På denne måten unngår en at store opplag blir liggende ubrukt. Publikasjonene hentes fram fra en videoplate, skrives med laserskriver og postes, evt. sendes over telelinje til brukeren.

En applikasjon som ofte nevnes i forbindelse med videoplateteknologi, er lagring av *leksika/oppslagsbøker*. Som vi alle vet er det dyrt å kjøpe leksika, de blir også fort for «gamle». Den som vil følge med savner en «oppdatert» versjon etter få år. Et av de største verk av denne type er The Encyclopædia Brittanica, med 23 bind som inneholder omlag 20 000 figurer/bilder. På en Philips DOR-plate som tar 2 Gbytes pr. side, vil hele verket kun ta 36 prosent av plassen. Dette er ved en konvensjonell presentasjon med tekst og stillbilder i svart og hvitt. Det er imidlertid ikke grenser for hvilke presentasjonsmåter en kan tenke seg - en kombinasjon av tale, film og figurer ville f.eks. være mer interessant. I USA arbeides det nettopp med et slikt leksikon. Et leksikon til bruk i den videregående skole og for universitetsstudenter overføres til videoplate. Det brukes både tekst, billedmateriale og filmsekvenser. Det planlegges å gjøre hele verket tilgjengelig over et to-veis interaktivt kabel-TV-nett som når seks amerikanske byer. Kostnadene for brukerne antas å bli på bare halvparten av hva en trykt utgave ville koste (Electronic Publishing Review, nr.1 1982 s.73).

Ordbøker kan forbedres ved at de automatiseres, dvs. at alle oppslagsordene kan framsøkes og vises automatisk. Dette gir mange fordeler i forhold til ordbøker i bokform. Oppslag blir lettere, raskere og kan gjøres mer avanserte. I tillegg kan en tenke seg nye funksjoner som stave-korrigering, synonym- og antonym-søking, og uttale i lydform. I framtiden kan slike ordbøker tenkes brukt til grammatikkkontroll og setningsbygging. Forskjellige spill, f.eks. anagramlaging (brukeren får oppgitt en liste med bokstaver og skal lage så mange ord som mulig av dem), ord- og bildegjetting (brukeren får oppgitt en definisjon eller et bilde og må si hva det forestiller) kan bli biprodukter. Videoplaten gjør det mulig og økonomisk forsvarlig å satse på slike applikasjoner.

Telefonkatalogen er kanskje vår mest brukte oppslagsbok. Produksjon og distribusjon av trykte papirutgaver levert hvert år er imidlertid både dyrt og arbeidskrevende, noe vi forbrukere har merket ved at vi ikke lenger automatisk får tilsendt en katalog pr. telefonapparat. Franskmennene er kommet lengst i forsøk med en elektronisk telefonkatalog. Abonnentene i flere større byer har fått utlevert en terminal istedenfor en papirkatalog. Denne bruker de til søking i en sentral database som jevnlig oppdateres. Databasen ligger på konvensjonelt masselager.

Hvis videoplaten gjør sitt inntog kan det ha innvirkning på flere måter. For det første kan den sentrale databaseoperatør forbedre sine tjenester ved å installere videoplateutstyr istedenfor magnetiske plater. En annen mulighet er at hver enkelt forbruker får sin egen plate hvert år, istedenfor en trykt katalog. Dette siste ville imidlertid avhenge av

om videoplatespilleren blir allemannseie. Uten masseproduksjon vil produktene bli for dyre for den vanlige forbruker. Det ville også avhenge av i hvor høy grad standardisering kan oppnås. Det bør bli minst like god standardisering som det vi har for vanlige videobåndspillere i dag. Og sist men ikke minst, vil brukerne godta denne nye formen for oppslagsbok? Det vil avhenge av flere ting: grad av brukervennlighet, kvalitet og kostnad. Hvis denne nye måten å slå opp på er lettere, bedre og billigere enn den gamle, vil publikum trolig godta den nye teknologien. (Er de ikke nødt til det?)

En annen type publikasjon som krever stadig oppdatering er *rutetabeller* av forskjellig slag. Også her er distribusjons- og trykkingsomkostningene store. Store materialkostnader (spesielt papir) kan innspares ved at tabellene lagres elektronisk.

Videoplaten vil sannsynligvis ha stor innflytelse på *on-line* informasjonsgjenfinning. Her i Norge er NSI databaseoperatør for databaser av denne type, og utenlands finnes de i hundrevis - av de mest kjente er kanskje LEXIS, MEDLINE og basene som nås gjennom Dialog. Disse databasene inneholder opplysninger om spesielle emner (f.eks. medisin), ofte i bibliografisk form (tittel på publikasjon, forfatter, utgivelsesopplysninger m.m.). Andre ganger er det såkalte faktadatabaser en har med å gjøre.

Det som karakteriserer disse databasetjenestene er at de forholdsvis ofte oppdateres (sentralt), de gir mulighet for en on-line bestilling av dokumenter (men off-line levering) og kostnadene deles av en stor brukergruppe. For brukerne er en ikke uvesentlig del av kostnadene telekommunikasjonsutgiftene. Derfor kan en faktisk tenke seg at databaser på videoplater blir spredt rundt til de store brukerne (f.eks. biblioteker) på regulær basis. Det kan også bli mulig å søke i flere store databaser samtidig. En sentral database-operatør kan tilby søking i et system bestående av mange videoplatespillere knyttet til en sentral prosessor, eventuelt et juke-boks-system med flere plater. En slik løsning vil bli vesentlig billigere enn en konvensjonell databasekonfigurasjon med magnetplater (Electronic Publishing Review nr.1, 1982 s.76).

I bibliografiske anvendelser lagres i dag oftest et «sekundært» dokument, f.eks. et abstract. Videoplatens store lagringskapasitet åpner for muligheten til å lagre også primær-dokumentet. Det vil imidlertid med dagens priser bli for dyrt å overføre primærdokumentet via telelinje til brukeren. Leveringen måtte sannsynligvis skje ved at videospilleren var tilknyttet en laserskriver. Utskriften sendes så pr. post til brukeren.

Hvis informasjonen (databasen) krever hyppig oppdatering (f.eks. daglig) er dagens utstyr ikke egnet. Når og hvis de overskrivbare platene kommer i produksjon, blir sannsynligvis situasjonen en annen. Teknologien krever også i dag at oppdateringen kan skje sentralt. Hvis mange informasjons-meglere går sammen i et fellessystem, vil trolig et

system av typen «videotex» (toveis kommunikasjon) med sentral databank egne seg bedre.

Et eksempel på en databaseoperatør som allerede benytter seg av videoplateteknologi er Pergamon International Information Corporation, som har lansert *Video Patsearch*. Ved hjelp av en mikromaskin, en videospiller og videoplater har en adgang til tegninger av alle patenter registrert i USA fra 1971 til nå. Platene oppdateres hvert kvartal av Pergamon. Brukerne abonnerer på systemet på en årlig basis.

Kan videoplaten benyttes for *undervisningsformål*? Det vi først og fremst vil diskutere her er det som blir kalt datamaskinstøttet undervisning. Det vil si at datamaskinen (læreprogrammer) i en viss utstrekning har overtatt lærerens og lærebokas rolle. Det er imidlertid ikke en triviell oppgave å lage læreprogrammer. De må være svært gode for at elevene skal godta dem, bli stimulert av dem og fremfor alt lære av dem. På den annen side kan datamaskinen ideelt sett gi hver enkelt elev en ideal-lærer, tilpasset den enkeltes nivå og behov. Til nå har en hatt små muligheter til å kunne vise tekst, bilde, film og lyd på samme skjerm. Det er klart at videoplateteknikken vil ha enorm innvirkning på kvaliteten av de læreprogrammer som vil lages i fremtiden, men det kreves kunnskap og fantasi for å utnytte det nye mediets potensiale fullt ut.

Calico Journal, et nytt tidsskrift som kom med sitt første nummer i juli 1983, rapporterer fra et interessant videoplateprosjekt, kalt «Montevidisco». Prosjektet er utviklet ved Brigham Young University i USA. Videoplateteknikken er brukt til å lage et læreprogram for spansk. Dette er gjort ved å simulere et besøk til en meksikansk by. Montevidisco inneholder 28 hovedsekvenser som hver har flere scener, og hver scene har minst 4 valgmuligheter. Dette medfører at studentene kan bevege seg gjennom byen (programmet) på mange forskjellige måter. Alt ettersom hvilken respons studenten gir på de forskjellige situasjoner han befinner seg i, havner han på ulike steder i byen. Hvis en taxifører spør om han ønsker drosje, kan studenten enten svare ja og kanskje havne på stranden, eller han kan fornærme sjåføren og havne på sykehus!

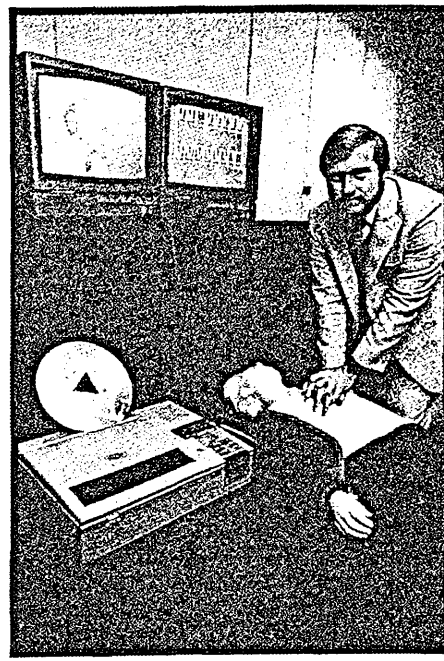
Filmsekvensene ble tatt opp i en meksikansk by. Det bød imidlertid på problemer å holde alle de forskjellige scener og muligheter fra hverandre, fordi det var uvant for produsenten å lage en produksjon med så mange alternative handlingsmønstre. De måtte faktisk skrive et datamaskinprogram for lettere å holde rede på hvilke stier i systemet som var ferdige, og hvor det gjenstod scener. Det ferdige produktet er blitt svært godt mottatt. Representanter for mange ulike land har tatt produktet i øyensyn. «Montevidisco» er hittil bare brukt i 2 forskjellige klasser, slik at grunnlaget for å trekke konklusjoner er meget spinkelt. Den første utprøving har imidlertid gitt grunnlag for en revisjon og forbedring av produktet, og det forhandles om en videre spredning.

Læring ved hjelp av videoplate trenger selvsagt ikke begrenses til de

vanlige skolefagene. Den som har skiftet bremseklosser på bilen for første gang, ville sikkert satt stor pris på en filmatisert veiledning, gjerne med mulighet for ekstra forklaring av vanskelige punkter. Andre tenkelige emner er f.eks. matlaging, håndarbeid, hagestell, dans, gitarspill og førstehjelp.

I USA satses det faktisk stort på opplæring i førstehjelp ved hjelp av videoplateteknikk. For at det ikke skal bli ren teori, har elevene anledning til å øve seg på en «elektronisk» dukke. Dukken har sensorer som sender signaler om hva eleven gjør med den til læreprogrammet. På denne måten kan eleven gis individuell korrigering. Dukkens hjerteslag og pust vises også på en separat monitor, slik at eleven kan følge med og se egne framskritt. (Byte, June 1982 s.108.) (Se figur f.)

Figur f.



Integrering av tekst, lyd, bilde og film, samt muligheten for alternative handlingssekvenser og brukerinteraksjon, åpner til og med muligheten for en ny kunstart. (Slik det skjedde da teknologiske framskritt gav oss filmen.) Kanskje framtidas forfattere vil gi sine romaner en slik uttrykksform? Det er ihvertfall klart at bruk av videoplateteknologi vil forutsette ny kunnskap, og fremfor alt være en utfordring for dem som skal lage produkter basert på den nye teknologien.

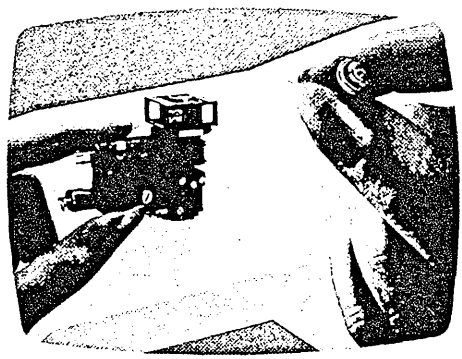
I Norge har museumsfolk, arkivinstitusjoner og biblioteker begynt å interessere seg for den nye teknologien. Så vidt vi vet er det bare Gruppe for bibliotekautomatisering ved RUNIT i Trondheim som er i gang med å forberede et konkret prosjekt. Her orienterer en seg nå og det er planer om senere å kjøpe utstyr og foreta prøver med masselagring av bibliografiske data. I Sverige er det bl.a. to humanistiske institusjoner som har interessert seg for slikt utstyr, nemlig Kulturarvet i Falun (jf. HD 2-83) og Armémuseet i Stockholm. Kulturarvet har for lengst gått til innkjøp, men Armémuseet vil vente på den overskrivbare platen. Grunnen til dette er faktisk ikke selve overskrivbarheten, men at utstyret gir museet muligheten til å fylle opp en plate i flere omganger. En av ulempene ved et «gammeldags» system som Kulturarvet har, er at platen må fylles helt opp med én gang (90 000 bilder). Platen lages nemlig fra en «master». Hvis institusjoner samarbeider om en plate, kan dette problemet til en viss grad omgås. Det gjenstår imidlertid å se om dette er en praktisk mulighet.

Noter

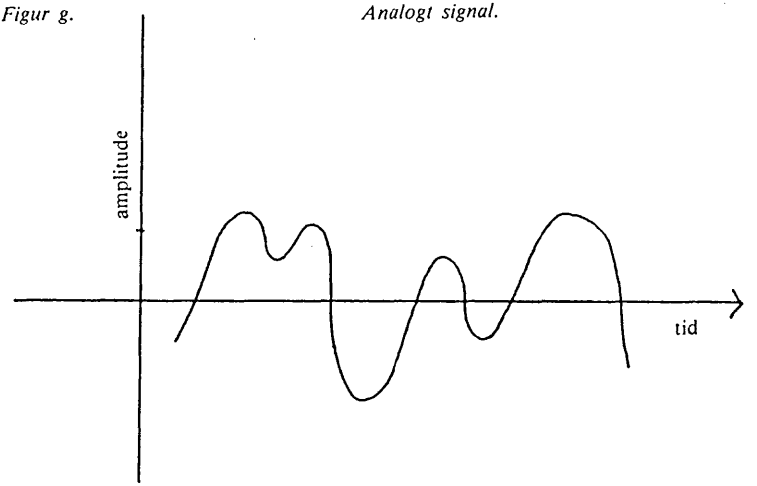
1. Et *analogt* signal er en svingning. (Se figur g.) Svingningen på figuren kan være et lydsignal eller et videosignal. Et *digitalt* signal består av tall, i praksis null og én (binært). En *tekst* lagres digitalt ved at hvert tegn får 8 binære siffer (bit) etter en kodetabell. For å lagre et *bilde* digitalt, må det gjennom en digitaliseringsprosess. Digitalisering kan man tenke seg gjort ved at bildet deles opp i mange små ruter. Hver rute gis et tall som angir fargen i ruten. Digitaliseringen foretas vha. en «scanner». Scanneren sveiper over bildet linje for linje. Hver linje deles så opp i enkeltruter. Ofte foreligger bildet som et videosignal. Da er det allerede scannet, men hver linje er representert som et *analogt* signal. Signalet forteller hvordan fargen og intensiteten skifter langs linjen. Digitaliseringen av et analogt signal består av en samplingsprosess, en kvantifiseringsprosess og en kodingsprosess. Samplingen går ut på å «se» på signalet med jevne mellomrom, kvantifiseringen består i å måle utslaget på hvert punkt, og kodingen i å representere utslaget som et tall (altså digitalt). (Se figur h.)
2. En diodelaser er en type laser som er mindre og billigere enn de tradisjonelle store gass-laserne. Utviklingen av diodelaseren har gjort det mulig å masseprodusere små videospillere til en rimelig pris. (Se figur i.)

Figur i.

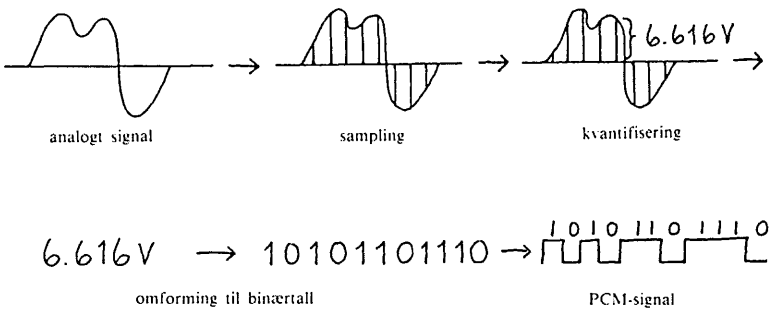
Diodelaser.



Figur g. *Analogt signal.*



Figur h.



Automatisk rotlemmatisering

Tove Fjeldvig og Anne Golden

1. Prosjekt for automatisk rotlemmatisering

Institutt for rettsinformatikk (IRI) (tidligere Institutt for privatretts avdeling for EDB-spørsmål) har i mange år drevet forskning omkring tekstsøkesystemer. I ett av disse prosjektene har vi spesielt tatt opp ulike lingvistiske aspekter knyttet til denne type systemer. Foreløpig har arbeidet vært konsentrert omkring utvikling av en metode for gruppering av ord med felles rot på tvers av ordklassene. Prosessen har fått navnet «automatisk rotlemmatisering», bl.a. for å skille den fra den mer vanlige lemmatiseringsprosessen som opererer innenfor de tradisjonelle ordklassegrensene.

Et *lemma* kan sammenlignes med et slags stikkord eller et oppslagsord i en ordbok. At to ord tilhører samme lemma, betyr enten at de er ulike bøyningsformer av samme grunnform (leksem) eller at de er to ulike skriftvarianter av samme leksikalske ord (f.eks. fram og frem). Et *rotlemma* vil derfor være en betegnelse på ord som har samme rot og samme semantiske betydningen når man ser bort fra den informasjon som ligger i selve bøynings- og avledningsendelsen.

2. Bakgrunn

Arbeidet med gruppering av ord med felles rot ble allerede påbegynt i 1979 som en aktivitet under prosjekt NORIS (34) ved IRI. Dette prosjektet, som var finansiert av Norges Teknisk-Naturvitenskapelige Forskningsråd og ledet av cand. mag. *Tove Fjeldvig*, tok sikte på å undersøke muligheten for enkle strategier for tekstsøking basert på argumenter i naturlig språk.

Blant de problemer man ønsket å belyse i dette prosjektet, var muligheten for automatisk utvidelse av søkeargumentet med alle aktuelle bøyningsformer til søkeordene. Også avledningsformene var aktuelle forutsatt at ordene representerte det samme innholdet (grunnidéen).

I et tekstsøkesystem vil i prinsippet alle ordene i dokumentene være søkbare. Det vil bl.a. si at en bruker må selv definere alle mulige bøyninger og avledninger av aktuelle søkeord. Hvis man for eksempel bare angir søkeordet «BIL», vil man ikke finne de dokumenter som inneholder ordene «BILEN», «BILER» eller «BILENE».

Man fant det også interessant å undersøke om en slik rutine representerte et alternativ til manuell høyre-trunkering, eller om den kunne inngå som et ledd i en rutine for automatisk trunkering.

Trunkering er en måte å spesifisere søkeord på ved å definere en viss følge av tegn som søkeordet skal inneholde. Alle ord som inneholder den definerte tegnstrengen anses kvalifisert som søkeord. Den mest vanlige form for trunkering er høyre-trunkering, hvor tegnstrengens høyre del er uspesifisert. Søkeargumentet BIL* (hvor * er brukt som trunkeringstegn) vil omfatte alle ord som begynner med bokstavene «bil», f.eks. **BIL**ER, **BIL**EN, **BIL**ENE, **BIL**HOLD, **BIL**DE, **BIL**LION. Ulempen med trunkering er at det medfører en del støy og ikke inkluderer vokalvekslinger (f.eks. sterke verb og uregelmessige substantiv).

Dessuten ville en slik rutine gjøre analyse av et søkeargument i naturlig språk lettere, og dermed også øke muligheten for et bedre søkegrunnlag.

I NORIS (34) ble det etterhvert et spørsmål om automatisk rotlemmatisering. Ved prosjektets opphør i 1981 fant vi resultatene såpass interessante, at vi ønsket å fortsette studiene.

Samtidig med prosjekt NORIS (34) pågikk det også et prosjekt LÆREBOKSPRÅK ved Nordisk Institutt (UiO) hvor man var opptatt av lemmatiseringsproblematikken.* Prosjektet var ledet av amanuensis *Anne Hvenekilde* og cand. philol. *Anne Golden* og finansiert av Kirke- og undervisningsdepartementet. Formålet med prosjektet var å kartlegge de høyfrekvente ordene i en del fagbøker for grunnskolen, slik at det kunne lages støttemateriell i norsk for innvandrerelever. Støttemateriellet skulle i første omgang konsentrere seg om vokabularet i fagbøkene, og man ønsket derfor å finne fram til de ordene som de fremmedspråklige elevene fikk mest nytte av å ha lært ved lesing av fagbøker i skolen. I dette arbeidet var det ikke tilstrekkelig å ta utgangspunkt i grafordenes frekvens og heller ikke lemmaets frekvens (dvs. den samlede frekvens for de ulike bøyningsformer av samme grunnord). Det riktigste bilde av ordforrådet fikk man ved å beregne frekvensen til et grunnord med alle dets avledninger. Med andre ord: det var nødvendig å rotlemmatisere ordene.

F.eks. hvis en fremmedspråklig elev har lært ordet «ANVENDE», vil hun (evt. han) også forstå ordene «ANVENDELSE» og «ANVENDELIG» så snart vedkommende også har lært noen enkle regler for ordlaging i norsk.

I dette prosjektet ble grupperingen av ordene foretatt manuelt, og det utviklet seg mange diskusjoner omkring hvilke ord som tilhørte samme semantiske rotlemma.

Ved de Nordiske datalingvistdager 1981 ble vi oppmerksom på vår felles interesse for automatisk rotlemmatisering, og et samarbeid ble etablert. Fordi den ene hadde kompetanse i edb og den andre i lingvistikk, kunne vi nå ta fatt på en rekke av de uløste problemer som vi hver for oss hadde stått overfor. Arbeidet med automatisk rotlemmatisering ble derfor intensivert, og i dag utgjør det et eget delprosjekt under NORIS (58). Dette prosjektet er finansiert av NTNF og tar sikte på å utvikle en «intelligent forsats» til tekstsøkesystemer.

3. Målsetning

3.1 Programsystem

Målsetningen for arbeidet med den automatiske rotlemmatiseringen var å utvikle et programsystem for rotlemmatisering som ikke var for ressurskrevende. Dette var spesielt viktig med tanke på implementering av et slikt program i tekstsøkesystemer, da responstiden i slike systemer spiller en meget viktig rolle.

Det var derfor utelukket å basere programsystemet på et manuelt utviklet leksikon. I steden valgte vi å basere oss på et sett med generelle regler for rotlemmatisering som var uavhengig av datamaterialet. Dette regelsettet kunne forsåvidt også inneholde ord, men disse måtte i tilfelle tilhøre lukkede ordklasser (f.eks. funksjonsord, sterke verb etc.) slik at det ikke oppsto behov for endring av regelsettet ved oppdatering av datamaterialet.

3.2. Rotlemmatisering

Rotlemmatiseringen skulle bidra til at ord med felles grunnform ble gruppert. Dette gjelder ikke ord som i vesentlig grad har fått endret sin betydning ved at de har fått lagt til endelser eller avledninger (f.eks. BEHOLDE og BEHOLDNING, KOMMUNE og KOMMUNIST, OPPDRAG og OPPDRAGELSE, STAT og STATISK.) Eksempel på ord som kan sies å tilhøre samme rotlemma er:

AMERIKA	ANTA
AMERIKAS	ANTAS
AMERIKANSK	ANTOK
AMERIKANSKE	ANTATT
AMERIKANER	ANTATTE
AMERIKANERE	ANTAKELSE
AMERIKANERNE	ANTAGELSE
AMERIKANISERE	ANTAGELSEN
AMERIKANISERT	ANTAKELIG
:	:
:	:

3.3. Homografseparering

Målsettingen omfatter ikke kartlegging av homografer. Dette problemområdet ble ansett for å være for omfattende innenfor rammen av prosjektet. Imidlertid vil en langt større del av homografene være interne homografer (dvs. homografer som har samme rotlemmatilhørighet) enn tilfellet er ved vanlig lemmatisering.

ARBEIDER (nomen agentis) og ARBEIDER (verb, presens) er eksempler på interne homografer i en rotlemmatiseringsprosess. De skal grupperes sammen og skaper derfor ikke noe problem. I en vanlig lemmatiseringsprosess ville disse regnes som eksterne homografer fordi de tilhører hvert sitt lemma.

Derimot vil eksterne homografer virke forstyrrende (f.eks. ordet HELT som både kan være et substantiv, et adverb og et verb i perfektum partisipp). Retningslinjen for grupperingen var at vi skulle forsøke å plassere ordet i rotlemmaet som vi regnet med hadde høyest frekvens, men generelt skulle rotlemmatiseringen aksepteres så sant ordet ble plassert i ett av de riktige rotlemmaene.

4. Gjennomføring

Prosjektet har følgende hovedaktiviteter:

1) Tilretteleggelse av datamateriale
2) Utvikling av et regelsett
3) Utvikling av ett programsystem
4) Testing

I utviklingen av regelsettet var det nødvendig med et eksperimentmateriale. I vårt tilfelle var det naturlig å ta utgangspunkt i det materialet som allerede var tilgjengelig, og eksperimentmaterialet ble derfor sammensatt av ulike fagbøker for grunnskolen (geografi, fysikk og historie) og 2 juridiske dokumentsamlinger (tinglysingsavgjørelser og sammendrag av lagmannsrettsavgjørelser i familie-, skifte- og arverett). Tilsammen besto korpuset av ca. 1/2 mill. løpende ord, hvorav de juridiske samlinger utgjorde ca. 2/3. Vi fant det hensiktsmessig å fjerne skrivefeil, utenlandske ord, nynorske ord, forkortelser og noen ord med gammel skriveform. Når det gjaldt navn, beholdt vi bare egennavn som hadde substantiv som siste ledd og navn på land og verdensdeler. Antall ulike ord i korpuset ble som følge av dette redusert fra ca. 29.000 til ca. 25.000.

Tyngden i prosjektet har helt opplagt ligget i spesifiseringen av regelsettet. Gjennomføringen kan deles i tre stadier:

a) Forslag til hovedregler ble satt opp på bakgrunn av en systematisering av formverket i norsk.
b) Hovedregler ble testet og spesialregler ble innført.
c) Reglene ble vurdert ut fra deres hyppighet.

Arbeidet har nærmest tatt form av en «feedback-prosess». Vi startet med et sett med hovedregler. Disse ble så testet på eksperimentmaterialet, og ut fra en vurdering av feilene ble spesialregler satt opp. Vi gjentok så prosessen med det nye regelsettet, og fortsatte inntil vi sto igjen med en fullstendig systematisering og kategorisering av alle ordene i eksperimentmaterialet. Før det endelige regelsettet ble fastsatt, ble det foretatt en vurdering av de enkelte reglers «eksistensberettigelse» på grunnlag av hvor hyppig de ble brukt dvs. hvor mange ulike og løpende

ord de dekket. Særlig gjaldt dette spesialreglene.

I overensstemmelse med målsettingen ble det lagt vekt på at regelsettet skulle være så uavhengig av korpuset som mulig. Det var allikevel vanskelig å unngå at enkelte regler ble noe «preget» av vårt materiale, f.eks. spesialreglene. For å få en generell bedømmelse av regelsettet til slutt, ble det testet mot et helt tilfeldig valgt materiale (barneboken «Ole Brumm»).

5. Regelsettet

Den ordbehandlingen som regelsettet dekker, kan deles i 7 kategorier:

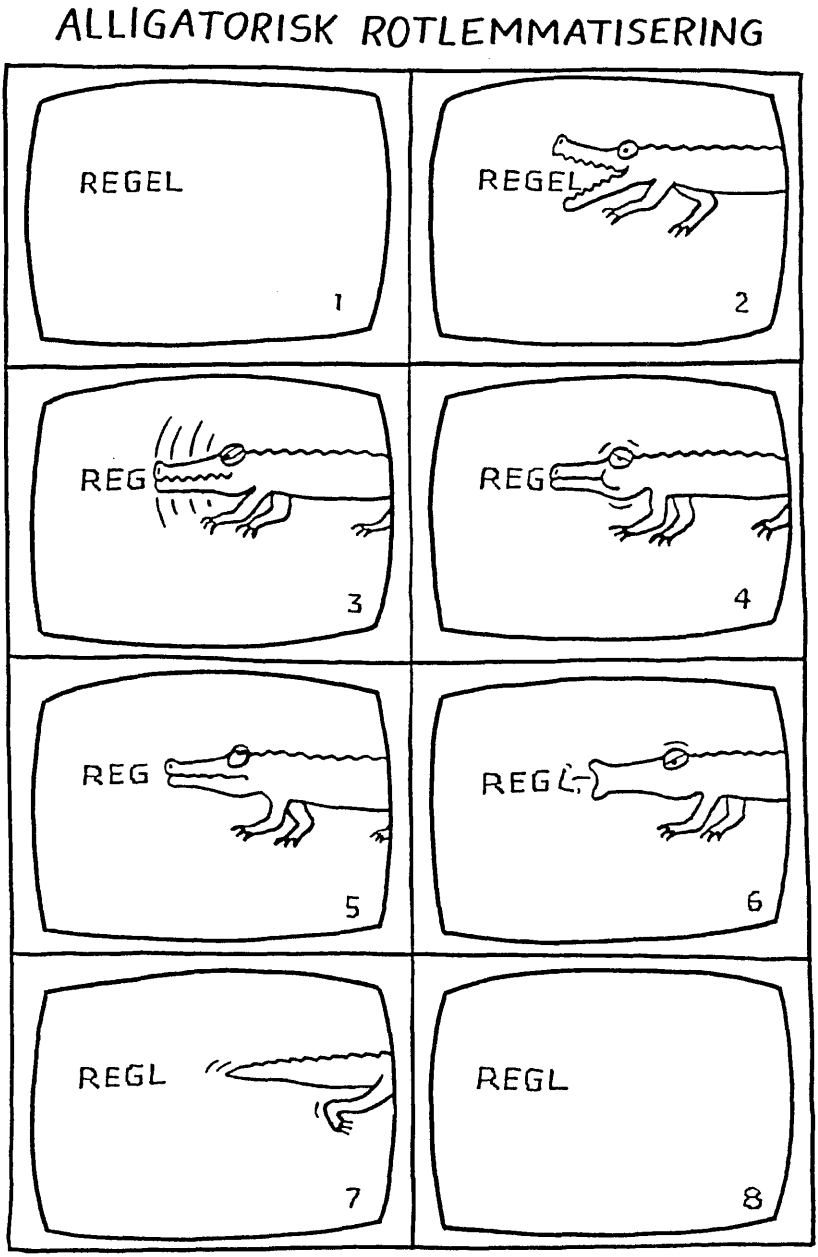
1) fjerning av bøyningsendelser
2) fjerning av avledningsendelser
3) nøytralisering av vokalvekslinger
4) nøytralisering av konsonantforenklinger og -fordoblinger
5) nøytralisering av stavelsessammentrekninger
6) nøytralisering av skriftvarianter av samme leksikalske ord
7) markering av ord som får felles oppslagsform som følge av vår behandling uten at de tilhører samme rotlemma

En del ord må behandles med hensyn til flere av disse kategoriene samtidig. Det finnes derfor forskjellige typer regler. En type er «enkel» og tar bare for seg en kategori av gangen. Disse reglene er i noen tilfeller temporære, dvs. de er kodet slik at de sender ordet til viderebehandling i motsetning til de endelige reglene som avslutter behandlingen av ordet. Andre regler er sammensatte, de dekker flere av kategoriene på en gang.

Etter en systematisk gjennomgåelse av de forskjellige ordklassers paradigmer, satte vi opp et forslag til fjerning av bøyningsendelsene (pkt. 1). Likeledes valgte vi ut en «normalform» når det gjaldt paradigmer som inneholdt vokalvekslinger (pkt. 3), konsonantforenklinger/-fordoblinger (pkt. 4) og stavelsessammentrekninger (pkt. 5). Ord som forekom hyppig i forskjellige skriftvarianter (f.eks. fram/frem, nå/nu) ble lagt inn spesielt (pkt. 6).

> De sterke verbene FINNE og VINNE har vokalvekslingene i-a-u. Vi regner infinitivs-/presensformen som normal form og erstatter preteritums- og perfektumsformen med denne, dvs. -ANT og -UNNET strykes og -INN settes i stedet. (For å forenkle reglene strykes alltid e'en når den er utlyd).

Avledningsendelsene ble vurdert i forhold til hvor stor grad de endret det semantiske innholdet av ordet, og i første omgang ble de aller fleste lagt inn i regelsettet med beskjed om at de skulle fjernes (pkt. 2). Vi undersøkte også hvilke prefikser (preposisjoner og adverb) som forekom hyppig i sammensetninger med sterke verb (f.eks. **INNGÅ**) og funksjonsord (f.eks. **DERPÅ**). Disse ble samlet i 2 forskjellige grupper og skulle hjelpe til å bestemme hvorvidt et ord var et sterkt verb (som måtte nøytralisere vokalvekslingen) eller et funksjonsord (som i de fleste tilfeller skulle beholde den formen det hadde).



Tegning: Øystein Reigem

De reglene vi så kom fram til, kjørte vi ut på vårt eksperimentmateriale for å kartlegge omfanget av følgende problemer:

1) ord med «falske» endelser
2) ord med større uregelmessighet i rotlemmakomponentene enn hovedreglene dekket
3) uheldige grupperinger

1) En falsk endelse

En falsk endelse er en bokstav eller en bokstavkombinasjon som er en del av stammen, men som har en form som er identisk med en bøynings- eller avledningsendelse. I prinsippet kan alle bøynings- og avledningsendelser ha en «tvilling» som er falsk, men det er stor forskjell på hyppigheten av forekomstene av disse falske endelsene. Eksempler på falske endelser er -EN i LAKEN, -ER i METER, -S i PRIS og -A i KOLLEGA. Hvis målsettingen hadde vært å finne fram til stammen i ordene, måtte disse falske endelsene beholdes. Men med vår målsetting kan vi tillate oss å la disse bli fjernet *så lenge oppslagsordet for rotlemmaet* (dvs. det som blir igjen av stammen) *blir entydig.*

-A'en i utlyd kan være en bøyningsendelse (bestemt form entall hunkjønn eller bestemt form flertall intetkjønn) og skal da fjernes. Men den kan være en del av stammen (eks. KOLLEGA) eller den kan tilhøre et sterkt verb i preteritum (INNLA). Når A'en er del av stammen, burde den beholdes, når den er utlyd i forbindelse med et sterkt verb, burde det sterke verbet forandres til «normalformen» (se over). Dette har vi løst på følgende måte: Hovedregelen er at A'en fjernes i utlyd. At den dermed blir fjernet i ord som KOLLEGA, løser vi ved også å fjerne den i de andre bøyningsformene i ord med -A som siste bokstav i stammen (- AEN, -AER, -AENE i eksemplet med kollega). Så lenge oppslagsordet er entydig, er vårt krav oppfylt. Ved endelser som har en bokstavkombinasjon som kunne tilsi at de er sterke verb, kaller vi opp prefikslisten for sterke verb og sjekker ordets begynnelse mot denne. Det viser seg nemlig at svært mange av de sammensatte sterke verbene nettopp består av en prefiks + verbet (det viktigste unntaket er LEGGE som også danner forholdsvis mange sammensetninger med substantiv). Hvis vi får tilslag på den aktuelle prefikslisten, blir ordet oppfattet som et sammensatt sterkt verb og behandlet deretter. I eksempelet med INNLA ville INN være å finne blant prefiksene for verb, og ordet ville forandres til INNLEGG, mens f.eks. KULA ikke ville få tilslag på prefikslisten og ville følge hovedregelen.

Hvis derimot oppslagsordet for et rotlemma faller sammen med oppslagsord for andre rotlemma, kan ikke de falske endelsene fjernes. I disse tilfellene må vi legge inn spesialregler for å kunne skille mellom de ekte og de falske endelsene.

I noen tilfeller sløyfet vi hovedregelen, og de ordene som hadde denne bokstavkombinasjonen som en virkelig avlednings- eller bøyningsendelse, fikk spesialregler. I andre tilfeller var det ordene med falske endelser som fikk spesialreglene. Metoden vi valgte var alltid den som gav færrest regler totalt sett.

2) Ord med større uregelmessighet i rotlemmakomponentene enn hovedreglene dekket

En del ord som klart tilhører samme rotlemma, viser større uregelmessighet enn våre hovedregler tilsier. Dette gjelder i første omgang fremmedord, lånt fra latin og gresk. I slike tilfeller måtte vi legge inn noen spesialregler som gikk forbi suffiksgrensen og behandlet stammen i ordet.

PRODUSERE og PRODUKSJON tilhører samme rotlemma og skal derfor grupperes sammen. Ved å fjerne -ERE og -SJON ville vi stå igjen med oppslagsordene PRODUS og PRODUK som må tillempes hverandre. I dette tilfellet vil PRODU være en entydig oppslagsform, og vi la inn regler som fjernet S'en og K'en i forbindelse med disse suffiksene.

3) Uheldige grupperinger

Uheldige grupperinger er ord som tilhører forskjellige rotlemmaer, men som får felles oppslagsform uten at de i utgangspunktet er homografer. I mange tilfeller dreier det seg om innholdsord som får samme form som et funksjonsord. I slike tilfeller har vi lagt inn og markert de aktuelle funksjonsordene (de tilhører jo en del av ordforrådet som ikke ekspanderer), slik at vi kan skille gruppene fra hverandre.

E'en i utlyd blir alltid fjernet, og ordet MENE vil derfor få rotlemmaet MEN. For at det ikke skal bli gruppert sammen med konklusjonen MEN, har vi lagt inn konjunksjonen og markert denne. Dermed får vi skilt de to rotlemmaene fra hverandre.

6. Programsystem for automatisk rotlemmatisering

6.1 Oversikt

Figuren på s. 30gir en oversikt over programsystemet for automatisk rotlemmatisering.

Inndata til programmet er ett eller flere ord, f.eks. en frekvensordliste som i vårt tilfelle. Resultat er en rotlemmatisert liste hvor hvert ord er angitt med rotlemma og en del tilleggsinformasjon, f.eks. hvilke regler som er brukt og om ordet er et funksjonsord. Det siste er nyttig informasjon ved gruppering av ordene fordi vi ikke ønsker å gruppere funksjonsord sammen med andre ord (jfr. eksemplet med MEN og MENE).

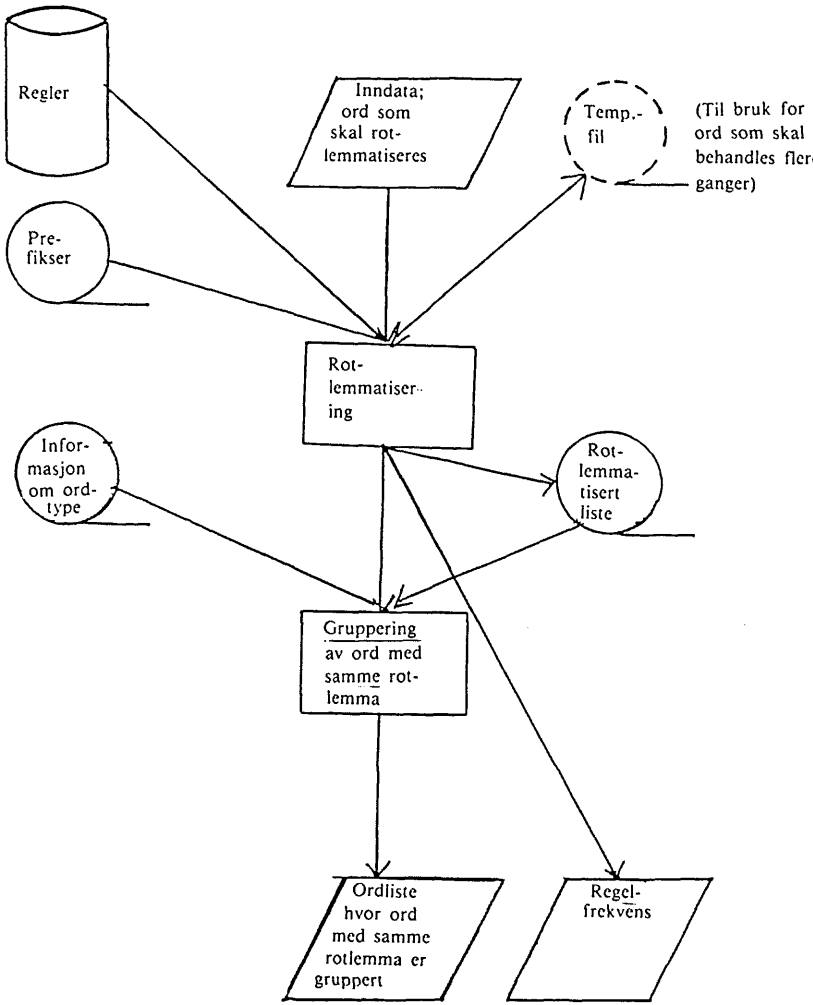
Som del av resultatet får man også en oversikt over hvor hyppig den enkelte regel er brukt. F.eks. endelsen -EN er brukt 3399 ganger, mens -ENE er brukt bare 1681 ganger. Denne informasjon var til stor nytte for oss ved spesifisering av regellisten, og den kan også være interessant for dem som ønsker å studere suffiksene i et datamateriale.

Den rotlemmatiserte listen blir til slutt gitt som input til et program som grupperer alle ord med samme rotlemma. Dette programmet beregner også den samlede frekvens for rotlemmaene.

6.2 Spesifisering av reglene

Hver regel inneholder en tegnstreng, en typebetegnelse, en ordre og et krav.

Oversikt over programsystemet for automatisk rotlemmatisering



Tegnstrengen består av ett eller flere tegn som skal sjekkes mot ordets høyre del. I enkelte tilfeller kan eller må tegnstrengen utgjøre hele ordet.

Typebetegnelsen angir om ordet f.eks. er et funksjonsord eller et sterkt verb. Denne informasjonen hindrer at bestemte typer ord (f.eks. funksjonsord) blir gruppert sammen med andre ord.

Ordren gir informasjon om hvor stor del av ordet som eventuelt skal fjernes og hvilken tegnstreng som eventuelt skal legges til. Dessuten gir ordren beskjed om ordet skal behandles på nytt etter at ordren er utført (f.eks. ord med genitivs *s*).

Et generelt *krav* til rotlemmaet er at det minst må bestå av to tegn, hvorav det ene må være vokal. I tillegg kan hver regel stille krav til

a) hvor stor del av ordet tegnstrengen skal utgjøre (f.eks. hele ordet eller bare en del av det),
b) ordets begynnelse (f.eks. at det skal være en prefiks),
c) at tegnstrengen ikke er etterfulgt av andre tegn (dvs. at ordet ikke har vært behandlet tidligere). I slike tilfeller kaller vi regelen lukket.

Eksempler på regler og bruken av dem er gitt på s. 32.

6.3 Algoritme

Programmet behandler ordet bakfra. Etterhvert som det beveger seg et tegn mot venstre, sjekkes ordets høyre del mot tegnstrengen i regelen. Dette gjentas for hver regel inntil ordets høyre del finner en regel som passer, eller at alle reglene er sjekket. I det siste tilfellet får ordet et rotlemma som er lik ordet.

Hvis regelen passer – dvs. at tegnstrengen er lik ordets høyre del eller hele ordet – sjekkes først regelens krav. Er disse også tilfredsstilt, utføres ordren. Hvis ikke, fortsetter søkeprosessen nedover regellisten.

I de tilfeller et ord skal behandles på nytt igjen, merkes ordet og legges ut på en temporær fil. Denne filen behandles til slutt som om den var en vanlig inputfil. Det er ingen grenser for hvor mange ganger et ord kan behandles før rotlemmaet er bestemt.

Regellisten er organisert som en kjedet fil for å gjøre søkingen mer effektiv. Dessuten er både inndata og regellisten (tegnstrengen) sortert bakfra i samme rekkefølge, slik at programmet ikke starter øverst på resultatlisten for hvert nytt ord. Det finnes mange alternative søkeprosesser som er mer effektive, og programsystemet BETA ville f.eks. ha vært velegnet i dette tilfellet (jfr. Benny Brodda 1982 «The BETA system», foredrag holdt på COLING i 1982).

6.4 Eksempel på automatisk rotlemmatisering

Vi ønsker å rotlemmatisere VARE og GÅRDEIERNES og forutsetter at reglene på s. 32 gjelder.

Behandling av VARE:

1. gang passer regel 6 og endelsen -E blir fjernet.
   Resultat: VAR.
   Ordren gir beskjed om at ordet skal behandles på nytt.
2. gang stemmer ordet overens med tegnstrengen i regel 3, men kravene til regelen er ikke tilfredsstilt (jfr. krav c). Søkingen fortsetter og regel 4 passer. Ordet tilfredsstiller kravene og får rotlemmaet VAR.

Regel nr.	Tegn-streng	Type-betegn.	Ordre			Krav a): Tegn-strengen må ut-gjøre...	Krav b): Sjekk prefiks-liste	Krav c): Lukket?
			Slett	Legg	Behandl.			
1	S		1		Ja	høyre del av ordet		
2	ER		2		Nei	høyre del av ordet		
3	VAR	sterkt verb	2	ÆR	Nei	hele ordet		Ja
4	R		0		Nei	høyre del av ordet		
5	ERNE	subst. eller verb	2		Ja	høyre del av ordet		
6	E		1		Ja	høyre del av ordet		
7	LA	sterkt verb	1	EGG	Nei	høyre del av ordet el. hele ordet	Ja	Ja
8	A	subst. eller verb	1		Nei	høyre del av ordet		

Behandling av GÅRDEIERNES:

1. gang passer regel 1 og endelsen -S blir fjernet.
   Resultat: GÅRDEIERNE
   Regelen inneholder ikke typebetegnelse, men gir beskjed om at ordet skal behandles på nytt.
2. gang stopper prosessen ved regel 5 og endelsen -NE fjernes.
   Resultat: GÅRDEIER og typebetegnelse verb eller substantiv.
   Også denne regelen gir beskjed om at ordet skal behandles på nytt.
3. gang passer regel 2 og -ER fjernes. Ordet får rotlemmaet GÅRDEI og typebetegnelsen beholdes.

7. Resultat

7.1 Oversikt

Eksperimentmaterialet besto av 489.382 løpende ord og 23.890 ulike graford. Av disse ble 685 (2,8%) ulike graford gruppert feil. Tilsammen utgjorde disse 6.740 løpende ord - dvs. 1,4% av det totale antall ord i korpuset.

7.2 Typer feil

Feilene er inndelt i 3 hovedkategorier.

a) «Tunge feil» - (utgjør 2,3%)
b) «Lette feil» - (utgjør 0,3%)
c) «Vriene feil» - (utgjør 0,3%)

Tunge feil har vi valgt som eksempel på ord som ikke er behandlet i samsvar med de 7 kategoriene i avsnitt 5. Disse kan inndeles i 3 undergrupper.

a1) Rotlemmaet inneholder endelser som burde ha vært fjernet eller at rotlemmaet ikke er fullstendig. Det siste tilfellet gjelder spesielt avledningsendelser som fører til at ordet får endret sitt semantiske innhold i forhold til ordstammen, for eksempel:

DEFINISJON	får	rotlemma	DEFINISJON,	mens	ønsket	rotlemma	er DEFI
SYKT	"	"	SYKT,	"	"	"	" SYK
SMERTE	"	"	SMER,	"	"	"	" SMERT

a2) Rotlemmaet blir for lite og derfor tvetydig, for eksempel:

FETTER	får	rotlemma	FETT,	mens	ønsket	rotlemma	er FETTER
SETET	"	"	SE,	"	"	"	" SET
LEVERE	"	"	LEV,	"	"	"	" LEVER

a3) Ordet inneholder stavelsessammensetninger som ikke blir nøytralisert, for eksempel:

USSEL	får	rotlemma	USSL,	mens	ønsket	rotlemma	er USL
SYKKEL	"	"	SYKKL,	"	"	"	" SYKL

Lette feil omfatter ord hvor avledningsendelsen ikke er fjernet, men hvor endelsen i en viss grad endrer det semantiske innholdet til ordet i forhold til ordstammen. Endringen er allikevel ikke så stor at vi vil beholde endelsen, eksempelvis:

ADRESSAT	er	ikke	gruppert	sammen	med	ADRESSE
BILIST	"	"	"	"	"	BIL
GARANTIST	"	"	"	"	"	GARANTI, GARANTERE

Vriene feil gjelder ord som har så avvikende skrivemåte i de ulike former, at det kreves spesialbehandling i hvert enkelt tilfelle, eksempelvis:

KONTO og KONTI
EPILEPSI og EPILEPTISKE
FØDSELSDATO og FØDSELSDATUM

7.3 Homografer

Målsetningen for den automatiske rotlemmatiseringen omfattet ikke homografseparering (jfr. også pkt. 3.3). Ord med felles rot er derfor plassert i samme gruppe, selv om de har avvikende innhold. Dette har vi ikke regnet som feil. Det samme gjelder selv om ordet i sin bøyde form er entydig, f.eks.:

RETTSLIG	som	har	fått	rotlemmaet	RETT	(homograf	en(varm)rett)
VARIG	"	"	"	"	VAR	("	"	(mat)vare)
MUNNING	"	"	"	"	MUNN(	"	"	munn)

7.4 Rotlemmatisering av et tilfeldig valgt materiale

For å få en så generell bedømmelse av den automatiske rotlemmatiseringen som mulig, ble regelsettet helt til slutt utprøvd på et tilfeldig valgt materiale. Vi valgte barneboken *Ole Brumm* som består av 19.725 løpende ord og 2.369 ulike graford.

Resultatet viste at ca. 97,6% av de ulike ordene ble plassert i riktig gruppe.

8. Videreføring

Språkbruken i tekster fra forskjellige fagområder kan variere sterkt, og et regelsett som har som mål å være tekstuavhengig må nødvendigvis bli ganske omfattende. Dessuten vil kravet til korrekthet variere alt etter hva rotlemmatiseringen skal brukes til. I de to konkrete problemstillingene vi tok utgangspunkt i, var korrekthetskravet forskjellig. I det førstnevnte – tekstsøkingsproblemet – er det viktig at de ordene som blir vurdert som søkeord blir gruppert riktig, mens det ikke gjør noe om det forekommer feilgrupperinger blant støyordene. Den andre problemstillingen – utskillingen av høyfrekvente ord – krever derimot at alle typer ord blir gruppert riktig. Hvis ikke ville frekvensen forskyve seg, og man

vil ikke finne fram til de gruppene man ønsket. Man kunne derfor tenke seg at et regelsett ble delt inn i forskjellige lag eller pakker. Utgangspunktet kunne være en bunke med basisregler, og man kunne tilføre lag med spesialregler som gav større korrekthet. Dessuten kunne man tenke seg spesielle fagpakker eller genre-pakker, f.eks. en juss-pakke en fysikk-pakke, en medisin-pakke eller en språkkonservativ-pakke og en språkradikal-pakke. På denne måten kunne man få et regelsett som var tilpasset både teksten og bruken.

En del av reglene er markert med hensyn til ordklassetilhørighet. Alle reglene er markert enten som funksjonsord eller som innholdsord. Denne markeringen er først og fremst lagt inn for å skille mellom ord som ellers ville bli gruppert sammen. Dessuten har det vært en hjelp under arbeidet å vite hvilke ord en regel er ment å dekke. Dette arbeidet kunne lett utvides, og en automatisk markering av ordklassetilhørighet skulle være innen rekkevidde.

* Se HD 1-83 (Red.anm.)

Grammatical tagging of the LOB Corpus: A progress report

Stig Johansson

1 Introduction

The Lancaster-Oslo/Bergen (LOB) Corpus is a collection of British English texts in machine-readable form. In size and composition it is comparable with the Brown Corpus of American English (see Johansson et al. 1978). The present paper reports on the grammatical tagging of the LOB Corpus, a project which is nearing its completion. The project has been undertaken by researchers at the universities of Lancaster and Oslo, in cooperation with the Norwegian Computing Centre for the Humanities at Bergen. The principal members of the research team have been: *Geoffrey Leech*, *Roger Garside*, *Eric Atwell* and *Ian Marshall*, University of Lancaster; *Stig Johansson* and *Mette-Cathrine Jahr*, University of Oslo; *Knut Hofland*, Norwegian Computing Centre for the Humanities. The project has been supported by the Social Science Research Council in the United Kingdom and by the Norwegian Research Council for Science and the Humanities.

It is not possible to give full details on the project in this brief article. For more information, consult the papers and reports listed at the end of the article.

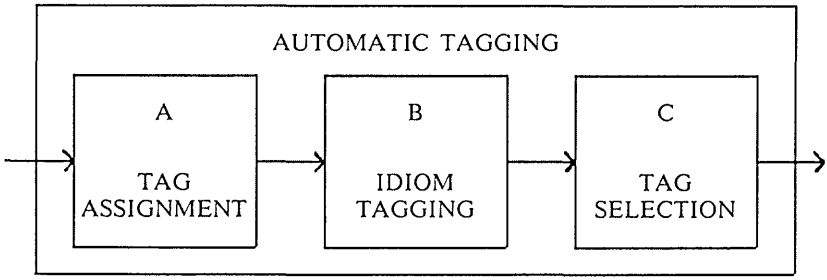
2 Aims of the project

In its raw state the LOB Corpus has already proved to be a valuable research tool (see my previous paper in *Humanistiske data*, Johansson, 1981, and the bibliography in Johansson, 1982). One of the aims of the project has been to provide an even richer data base which can make possible more sophisticated grammatical studies. A second aim has been to develop a system for automatic grammatical analysis. As the project has evolved, particularly at the University of Lancaster, the emphasis has shifted more and more towards the latter aim.

3 An outline of the tagging system

As a starting-point we took the system of automatic grammatical tagging developed for, and successfully applied to, the Brown Corpus (see Greene and Rubin, 1971, Francis, 1980, Francis and Kučera, 1982). Each word in the Brown Corpus was provided with one of 87 tags (e.g. NN = singular noun, IN = preposition). The system assigning the tags, henceforth referred to as BROWN, contained three main components: a word list assigning tags to individual word forms, a «suffix» list

Figure 1 The LOB tagging suite (quoted from Leech et al. 1983)



assigning tags to words ending in particular character sequences, and a set of context-frame rules selecting the appropriate tag in each context for forms which were not unambiguously tagged by the word list and the «suffix» list. BROWN achieved an overall success in automatic tag assignment in the region of 80 per cent.

The LOB tagging project builds on the experiences from the tagging of the Brown Corpus. The tag set is basically the same, although a good number of new distinctions have been introduced (in all 134 tags as compared with 87 for BROWN); see the Appendix. The tagging is done through a set of programmes which we shall refer to as the LOB tagging suite (see Figure 1). In addition, there is a manual pre- and post-editing stage. Of the three main components in Figure 1 only A is comparable with BROWN.

The Tag Assignment program contains a word list and a «suffix» list, both considerably larger than in BROWN (7,000 vs. 3,000 words and 660 vs. 450 word endings). A new feature is that all tags for a word or word ending are consistently ordered according to frequency, based on studies of the tagged Brown Corpus. The probabilistic aspect is characteristic of the LOB tagging suite as a whole. Apart from the word list and the «suffix» list, the Tag Assignment program contains special routines for words ending in *-s*, words beginning with capital letters, etc. Words which are not assigned one or more tags by the word list, the «suffix» list, or the special routines are given the default tags NN (= singular noun), VB (= verb, uninflected form), JJ (= adjective).

The Idiom Tagging program was introduced to deal with idiosyncratic word sequences like the pronoun *each other* or the conjunction *in order that*. The program, which contains a list of «idioms» with their associated tags, assigns tags to sequences of words rather than individual words. In the LOB tagging project it plays a fairly modest role.

After the operation of the Tag Assignment and Idiom Tagging programs many words are provided with two or more tags. It remains for the Tag Selection program to identify the correct tag in each context. The Tag Selection program, which replaces the BROWN

Figure 2 Sample output of the LOB tagging suite

		Also eliminated:
it	PP3	
is	BEZ	
not	XNOT	
'	'	
by	IN	RI%
a	AT	
long	JJ	RB, VB%
way	NN	
'	'	
her	PP$	PP3O
first	OD	RB%
experience	NN	VB@
of	IN	
exhibiting	(NN@/ 41 VBG/ 38 JJ@/ 21	
'	'	
for	(IN)/ 56 CS@/ 44	
Miss	NPT	
Smith	NP	
has	HVZ	
been	BEN	
painting	VBG	NN@,JJ@
most	(AP)/ 70 RBT@/ 18 QL/ 12	
of	IN	
her	PP$	PP3O
life	NN	
'	'	
striving	(VBG)/ 87 JJ@/ 9 NN@/ 4	
to	(TO)/ 98 IN/ 2	
express	(VB)/ 98 NN/ 2 JJ/ 0 RB@/ 0	
in	(IN)/ 65 RP%/ 35	
oils	NNS	VBZ
on	IN	RP
hardboard	NN	
the	ATI	
ideas	NNS	
that	(WP)/ 36 CS/ 0 DT/ 0 QL%/ 0	
come	(VB)/ 36 VBN/ 0 NN%/ 0	
too	(QL)/ 36 RB@/ 0	
fast	(RB)/ 36 JJ/ 0 VB@/ 0	NN@
for	(IN)/ 36 CS@/ 0	
her	(PP3O)/ 36 PP$/ 0	
to	(TO)/ 36 IN/ 0	
cope	(VB)/ 36 NN/ 0	
with	IN	

Note

The probabilities for the tags in the ambiguous sequence at the end do not add up, presumably because of an overflow in the buffer storing the different alternatives (a very high number in this case).

context-frame rules, was developed at the University of Lancaster. It is the principal new feature in the LOB tagging suite and will therefore be dealt with in greater detail.

4 The Tag Selection program

There were several problems with the BROWN context-frame rules. In the first place, they only took the immediate context into account, more specifically the word to be disambiguated ±2. Secondly, the rules only worked when one or more of the words in the context were unambiguously tagged: The new Tag Selection program overcomes these difficulties.

The Tag Selection program is based on the probability of tag combinations, initially extracted from a study of the frequency of tag sequences in the Brown Corpus. Given a sequence of words with more than one tag the program computes all possible combinations and assigns a probability to each. Probabilities for individual tags are derived from the probability of the sequences in which they occur. The procedure is explained in detail in Atwell (1983).

An example of the output of the Tag Selection program is given in Figure 2. In the output tags are ordered according to probability. The figures accompanying the tags are probability values expressed in per cent. For example, *express* has four possible tags; the probability for VB (= verb, uninflected form) is 98, for NN (= singular noun) 2, for JJ (= adjective) 0, for RB (= adverb) 0. Parentheses indicate the preferred tag. Certain tags which are judged as highly improbable are not specified in the output. They are given in the righthand column in Figure 2. The symbols @ and % are rarity markers accompanying particular tags. For example, RB@indicates that this word is rarely used as an adverb. They play an important role in tag selection.

In the example taken up in Figure 2 the program selected (i.e. assigned the highest probability to) the correct tag in all the cases where a word had more than one tag. This is quite remarkable. In the sequence *the ideas that come too fast for her to cope with* the number of possible tag sequences is well over 1,500. Still, the program has selected the correct tag for each word.

The LOB tagging suite achieves an overall success rate of 96-97 per cent, i.e. only 3-4 words out of a hundred are mis-tagged. Errors occur in examples like:

Gaunt was a disappointed and now even *despised* failure. (*despised* identified as VBD, = verb, past tense)

The rich *still* supplied the traditional revenues. (*still* identified as NN, = singular noun)

So he was very much afraid of a charge of over-presumptuousness had he *let* his book either be published at home or ... (*let* identified as VBD, = verb, past tense)

... unless he had first put his thesis before his fellows and *confirmed*

it by visual demonstration. (*confirmed* identified as VBD, = verb, past tense)

... the expenses of *waging* war (*waging* identified as JJ, = adjective)

Errors occur especially in cases of ellipsis and with unusual word orders. They also occur frequently in cases where the human analyst has difficulties in drawing a borderline (e.g. between the various uses of *-ing* and *-ed* forms of words, between adjectives and adverbs, etc.). A study of errors made by the Tag Selection program could be of great interest for the grammarian.

5 Developments

The complete LOB Corpus has now been tagged and the project is currently in the post-editing stage. In spite of the high success rate, the output of the programs must be checked and corrected manually. This is a laborious and time-consuming task. After post-editing has been completed, the aim is to make available the tagged corpus to researchers through ICAME. Lemmatized word lists will be produced as well as a concordance with key words sorted according to tag. Work on automatic grammatical analysis will be continued at the University of Lancaster, where the research team has been given a new grant to explore the possibilities of syntactic parsing, based on the principles of the Tag Selection program.

The use of the LOB tagging suite thus extends beyond the immediate goal of tagging the LOB Corpus. A trial run on a text not included in the corpus resulted in a success rate of approximately 90 per cent (see *ICAME News* 7, 1983, p. 7f.), although the text had not been subjected to the usual pre-editing. The programs can therefore be successfully applied to English texts in general. Special uses which have been envisaged include the application of the programs in the development of an automatic context-sensitive spelling checker and an automatic grammar and style checker (see Atwell, 1983).

Experiences from the LOB tagging project are very encouraging. It is uncertain to what extent the programs can be adapted to deal with other languages. It is certainly true, however, that they are well suited to deal with a language like English, where there is a great deal of homonymy and a fairly fixed word order.

References

Atwell, Eric. 1983. «Constituent-Likelihood Grammar», *ICAME News* 7, 34-67.

Francis, W. Nelson. 1980. «A Tagged Corpus - Problems and Prospects». In S. Greenbaum, G. Leech, and J. Svartvik, eds, *Studies in English Linguistics - for Randolph Quirk*. London: Longman. 192-209.

Francis, W. Nelson and Henry Kučera. 1982. *Frequency Analysis of English Usage: Vocabulary and Grammar*. Boston: Houghton Mifflin.

Greene, Barbara B. and Gerald M. Rubin. 1971. «Automatic Grammatical Tagging of English». Providence, R.I.: Department of Linguistics, Brown University.

ICAME News. Newsletter of the International Computer Archive of Modern English

(ICAME). Norwegian Computing Centre for the Humanities, Bergen.
Johansson, Stig. 1981. «Current Work on the LOB Corpus», *Humanistiske data* 1, 1981, 19-21.
Johansson, Stig. 1982. (ed.) *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.
Johansson, Stig, Geoffrey N. Leech, and Helen Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus, for Use with Digital Computers*. Oslo: Department of English, University of Oslo.
Johansson, Stig and Mette-Cathrine Jahr. 1982. «Grammatical Tagging of the LOB Corpus: Predicting Word Class from Word Endings». In Johansson (1982), 118-46.
Leech, Geoffrey N., Roger Garside, and Eric Atwell. 1983. «The Automatic Grammatical Tagging of the LOB Corpus», *ICAME News* 7, 13-33.
Marshall, Ian. 1982. «Choice of Grammatical Word-Class without Global Syntactic Analysis for Tagging Words in the LOB Corpus». Lancaster: Department of Computer Studies, University of Lancaster.

APPENDIX: A selection of tags from the LOB tag set

&FO	formula
AT	singular article (*a*, *an*, *every*)
ATI	singular or plural article (*the*, *no*)
CD	cardinal numeral
CS	subordinating conjunction
DT	singular determiner
DTI	singular or plural determiner
IN	preposition
JJ	adjective
JJB	attributive adjective
NNU	unit of measurement unmarked for number (e.g. *ft.*, *m.p.h.*)
NN	singular common noun
NNP	singular common noun with word-initial capital (e.g. *Irishman*)
NP	singular proper noun
NPL	singular locative noun with word-initial capital (e.g. *Square*)
NPT	singular titular noun with word-initial capital (e.g. *Mr*, *Lord)*
NR	singular adverbial noun (e.g. *north*, *home*)
OD	ordinal numeral
PP1A	*I*
PP2	*you*
PP3	*it*
PP3A	*he*, *she*
PP3O	*him*, *her*
PP$	possessive determiner
QL	qualifier (e.g. *very*, *more*)
RB	adverb

RI	prepositional adverb (homograph of preposition)
RP	prepositional adverb which can also be a particle
VB	verb, uninflected form
VBD	past tense verb
VBN	past participle
VBZ	verb, 3rd person singular present tense
WP	WH-pronoun

Note

A number of punctuation tags represent themselves: . , ? etc. The letter *S* added to a tag marks it as plural, e.g. NNS = plural common noun. The dollar sign added to a tag marks it as genitive or possessive, e.g. NNS$ = genitive plural common noun. R and T added to JJ and RB indicate comparative and superlative forms, respectively.

Edb og språknormering

Aagot Landfald

Det har vist seg at bruken av edb har vært til god hjelp i arbeidet med språknormering. I Norsk språkråds sekretariat har vi mange slags listeprodukter som er laget på grunnlag av et språkmateriale som finnes i edb-format. De fleste av disse ordlistene skriver seg fra vårt samarbeid med Prosjekt for datamaskinell språkbehandling (PDS) i Bergen. Men i de senere årene har vi også fått noen listeprodukter på grunnlag av et skjønnlitterært språkmateriale som Norsk tekstarkiv tilrettelegger i edb-format for oss.

Norsk ordregistrant

I samarbeid med PDS har Norsk språkråd utarbeidet Norsk ordregistrant, som er en ordliste i dataformat. Arbeidet med å lage ordregistranten tok til alt i 1965 og 1966. I Norsk språknemnds årsmelding for 1965 står det: «Det har vært forberedende drøftinger av et prosjekt for utarbeiding av en offisiell ordliste eller ordregistrant for begge mål. Dette er en gammel tanke som har fått ny aktualitet ved den plan om behandling av ordtilfanget ved hjelp av datamaskin, som bl.a. har vært drøftet på de nordiske språkmøtene i 1963 og 1964. For en systematisk jevnføring av ordbruk og skrift- og bøyningsformer i våre to mål vil en slik ordregistrant overført til datamaskin være et uvurderlig hjelpemiddel.» Om den videre behandlingen av ordregistranten kan en lese i årsmeldingene fra Norsk språknemnd og senere fra Norsk språkråd. *Kolbjørn Heggstad*, som har ledet det edb-tekniske arbeidet med ordregistranten, har redegjort for dette arbeidet og for form og omfang av registranten i heftet «Datamaskinell språkbehandling PDS 1967 - 1976». PDS har også laget en beskrivelse av ordregistranten i «Norsk ordregistrant 1972. Kodesystemet», 1972.

I stor utstrekning arbeider Norsk språkråd med enkeltord, hvordan de skrives og bøyes. Nettopp til ordstudier kan en på en enkel måte dra nytte av de forskjellige listeproduktene vi kan få ut av ordregistranten. Ordregistranten inneholder det sentrale ordforrådet i bokmål og nynorsk, og ordene er innført i sin offisielle form, dvs. skrevet og bøyd i samsvar med offisiell rettskrivning. Vi kan på grunn av kodesystemet få ut for eksempel alle substantiv som er hankjønn i bokmål, eller vi kan få ut alle substantiv som har ulikt genus i bokmål og nynorsk. Det kommer stadig spørsmål om behandlingen av nye ord eller av ord som det aldri har vært tatt standpunkt til. Ved hjelp av forskjellige typer utskrifter kan vi sette slike ord inn i en sammenheng. Vi kan sammenligne med andre ord med lignende form eller med lignende betydning. (Ordregistranten inneholder ikke koder for betydning, men de listene

som kan utskrives på formelle kriterier, kan være grunnlaget for videre, «manuelt», arbeid med å skille ut betydningsgrupper.)

I årenes løp har vi fått mange store og små lister på grunnlag av Norsk ordregistrant. Mange av listene ble opprinnelig laget som et ledd i opprettelsen av registranten. Men det har vist seg at listene også kan brukes i forbindelse med enkelte rettskrivningssaker. Noen lister har vi bestilt spesielt for å bruke dem som hjelpemiddel i normeringsarbeidet.

Disse listene omfatter hele ordregistranten:

bokmål og nynorsk - forlengs alfabetisk
bokmål og nynorsk - baklengs alfabetisk
bokmål og nynorsk - kodesortert
bokmål - alfabetisk
nynorsk - alfabetisk
bokmål - kodesortert
nynorsk - kodesortert

Listene ovenfor har kode ved hvert enkelt ord. Koden viser om et ord er felles for bokmål og nynorsk, eller om det er et bokmålsord eller et nynorskord. Den viser om den oppførte formen er hoved- eller sideform, og angir ordklasse og bøyning. Ord med uregelmessig bøyning har en kode som viser det. Vi har også et par andre lister som PDS har laget på grunnlag av ordregistranten:

bokmål baklengsordliste (Norske språkdata nr. 2)
nynorsk baklengsordliste (Norske språkdata nr. 3)

I disse to listene har ordene ordklassemarkering, men ellers ikke grammatiske opplysninger.

Det sier seg selv at lister der ordene er ordnet baklengs alfabetisk eller i grupper etter bøyningstype, er til uvurderlig nytte i språknormeringen. Den baklengssorterte listen setter oss i stand til å vurdere ord med samme avledningsendelse under ett. Vi kan se alle ord som ender på -el i sammenheng. Koden ordene er forsynt med, viser oss hvordan de bøyes. Eller vi kan se hvordan ord på -ning behandles i bokmål og i nynorsk. Det er lett for oss å plukke ut adjektiv på -sk som er ubøyd i intetkjønn (H A 4 SK NORSK), og lignende adjektiv som har -t i intetkjønn. (Slike adjektiv er blitt regulert i 1982. I arbeidet som gikk forut for rettskrivningsvedtakene, hadde sekretariatet i Språkrådet ved hjelp av ordregistrantmaterialet laget oversikter med sammenligning av bøyningsformene i bokmål og nynorsk.)

Før den nye bokmålsrettskrivningen av 1981 ble vedtatt og satt ut i livet, hadde et særutvalg arbeidet i flere år. De punktene i rettskrivningen som særutvalget særlig arbeidet med, var *bøyningen av hunkjønnsord* og *bøyningen av svake verb*. Dette var de mest omfattende punktene. Det forelå mange forslag til oppmyking av de gjeldende reglene for at tradisjonelle former (f.eks. bok**en**, døre**n**; dekk**et**, eie**t**) igjen skulle bli en del av offisiell rettskrivning. Særutvalget måtte ta standpunkt til et stort antall ord der det forelå forslag om endringer.

Men det var rimelig å se disse ordene i en sammenheng og å sammenligne med systemet i nynorsk. Her kom ordregistranten til nytte under det forberedende arbeidet. Vi kunne få lister over alle hunkjønnsord som på bokmål hadde obligatorisk a-ending. Disse listene fikk vi både forlengs og baklengs alfabetisert. Listene viste seg å være et godt utgangspunkt for videre forberedende arbeid. Vi fikk fanget opp ord som burde behandles i sammenheng med ord det allerede var kommet forslag om at særutvalget skulle behandle. Derved fikk en en viss sikkerhet for at det ikke ble gjort endringer som førte til urimelige systembrudd.

Utskriftene av de svake verbene i registranten ble svært omfattende. Vi ønsket å se verbene både i lys av fonologiske strukturtyper og i sammenheng med bøyningssystemet i bokmål, nynorsk og tradisjonelt riksmål. Vi bestilte utskrifter av alle svake verb i bokmål

a) ordnet i knipper etter bøyningstype (forlengs og baklengs)
b) ordnet i typer etter fonologisk struktur (baklengs)

Siden ordregistranten er en registrant over *offisielt* norsk språk, gav den ikke opplysninger om bøyningsformene i andre varianter av norsk. Så det måtte gjøres mye sammenlignende arbeid på tradisjonell måte også, men listeproduktene fra registranten var et praktisk utgangspunkt for arbeidet.

I arbeidet med verbene kunne vi sammenholde bøyningstyper med strukturtyper. Som eksempel brukes her en liten undersøkelse av strukturtype 3. Denne strukturtypen var verb med stammestruktur på vokal + konsonant (f.eks. *bake*, *bone*). I registranten omfattet denne strukturtypen 2168 verb. Av disse var 1033 ere-verb (fremmedord på -ere, f.eks. *dominere*). Ere-verbene var ikke noe problem, så dem sjaltet vi ut. Av de verbene som stod igjen, hørte 46,5% til bøyningstypen med -te i preteritum (hvis en også hadde regnet med ere-verbene, ville prosenten blitt over 70). Ellers fordelte verbene seg med 28,7% på «a-verb» (verb med bøyningen enten -a/-et, -a eller -et), 16,9% på verb med «blandet» bøyning -a/-te, 3,3% på verb med «blandet» bøyning -a/-de og 3,5% på verb med -de i preteritum.

Det var ikke klart at stammestrukturen v + k betinget preteritum på -te. De listene vi hadde, satte oss i stand til å studere verbene med hensyn til hvilken konsonant stammen endte på. Ved en slik finfordeling av denne verbgruppen kunne en få øye på verb som tydelig skilte seg ut fra de verbene de hadde fonologisk likhet med. Verb som slik skilte seg ut (ved at de hadde en annen bøyningstype enn den som var vanlig ved slik struktur), var det rimelig å vurdere nærmere.

De strukturtypene som hadde best samsvar med en bøyningstype, var typene der stammen hadde to ulike konsonanter (f.eks. *knikse*) eller to like konsonanter (f.eks. *kjemme*). Der både stammestrukturen, den vanlige praksis i bokmåls-/riksmålstradisjonen og systemet i nynorsk pekte i samme retning, var det naturlig å foreslå regulering av bokmålsformer som skilte seg ut. Av *St.meld. nr. 100 (1980-81) Endringer i*

rettskrivningen og læreboknormalen for bokmål kan en se at mange av de verbene som ble regulert,hørte til disse strukturtypene (f.eks. *banne, brekke, dekke, ense, knikse, krinse*).

Planen om å få utarbeidet offisielle ordlister for begge språkformer på grunnlag av ordregistranten er ennå ikke realisert. I årenes løp har det vist seg at arbeidet med å lage en helt riktig ordregistrant i edb-format reiser mange språklige spørsmål som må løses først. Det kreves fullstendige og entydige opplysninger om alle ordenes skrivemåte og bøyningsformer. I vanlige ordlister er det med mange sammensatte ord uten at det gis fullstendige grammatiske opplysninger. Registrantens kodesystem tvinger frem at mange ord som aldri tidligere har vært vurdert av noe normeringsorgan, blir vurdert. En kan trygt si at det møysommelige arbeidet med ordregistranten er blitt en del av Språkrådets kontinuerlige språkrøkts- og normeringsarbeid.

Avismateriale

Fra PDS har vi fatt lister som bygger på et større avismateriale. De listene vi har, er:

Norsk grunnvokabular (Norske språkdata, nr. 1)
Frekvenssortert utskrift av de 10.000 vanligste ordene
Fullstendig ordliste - forlengs alfabetisk
Fullstendig ordliste - baklengs alfabetisk

Avismaterialet omfatter ca. 1,1 mill. ord, og det er fra perioden 1968-1973. Oslo-aviser, provinsaviser og NTB er med. I alle de listene vi har, er hvert ord forsynt med et tall som opplyser hvor mange ganger ordet forekommer i avismaterialet.

Dette materialet er nyttig for oss på en annen måte enn listene fra ordregistranten er det. Ved normering er det mange momenter det er naturlig å trekke frem. Hensynet til fonologisk struktur og grammatisk system er bare ett. En kommer aldri utenom hensynet til faktisk språkbruk, usus.

Ved ethvert rettskrivningsspørsmål som kommer opp til behandling, forsøker vi å skaffe opplysninger om hvordan ordet faktisk blir behandlet. Takket være avismaterialet har vi i noen tilfeller kunnet finne sikre opplysninger om dette. (Dette gjelder ved svært *vanlige* ord, de fleste ord forekommer få ganger, selv i et så stort materiale.)

Da Norsk språkråd vurderte en revisjon av bokmålsrettskrivningen, utarbeidet vi noen oversikter på grunnlag av avismateriale. Oversiktene skulle vise hvilke former av hunkjønnsord og av noen enkeltord som var de vanligste i avismaterialet. Oversiktene er fra 1978, og bygger på Dagbladet 1950, aviser fra 1968 (Aftenposten, Dagbladet, Morgenbladet) og Adresseavisen 1973. Noe av dette materialet var identisk med det avismaterialet PDS hadde, noe skrev seg fra NAVFs EDB-senter for humanistisk forskning. Avisundersøkelsen bød egentlig ikke på de store overraskelsene. Den viste at hunkjønnsord med -en i bestemt form både

forekom flere ganger og hadde bedre spredning i materialet enn de tilsvarende ordene på -a. Allikevel kunne undersøkelsen ha en viss interesse, fordi den viste *hvilke* ord som av og til forekom med -a, og hvilke som aldri gjorde det. Undersøkelsen demonstrerte at mange enkeltord som hadde hatt en «radikal» form i offisiell rettskrivning siden 1938, praktisk talt ikke ble brukt i denne formen. Det kunne gjelde ord som *dokke, sløkke, vogge, mjøl, lyge.*

Skjønnlitterært materiale

Språkrådet har satt i gang registrering av et større skjønnlitterært materiale for å kunne undersøke hvordan rettskrivningsreformene kan ha virket inn på forfatternes språkbruk. Materialet er valgt ut slik at det også skal kunne si noe om utviklingstendenser i språkbruken. De utvalgte tekstene blir overført til maskinleselig form av Norsk tekstarkiv. Overføringen begynte i 1981. Vi skal tilrettelegge et like stort materiale for hver av språkformene og har nå tilrettelagt bokmålsmaterialet. Nynorskmaterialet er også påbegynt, og vi har snart tilrettelagt 1/3 av det.

Språkrådet vedtok at det skulle legges *hele* tekster til grunn for undersøkelsen. Dette var ønskelig av flere grunner. En praktisk grunn var at Norsk tekstarkiv skulle kunne innlemme tekstene i sitt permanente arkiv, og der skal det registreres bare komplette tekster. De hele tekstene vil senere kunne benyttes til enkeltstudier av en forfatter eller av en roman. Vedtaket om å bruke hele tekster begrenser naturligvis antallet tekster. En kunne skaffet seg et bredere, mer representativt tekstgrunnlag dersom en hadde valgt en korpusmodell.

Vi har registrert 10 romaner på bokmål fra hvert av årene 1937, 1957 og 1977 og er i gang med registrering av et tilsvarende nynorskmateriale. Utvalgskriteriene er mekaniske. Utvalget fra 1937 skal være av forfattere født etter 1896, det fra 1957 skal være av forfattere født etter 1916, og det fra 1977 skal være av forfattere født etter 1936. Forfatterne skal altså høre til forskjellige «generasjoner»; ingen av de registrerte forfatterne var over 40 år da de gav ut det registrerte verket.

Romanene skal ha et sidetall på mellom 100 og 300 sider.

De verkene vi registrerer, er valgt ut fortløpende fra Norsk Bokfortegnelse blant de verkene som tilfredsstiller våre kriterier.

Det viste seg at utvalget av 10 romaner gav en nesten fullstendig dekning av bokmålsromaner som tilfredsstilte våre utvalgskriterier, når det gjaldt året 1957. Det året kom det ut få romaner. For årene 1937 og 1977 ble ikke dekningen så bred. Siden det blir gitt ut mindre på nynorsk enn på bokmål, får vi også bredde over det vi registrerer på nynorsk. Vi er der i gang med registreringen av bøker fra omkring 1977. Det viste seg at vi måtte lempe på utvalgskriteriene for å få et stort nok tekstgrunnlag. Så nynorskutvalget fra den yngste «generasjonen» blir romaner *og noveller* utgitt *1976-1980* av forfattere født etter 1936.

Før registreringen av et verk har vi innhentet skriftlig tillatelse fra forfatteren eller rettighetshaveren.

Når akkurat årene 1937, 1957 og 1977 er valgt, har det sammenheng med tidspunktet for større rettskrivningsendringer. Vi vil for eksempel prøve å få et bilde av hvordan rettskrivningsreformen av 1938 hadde slått igjennom i språket til en gruppe forfattere som var unge da 1938-rettskrivningen kom, og som gav ut bøker i 1957. (Det har vist seg at forfatterne av 1957-tekstene er født mellom 1918 og 1932. Fem av dem gikk på folkeskolen ennå i 1938.) Og vi kan få rede på hvor den offisielle normeringen har stått i forhold til samtidig litteraturspråk. Det har videre normeringsteoretisk interesse å finne ut hvilke typer av rettskrivningsendringer som lett slår igjennom, og hvilke typer som har liten gjennomslagskraft.

For at den planmessige undersøkelsen av materialet skal kunne komme i gang, trengs det konsentrert innsats av én eller flere personer, og materialet må tilrettelegges i form av indekslister. Til undersøkelse av syntaks og fraseologi må det lages konkordanser av ulike slag.

Hittil har vi fatt ni listeprodukter på grunnlag av bokmålsmaterialet. Det er forlengs alfabetisk, baklengs alfabetisk og frekvensordnet liste fra hvert av årene. Disse listene har vi allerede nå nytte av når vi behandler skrivemåte eller bøyning av enkeltord. Vi bruker dem på samme måte som listene på grunnlag av avismaterialet. I og med disse listene har vi fått et bredere grunnlag enn tidligere for å si noe om språkbruken når det gjelder enkeltord. Dessuten kan listene fra de tre årene fortelle oss noe om utviklingstendenser i ordbruken.

Jeg vil gi noen eksempler på faktisk og mulig bruk av listeproduktene.

I fjor drøftet fagnemnda i Språkrådet ordet *dobbel(t)*. Etter godkjente ordlister er det valgfritt om en skriver *dobbel* eller *dobbelt*, men som førsteledd i sammensetninger skal det alltid brukes *dobbelt-*. Fagnemnda vurderte om det burde være valgfritt med og uten *t* også når ordet var førsteledd i sammensetninger. (Saken endte med at gjeldende praksis ikke ble endret.) Da denne saken ble lagt frem, kunne vi lettvint skaffe oss rede på hvordan bruken av ordet som førsteledd var. Vi fant at avismaterialet hadde 25 ord med *dobbelt-* som førsteledd, fem med førsteleddet *dobbel-*. I bokmålsromanene fra 1977 fant vi bare *dobbelt-* som førsteledd (20 ganger).

I forbindelse med etterarbeidet etter den nye bokmålsrettskrivningen drøftet bokmålsdelen av fagnemnda om det var mulig å redusere valgfriheten innenfor rettskrivning og læreboknormal ved å sjalte ut lite brukte varianter. Vi hadde godt grunnlag for å si noe om hva som var mye eller lite brukt. Det kunne også synes praktisk om enkelte sideformer som er bare rent ortografiske varianter, kunne utgå. (Det kom lite ut av disse drøftingene, da mange av de sideformene det gjaldt, hadde språkpolitisk betydning for flere medlemmer av Språkrådet. Det viste seg å være liten vilje til å innskrenke valgfriheten dersom dette

gikk ut over tilnærmingsformer til nynorsk eller talemålsformer).

I språkrådssekretariatet var vi inne på tanken om formene [*ba*], [*ga*], [*sto*] og [*dro*] kunne tas ut av rettskrivningen. Formene med konsonant (*bad*, *gav*, *stod* og *drog*) hadde tradisjonen på sin side, og de var felles for bokmål og nynorsk, dansk og svensk. Disse preteritumsformene har hatt en litt omskiftelig rettskrivningstilværelse. Konsonantformene var eneformer frem til 1938-rettskrivningen. I nynorsk ble da [*sto*] innført som sideform, ellers ble det ikke endret noe. Men i bokmål ble *ba*, *dro*, *ga* og *sto* gjort til sidestilte former. Ved rettskrivningsendringen av 1959 ble bokmål endret slik at de konsonantløse formene ble redusert til sideformer.

Vi undersøkte i avismaterialet og skjønnlitteraturen hvordan formene ble brukt. Resultatet var slik:

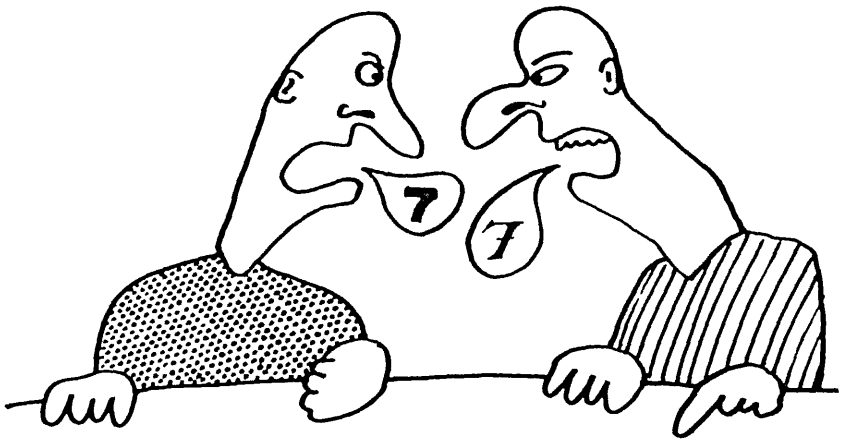
	ba	**bad**	**dro**	**drog**	**ga**	**gav**	**sto**	**stod**
1937	0	95	0	69	0	175	0	617
1957	49	38	96	15	175	59	442	305
1977	137	16	95	2	186	34	371	186
Avis	48	59	42	11	153	131	152	112

Vi ser at de formene som ble innført med rettskrivningen av 1938 ikke hadde grunnlag i skjønnlitterært språk. Men disse lydrette formene ble allikevel tatt i utstrakt bruk, ser vi av romanmaterialet fra 1957.

At de i 1959 ble redusert til sideformer, viste seg ikke å ha noen virkning. Tallene fra 1977 og fra avismaterialet viser at de konsonantløse formene brukes mer enn de tradisjonelle med konsonant. At tendensen virkelig er slik, ble vi enda mer overbevist om ved å se litt nærmere på datagrunnlaget. Nesten alle *bad*, *gav* og *drog* skriver seg fra én og samme forfatter av 1977-generasjonen, og formen *stod* skriver seg hovedsakelig fra to av de ti forfatterne. Tilbaketoget fra normeringsorganets side i 1959 hadde altså ikke greid å snu retningen. (Det er ikke nå gjort endringer, så formene med konsonant er fremdeles tillatt innenfor offisiell rettskrivning.)

I forbindelse med vår rådgivning om språkbruk har vi kunnet skaffe oss enkelte opplysninger om flerordsuttrykk. Dette kan vi ikke gjøre ved hjelp av de listeproduktene vi har til rådighet. Men våre kontaktpersoner ved PDS og Norsk tekstarkiv har ved et par anledninger gått inn i tekstmaterialet og skaffet oss noen minikonkordanser. Vi var for eksempel interessert i å se hvordan ordene *innen* og *innenfor* ble brukt. Materialet vi fikk ut, var ikke så stort, men det viste tydelig at det ikke er vanlig å opprettholde noe fast semantisk skille mellom disse ordene. Etter å ha vurdert dette materialet fant vi det urealistisk å prøve å få gjennomført *innenfor* i stedsbetydning og *innen* i tidsbetydning.

Det har vært hevdet at «den nye tellemåten» er av de språkendringer som ikke har slått igjennom. Dette kan vi si noe om ved hjelp av de



Tegning: Øystein Reigem

dataordlistene vi har. Samtidig får vi opplysninger om bruken av visse tallord.

Først litt historikk:

I 1951 ble det innført «ny uttale og skrivemåte av visse talord» ved et rundskriv fra Kirke- og undervisningsdepartementet.

Etter dette skulle tallordene nå uttales slik at en sa tierne før enerne (*femtifire* istedenfor *fireogfemti*). Og formene *sju*, *tjue*, *tretti*, *firti* og *sytti* skulle være eneformer i begge mål. (Med et nytt rundskriv ble *førti* satt inn istedenfor *firti*.)

Tallordet *sju* ble tillatt ved siden av *syv* allerede i 1917. I 1938 ble *sju* eneform.

Tyve er den tradisjonelle formen, men i 1917 kom formen *tjuge* inn i bokmål ved siden av *tyve*. I 1938 ble *tjue* eneform. Men i rettskrivningsreglene fra 1938 står det også at *syv* og *tyve* ikke skulle «regnes for feil i skriftlig arbeid i skolen». (Dette står i en parentes, og det ser ut til å ha vært en sovende regel. I skolen har *sju* og *tjue* vært vurdert som det eneste riktige.) Tallordet for 30 hadde fra gammelt av formen *tredive*. Ved rettskrivningsendringen i 1917 ble formene *tredve* og *tretti* innført isteden. *Tredve* gikk ut da den nye tellemåten ble innført i 1951.

For tallet 40 finner en i *Norske retskrivnings-regler 1907 fyrti, fyrretyve* og *firti = fyrti*. I 1917 ble oppføringen *firti el. fyrti (førr)*. I 1938 ble det *førti el. førr*.

Da vi undersøkte bruken av tallordene og praktiseringen av tellemåten i romanmaterialet vårt, fikk vi dette resultatet:

	syv/sju	tyve/tjue	tredve/tretti	firti/førti	gammel/ny t.
1937	30 2	54 -	33 -	12 1	
1957	32 7	39 1	16 4	- 8	20 1
1977	27 14	38 13	19 9	- 9	37 8

(Det ene eksempelet på «ny tellemåte» i materialet fra 1957, var «tjuett».)

Vi ser på enda et par ord som er enkle å identifisere:

I 1959 ble rettskrivningen endret slik at det skulle skrives *plassere* (mot før *plasere*. Den stumme h-en ble tatt bort i noen ord (bl.a. ble *hverken* endret til *verken*).

Vi prøver å skaffe oss et bilde av hvordan disse skrivemåtene har slått igjennom ved å se på ordlistene fra avismaterialet og fra bokmålsromanene fra 1977:

AVIS:	hverken 101 - verken 33	plasere 96 - plassere 81
1977:	hverken 42 - verken 10	plasere 20 - plassere 49

Det er klart at *hverken* har overlevd vedtaket og lever i beste velgående. (Siden vi her bygger på nakne ordlister – ikke konkordanser – må vi dessuten ta forbehold om at ordet *verken* kan være substantivet *verk* i bestemt form eller substantivet *verken*.) Verbet *plassere* med to s-er har slått bedre an, ser det ut til. Vi legger også merke til at de unge forfatterne brukte denne skrivemåten mer enn avisskribentene. Men også i avisene er *plassere* blitt tatt i bruk.

Noe av årssaken til at disse ortografiske formene har slått ulikt an, kan ligge i måten de er blitt behandlet på i Riksmålsordlisten; der er *hverken* og *plassere* nå oppført som eneformer.

*

Disse små undersøkelsene jeg har nevnt her, er ment å vise at det materialet vi har fått tilrettelagt og fått listeprodukter av, allerede nå gir oss mulighet for å oppfylle noen av de ønskene som lå til grunn da språkundersøkelsene ble satt i gang. Når vårt nynorske skjønnlitterære materiale er ferdig overført til datamedium, vil vi ha et godt utgangspunkt for å undersøke utviklingen av ordforrådet i nynorsk. Vi vil også kunne si noe om frekvensen av visse ordformer i moderne nynorsk, men dette vil vi da allerede vite en del om på grunnlag av den nynorske frekvensordlisten som blir utarbeidet ved Norsk tekstarkiv.

Som det går frem av det foregående, har vi i Språkrådet tatt resultater av datateknologien i bruk i vårt daglige arbeid. Vi har selv datautstyr, men får listeprodukter fra våre samarbeidspartnere (PDS og Norsk tekstarkiv), som har ansvaret for den tekniske siden. *Vi* er ikke edb-eksperter, men vi har skaffet oss så pass rede på edb-behandling av språkmateriale at vi er i stand til å ha en realistisk oppfatning av hva vi kan få svar på ved hjelp av edb.

3. Nordiske forum for edb-bibliotekarer: Automatisk indeksering

Øystein Reigem

Det 3. nordiske forum for edb-bibliotekarer ble holdt på Fjellhvil Hotell, Norefjell i tiden 26.-28.9.83. Det sentrale temaet på seminaret var automatisk indeksering og andre metoder for forbedring av maskinell søking i bibliografiske data. Til å forelese om dette hadde man hentet en av de ledende personer på feltet, nemlig professor *Gerard Salton* fra Cornell University i Ithaca i staten New York.

Indeksering

I bibliotekene snakker man om klassifisering og indeksering. Klassifisering vil si å plassere et objekt (bok, tidsskrift, rapport, osv.) i en kategori som karakteriserer innholdet. Det finnes omfattende og kompliserte klassifikasjonssystemer hvor man kan spesifisere emner i stor detalj. Mange har hørt om Deweys desimalklassifikasjon og UDK-systemet. Et problem med slike systemer er at de er én-dimensjonale. Verdens kunnskap er delt opp i emner, som igjen er delt i underemner, osv. Imidlertid må det nødvendigvis finnes emner som går på tvers av denne inndelingen. I Dewey er f.eks. emnet edb spredt utover hele systemet. Ofte vil også en publikasjon krysse flere emneområder. Kodene som blir brukt for hvert emne, er i tillegg uleselige for legfolk.

Disse problemene bøter en på ved indeksering. Indeksering vil si å gi en publikasjon ett eller flere emneord, dvs. termer som karakteriserer innholdet. Ved å tillate flere emneord for en publikasjon, løser man problemet med én-dimensjonaliteten. Emneordene - eller indekseringstermene - kan enten velges fritt eller tas fra et fast vokabular. Et fast vokabular kan enten være «flatt», dvs. at alle termene er uavhengige og likeverdige, eller det kan ha en struktur. Et strukturert vokabular kalles en tesaurus. Struktur vil i dette tilfellet si at det er relasjoner mellom termene. Termer kan være synonyme eller beslektede, og spesifikke termer er underordnet mer generelle termer. Det finnes også sinnrike systemer som gir regler for hvordan indekstermer for en publikasjon kan settes sammen til setningsaktige konstruksjoner der rekkefølge og spesielle funksjonsord gir viktige opplysninger om hver enkelt terms rolle. PRECIS er et slikt system.

Automatisk indeksering

Det er altså lagt mye arbeid i å lage gode manuelle indekseringssystemer. Ifølge Salton egner imidlertid ikke mennesker seg til indeksering. For å gjøre et godt indekseringsarbeid, må du nemlig

- ha god innsikt i emnene publikasjonene omhandler
- kjenne brukerne av samlingen
- kunne spå om framtidig utvikling innen emnene
- kunne spå om framtidige brukerønsker
- ha god kunnskap om indekseringssystemet og vokabularet indekstermene hentes fra
- være i god og jevn form for å levere et konsistent arbeid
- ikke skifte filosofi over tid
- ha samme filosofi og tankegang som de andre indekserne som jobber med samme materiale

En indekserer bør altså ha overmenneskelige evner. Alternativet er da automatisk indeksering.

Hvordan foregår så automatisk indeksering? Utgangspunktet for en automatisk indeksering er vanligvis utdrag eller sammendrag (abstracts) av de aktuelle dokumentene (publikasjonene). Et utdrag kan være en samling avsnitt som man mener står sentralt i dokumentet. Dersom dokumentene er korte (f.eks. avisartikler, gjenstandsbeskrivelser), brukes hele teksten. Et program trekker automatisk ut indekstermer av dette materialet. Indekstermene brukes så av et søkesystem. Her er en strategi for automatisk indeksering slik den er skissert av Salton:

1. Plukk ut de enkelte ordene fra teksten i dokumentene. Sil bort funksjonsord som konjunksjoner, preposisjoner, osv. De vil ha liten verdi for søkingen og krever mye ekstra plass. Silingen foretas best ut fra en på forhånd oppstilt liste av ord. Slike ord kalles stoppord. (Akkurat dette gjør de fleste tekstsøkesystemer i dag. Da blir på sett og vis alle ordene unntatt stoppordene indekstermer.)

2. Reduser de enkelte ordene til en rotform. Poenget er å få slått sammen bøyningsformer og avledninger slik at brukeren slipper å tenke på disse, samt å få en riktigere evaluering av hver terms (rotforms) viktighet (punkt 3). Selv ved å bruke metoder som betrakter hvert ord for seg kan man få til en god automatisk rotlemmatisering (se Tove Fjeldvigs og Anne Goldens artikkel i dette bladet). For bedre resultater må man gjerne trekke inn en syntaktisk/semantisk analyse av teksten. Men merk at det ikke er viktig at ordene blir redusert til en grammatikalsk korrekt rot. Det essensielle er at ord som hører sammen får samme rot og ord som ikke hører sammen får forskjellig rot. Indekstermene skal nemlig bare brukes internt i systemet og ikke vises til brukerne. (I et vanlig tekstsøkesystem takles som regel problemet med bøyninger og avledninger ved å tillate brukeren å søke på *starten* av ord. Eksempel: En kan søke på alle ord som begynner med NOR og dermed få med både NORGE, NORGES, NORSK, NORDMANN, NORDMENN osv., men dessverre også NORD-TRØNDELAG, NORDISK, NORMAL, NORMERT osv.)

Senere, under søking i det indekserte materialet, må ordene i brukernes spørsmål kunne sammenlignes med indekstermene. De må derfor i sin tur gjennom den samme reduksjon til rotform som dokumentteksten.

3. Finn ut hvor gode de enkelte termene (rotformene) er som indekstermer. Dette gjøres ved enkle opptellinger av hyppighet og spredning. At en term forekommer hyppig i et dokument, kan være et tegn på at den karakteriserer innholdet godt, og dermed er en god indeksterm. En term som forekommer i mange dokumenter, er sannsynligvis for generell som indeksterm eller har lite med innholdet å gjøre. På grunnlag av slike betraktninger har man kommet fram til en rekke formler som tilordner de enkelte termene vekt etter antatt betydning. Behold de beste termene etter beregninger ut fra en slik formel.

4. De beste termene ut fra en slik beregning vil i mange anvendelser være de som forekommer middels ofte totalt sett. De er sannsynligvis verken for generelle eller for spesielle. De lavfrekvente termene vil ofte være for spesifikke til å være særlig nyttige. (Vanligvis vil halvparten av de forskjellige ordene i et materiale forekomme bare én gang.) Men istedenfor å kutte disse termene helt ut, kan man forbedre dem ved å slå dem sammen i tesaurusklasser. Dette kan også gjøres automatisk. To termer anses da for å være beslektede dersom de har en tendens til å forekomme i de samme dokumentene. Ved søking på en term, trekkes da de beslektede termene også inn i søkingen.

5. De dårligste termene for indeksering vil vanligvis finnes blant de høyfrekvente, nemlig ord som forekommer jevnt i hele materialet. Disse termene kan imidlertid også forbedres. Det gjøres ved å slå dem sammen til fraser. Det vi vanligvis regner som en frase, er påfølgende ord, men i denne sammenhengen kan det også være nyttig å regne ord som står innen en viss avstand fra hverandre eller til og med forekommer i samme setning, som fraser.

Det er klart at et språk som engelsk med sine frasekonstruksjoner vil ha mer nytte av frase-strategien i punkt 5 enn f.eks. norsk, der sammensatte ord er mer vanlig. For norsk vil det under punkt 2 i tillegg til en rotlemmatisering være nyttig med en oppsplitting av sammensatte ord i enkeltord.

Dette kan virke som en lite sofistikert måte å indeksere på – en slags teknikkens okse i åndens glassbutikk. Ifølge Salton viser imidlertid eksperimenter med søking i samlinger av vitenskapelige artikler at automatisk indeksering er manuell indeksering totalt overlegen. I tillegg til de nevnte problemene ved manuell indeksering, forklarer Salton dette ut fra at automatisk indeksering gir et mye rikere utvalg av indekstermer. En manuelt indeksert publikasjon vil typisk ha 4-5, i

beste fall 10 indekstermer, mens antallet for en automatisk indeksert publikasjon ofte er i størrelsesorden 100.

Forbedring av selve søkingen

I tillegg til metoder for automatisk indeksering, har Salton strategier for forbedring av søking, som også kan benyttes på konvensjonelle tekst-søkesystemer. Konvensjonelle systemer benytter seg vanligvis av såkalt boolesk søking, dvs. at spørsmål stilles som booleske (eller logiske) uttrykk. Eksempel: Spørsmålet «BAKSTEHELLE *og* (STEIN *eller* SKIFER)» gir som svar alle dokumenter som inneholder ordet BAKSTEHELLE og samtidig minst ett av ordene STEIN eller SKIFER. Slike booleske spørsmål kan lett få svar som ikke er helt i samsvar med vår intuitive oppfatning av saken. Et eksempel er spørsmål på formen «A *eller* B *eller* C *eller* ... *eller* N». I svaret fra systemet behandles et dokument som inneholder alle de søkte termene (og som dermed sannsynligvis er meget relevant), likt med et dokument som bare inneholder én eneste av termene (og som dermed kanskje bare er «støy»). Tilsvarende vil systemet for spørsmål av typen «A *og* B *og* C *og* ... *og* N» forkaste dokumenter som inneholder alle termene på én nær, på lik linje med dokumenter som ikke inneholder noen av termene.

Det har lenge eksistert søkestrategier som takler disse problemene bedre enn boolesk søking. Salton har arbeidet mye med en av disse, såkalt «vektorsøking», men kunne nå presentere en metode som er enda et hakk bedre. En kan si at de booleske *og*- og *eller*-operatorene fra konvensjonelle systemer er myket opp. De oppmykede operatorene gir en rangering av dokumentene slik at det ikke bare blir et enten-eller som ved boolesk søking. Faktisk gir metoden oss en hel skala av mer eller mindre oppmykede *og*- og *eller*-operatorer som strekker seg fra ren boolesk søking til vektorsøking. I praksis kunne en da velge ut en boolesk og en oppmyket versjon av *og* og *eller*, eventuelt bare bruke oppmykede operatorer. Følgende vesle tabell viser bruksområdene for de forskjellige operatorene:

operator:	bruksområde:
boolesk *eller*	strengt synonyme termer. Tilslag på én term er like bra som tilslag på alle.
oppmyket *eller*	beslektede termer. Tilslag på flere er bedre enn tilslag på én.
boolesk *og*	obligatoriske termer. Må ha tilslag på alle.
oppmyket *og*	«tentative» termer. Tilslag på flest mulige.

En annen forbedring av søkingen får man ved å legge inn såkalt «relevans-feedback» i systemet. Det virker på følgende måte: Brukeren blir presentert for et lite antall dokumenter (f.eks. 10) som svar på spørsmålet sitt, og blir bedt om å klassifisere hvert enkelt som relevant eller ikke-relevant. Deretter undersøker systemet disse dokumentene for å se hvor godt spørsmålet passer. Spørsmålet blir så reformulert av systemet i samsvar med dette. Hvis det finnes termer som har en tendens til å forekomme bare i de relevante dokumentene, blir disse gitt mer vekt eller lagt til i det reformulerte spørsmålet. Søketermer som stort sett forekommer i de ikke-relevante dokumentene blir sløyfet. Spørsmålet blir så utført på nytt, og hele prosessen går om igjen til brukeren er fornøyd.

Eksperimenter der disse to metodene er kombinert, har vist til dels dramatisk forbedring av søkesystemets yteevne.

Automatisk indeksering i praksis

Til tross for lovende eksperimentresultater er automatisk indeksering ikke tatt i bruk i noe større system. Det er imidlertid ikke noen tekniske hindringer for dette. Problemene er på den menneskelige, organisasjonsmessige og miljømessige siden. Salton sier han finner det vanskelig å brolegge kløften mellom bibliotekarer og informatikere. En vanskelighet er det også at mye søking foregår med antikverte systemer som er vanskelige å endre. Og når det gjelder store on-line søkesystemer, kan det være store organisasjonsapparater bygd rundt dem, men det finnes lite personell på edb-siden. Salton mener dessuten at hans oppgave ikke strekker seg lenger enn til å verifisere og presentere sine teorier.

På seminaret ble det fra deltakerne ytret tvil om hvor god automatisk indeksering vil være på materiale av andre typer enn eksperimentmaterialene (f.eks. bøker) og på mindre homogent materiale. Salton mente at ytelsen i en del tilfelle vil svekkes, men at automatisk indeksering nesten alltid vil være en brukbar teknikk. Et viktig problem for bibliotekene i Norden i dag er også at de mangler maskinleselig grunnlag for automatisk indeksering. Salton hevdet imidlertid at det bare er et tidsspørsmål før maskinleselig tekst eller sammendrag blir tilgjengelige også her.

Referentens konklusjon

Det er viktig å komme i gang i Norge med å prøve ut metoder for automatisk indeksering og forbedret søking. Mange miljøer og anvendelser vil ha nytte av disse metodene, og slett ikke bare biblioteker!

Takk til Tove Fjeldvig for faglig korrekturlesning.

Litteratur

Gerard Salton and Michael J. McGill: Introduction to Modern Information Retrieval (McGraw-Hill 1983). (Dette er en god bok om gjenfinning, inkludert automatisk

indeksering. Den som vil vite noe om «oppmykede» operatorer, må imidlertid gå til en av de to andre referansene.)
G. Salton, E.A. Fox and N.H. Wu: Extended Boolean Information Retrieval (Technical Report TR82-511, Cornell University, Department of Computer Science, Ithaca, New York).
G. Salton, C. Buckley and E.A. Fox: Automatic Query Formulation in Information Retrieval (Technical Report TR82-524, Cornell University, Department of Computer Science, Ithaca, New York).

Systemutvikling: Informatikkens grense mot de «myke» fagene

Tone Bratteteig

Artikkelen omhandler min hovedoppgave «Kommunikasjon i systemutvikling» fra Institutt for informatikk, Universitetet i Oslo. Jeg starter med å beskrive informatikkfaget - og informatikeres perspektiv på virkeligheten. Videre gir jeg en kort beskrivelse av min (informatiske) forståelse av fag som omhandler kommunikasjonsteori. Arbeidet i hovedoppgaven er et forsøk på å kombinere disse fagområdene, og artikkelen avsluttes med noen av mine erfaringer fra dette arbeidet.

Artikkelen føyer seg inn i serien om tverrfaglighet - edb og humaniora - som har pågått i de siste numrene av HD.

*

Naturvitenskapene studerer forskjellige aspekter ved prosesser og fenomener i natur og samfunn:

- deres identifikasjon og egenskaper; d.v.s. fenomenologi,
- deres samspill med andre aspekter ved virkeligheten; d.v.s. grensesnittet med andre fag,
- hvordan de kan forstås og beskrives; d.v.s. analyse,
- hvordan de kan utformes, virkeliggjøres og omformes; d.v.s. syntese.

Informatikken legger vekt på informasjonsaspektene ved virkeligheten. I likhet med de andre naturvitenskapene, ligger hovedvekten i faget på syntese-delen. Informatikere arbeider med konstruksjon av informasjonsprosesser ved å foreskrive strukturen prosessene skal foregå under. De ulike retningene innenfor informatikkfaget skiller seg fra hverandre i typen informasjonsprosesser som konstrueres.

I likhet med fag som kjemi og fysikk har syntese-delen av informatikken stått sterkt fra fagets begynnelse: informatikkfaget er bygget omkring elektronisk databehandling som verktøy.

Et edb-system er en informasjonsprosess som foregår (harmonisk) underordnet en struktur. Begrensningene strukturen legger for prosessen, medfører at denne underordnede prosessen ses som vel avgrenset og regulerbar, og at rammene og reglene for den er klart definert. Informatikeres arbeid består i konstruksjonen av maskinvare og programvare som skal gjøre edb-systemet til et godt redskap i en informasjonsprosess.

Systemutviklingsdelen av informatikken fokuseres også på syntese, men skiller seg ut fra resten av faget ved å legge stor vekt på grensesnittet til samfunnsfagene og de humanistiske fagene. I systemutviklingsteorien omhandles edb-baserte systemer; også organisasjonen rundt edb-systemet tas i betraktning[1]. Systemutvikling ses følgelig som

en form for organisasjonsutvikling. En systemutviklingsprosess omfatter de forandringene som foregår i organisasjonen fra det oppstår ønske om forandring (eventuelt ved hjelp av innføring av et edb-system) til et ferdig konstruert edb-basert system er i drift i organisasjonen. Systemutviklingsprosessen er en styrt, målrettet prosess med formål å forandre organisasjonen slik at dens evne til å forfølge bestemte mål forbedres.

En konkret systemutviklingsprosess forløper innenfor rammer som settes av organisasjon og systemutviklingsmetode.

Det finnes et utall systemutviklingsmetoder, d.v.s. forskrifter for systemutviklingsprosessens utforming[2]. De fleste slike metoder legger hovedvekten på retningslinjer for arbeidet, og består ofte av en samling teknikker med tilhørende (systembeskrivelses-)verktøy. Metodens teknikker og verktøy begrenser ofte hva slags systemer som er mulige å beskrive. Konstruksjonen av systemet bygger på systembeskrivelsen(e), følgelig begrenses hva slags edb-baserte systemer det er mulig å lage.

Metodene foreskriver ofte også organiseringen av systemutviklingsarbeidet. De fleste metoder er laget med en prosjektorganisering for øye. I prosjektgruppen forutsettes det som regel at en eller flere «brukere» (eller representanter for slike) skal delta. Dette legger igjen føringer på hvordan beskrivelsesverktøyet og -teknikken utformes.

En metode bygger alltid på en rekke (ofte implisitte) antakelser om organisasjonen og systemutviklingens mål. De aller fleste metoder har et harmoniperspektiv på organisasjoner, og legger dette til grunn for sine forskrifter; de tar f.eks. ikke hensyn til (eller de tar avstand fra) å vurdere divergerende syn på edb-systemer og systemutvikling i organisasjonen. Metodene er med andre ord med på å bygge opp under myten om den nøytrale og verdifrie teknologien - og den objektive edb-ekspertisen.

Det er nødvendig med en systemutviklingsteori som er uavhengig av eksisterende metoder for systemutvikling. En slik teori må ta utgangspunkt i systemutviklingsprosessen som arbeidsprosess (se note 1). Med denne betraktningsmåten er det mulig å skille ut noen dimensjoner som kan benyttes til å karakterisere en systemutviklingsprosess. Hovedhensikten med en slik prosess er forandring av organisasjonen ved innføring/endring av dens edb-baserte systemer. Innenfor rammene av en systemutviklingsprosess vil det foregå en rekke delprosesser som angår undersøkelse, konstruksjon, forandring, beslutning og kommunikasjon. Disse dimensjonene kan også brukes til å karakterisere andre typer arbeidsprosesser: systemutviklingsprosessen kan karakteriseres ved innholdet i hver dimensjon.

Systemutviklingsteoriens tilnærming til kommunikasjon går gjennom en betraktning av systemutviklingsprosessen som arbeidsprosess. En slik prosess utføres av flere mennesker, ofte med en prosjektorganisering av arbeidet. Prosjektgruppene består av mennesker med forskjellig bakgrunn og kunnskaper. Det er nødvendig med kom-

munikasjon både internt i en prosjektgruppe, og mellom denne og andre deler av organisasjonen.

I systemutviklingsteorien har det vært gjort lite med hensyn til å presisere/spesifisere hva kommunikasjon er og bør være. Min hovedoppgave er ment å være et bidrag til dette.

Tradisjonelt brukes begrepet kommunikasjon i informatikkfaget i forbindelse med sammenkobling av datamaskiner i nettverk. Begrepet brukes også når det er snakk om grensesnittet mellom datamaskinen og de(n) som bruker den. I systemutvikling vil det også være viktig å omtale kommunikasjon mellom mennesker.

I arbeidet med presisering av mellommenneskelig kommunikasjon i systemutviklingen, var det nødvendig for meg å gå utover informatikkfagets grenser: jeg valgte å hente kunnskaper om kommunikasjon fra fag som har dette som hovedemne. Mitt valg falt på sosio- og psyko-lingvistikk. I det følgende skal jeg kort redegjøre for min forståelse av disse fagene.

Sosiolingvistikk er den delen av språkvitenskapen som beskjeftiger seg med studiet av sosiale forholds innvirkning på språk og språkbruk. Faget har som utgangspunkt kommunikasjonsmodeller der komponentene «sender«, «kanal» og «mottaker» inngår. I disse modellene er kommunikasjonen retningsbestemt med utgangspunkt i at senderen har til hensikt å formidle informasjon. Dette gjør hun ved å representere informasjonen ved hjelp av en kode: hun gjør informasjonen om til data. Dataene sendes gjennom kanalen. Mottakeren mottar dataene, og tolker dem om til informasjon. I sosiolingvistikken deles kommunikasjonsprosessen opp på grunnlag av de forskjellige variantene av kommunikasjonsmodellen

(SENDER ⟶ KANAL ⟶ MOTTAKER)

Forskjellen mellom kommunikasjonsmodellene ligger i egenskapene som tillegges komponentene, og i detaljeringen av disse. Komponentene beskrives separat, og det er først og fremst forskjellige egenskaper ved deltakerkomponentene som omhandles. Utgangspunktet for analysen er meldingene som produseres (d.v.s. kommunikasjonsprosessens resultat). Ved å sammenholde egenskaper ved meldingenes form og innhold med egenskaper ved deltakere og kanal, påpeker en sammenhenger mellom språk (og språkbruk) og samfunnet språket eksisterer i. Språk og språkbruk varierer med deltakerkomponentenes sosiale egenskaper (etnisk gruppe, yrkesgruppe, sosial klasse, alder, kjønn o.s.v.), og med egenskaper ved kanalene som brukes (skrevet og talt språk, kroppsspråk o.s.v.). Slike systematiske språkvariasjoner danner utgangspunkt for analysen av språk og språkbruk i et samfunn.

I psykolingvistikk legges hovedvekten på analysen av hver enkelt deltakers språkforståelse og språkbruk. Faget konsentrerer seg om inn- og av-kodingsprosessene hos den enkelte deltaker. I likhet med sosiolingvistikken bruker psykolingvistikken meldingene som grunnlag for analysen.

Tegning: Øystein Reigem.

Ut fra min forståelse av tradisjonell sosio- og psyko-lingvistikk synes det som om fagene definerer kommunikasjon som informasjonsutveksling - og informasjon som tolkede data. Informasjonsutvekslingen deles opp som sekvenser av meldingsoverføringer.

Dette harmonerer med den tradisjonelle informatiske bruken av disse begrepene, slik at begrepene fra sosio- og psyko-lingvistikk også kan brukes til å betegne kommunikasjon mellom datamaskiner og kommunikasjon mellom menneske og datamaskin.

Kommunikasjon definert som informasjonsutveksling lar seg også beskrive ved hjelp av (det informatiske) system-begrepet. En systembeskrivelse av en kommunikasjonsprosess vil fokusere på strukturen for den, og beskrive kommunikasjonen som en begrenset og regulert prosess.

Tradisjonell sosio- og psyko-lingvistikk lar seg med andre ord lett kombinere med tradisjonell informatikk. Årsaken til dette ligger i den sterke oppdelingen og forenklingen av kommunikasjonsbegrepet. En forenkling av kommunikasjon til informasjonsutveksling gjør at mange aspekter ved mellommenneskelige kommunikasjonsprosesser må utelates. Oppdelingen av kommunikasjonsprosessen i tilstander fører til at det blir mulig å omtale egenskaper ved kommunikasjonen som isolerte og løsrevet fra resten av prosessen.

*

Det er naturlig å avgrense kommunikasjonen i systemutviklingsprosessen ved å se på de rammer som settes av organisasjon og systemutviklingsmetode.

Det er også naturlig å avgrense emnet ved å begrense temaet for kommunikasjonen. Kommunikasjonen i systemutviklingsprosessen omhandler i første rekke denne prosessens resultat: det edb-baserte systemet, og konsentrerer seg om konstruksjonen av dette. Systemutviklingslitteraturen (inkludert min hovedoppgave) beskjeftiger seg i hovedsak med systemutviklingsmetodenes teknikker og verktøy som hjelpemidler under dette konstruksjonsarbeidet.

Analysen av systemutviklingsmetodenes teknikker og verktøy gjøres hovedsaklig med utgangspunkt i de manifeste meldingene som produseres: skriftlige systembeskrivelser. I dette arbeidet kan enkle kommunikasjonsmodeller være nyttige: en kan vurdere hvilke egenskaper som kreves hos deltakerne i systemutviklingsprosessen for at de skal kunne produsere meldinger/systembeskrivelser ved hjelp av de teknikker og verktøy metoden foreskriver.

I informatikkfaget er hovedvekten lagt på konstruksjon. Analysene av systemutviklingsmetodenes teknikker og verktøy blir regnet som nyttige først når de kan bidra til arbeidet med å konstruere nye - forhåpentligvis bedre - teknikker og verktøy for beskrivelse av edb-baserte systemer.

I systemutviklingsprosessen foregår det kommunikasjon på flere

nivåer. I kommunikasjon om det planlagte edb-baserte systemet benyttes formaliserte teknikker og verktøy. Kommunikasjon om systemutviklingsprosessen som arbeidsprosess gjør ikke det.

Kommunikasjonen i arbeidsprosessen angår samarbeidet mellom personer fra flere yrkesgrupper om konstruksjonen av et (påtenkt) edb-basert system. Kommunikasjonen angår både samarbeidsprosesser, læringsprosesser o.s.v. - og informasjonsutvekslingsprosesser. Kommunikasjonen *inngår i* flere forskjellige samhandlingsprosesser, samtidig som den *er* samhandling.

For å få en forståelse av kommunikasjonsprosessen må også situasjonen kommunikasjonen foregår i, tas i betraktning. Med dette utgangspunktet er det vanskelig å omtale deltakerne, kanalen(e) eller situasjonen separat, løsrevet fra hverandre. Der er f.eks. vanskelig å skille egenskaper hos kommunikasjonsprosessens deltakere som enten sosialt eller psykisk betinget. Det er også vanskelig å skille deltakernes (iboende) egenskaper fra deres reaksjoner på situasjonen som kommunikasjonsprosessen foregår i.

Enkelte (nyere) teorier innenfor sosio- og psyko-lingvistikk framhever nettopp helheten som nødvendig for forståelsen av kommunikasjonsprosessen. I denne forbindelse er det naturlig også å trekke inn teorier fra disse fagenes grenser til psykologi, sosiologi og sosialpsykologi.

*

I det foregående har jeg forsøkt å beskrive hvordan min forståelse av kommunikasjon forandret seg i arbeidet med hovedoppgaven. Fordi det er mulig å omtale kommunikasjon - og teorier om kommunikasjon - ved hjelp av mer «tekniske» termer, og gjennom et informatisk perspektiv, tok det en stund før jeg så begrensningene dette la på forståelsen av kommunikasjonsprosesser.

Kommunikasjon er en sosial prosess. Det er følgelig vanskelig å få oversikt over en kommunikasjonsprosess uten å bruke forenklede modeller av noen av aspektene ved den. Ved kategorisering og formalisering vil en kunne trekke ut de egenskapene ved prosessen som synes viktige. Innføringsbøkene i sosio- og psyko-lingvistikk bruker denne framgangsmåten (som de fleste andre innføringsbøker).

De mer tradisjonelle teoriene innen sosio- og psyko-lingvistikk ligger nærmere lingvistikk, og er mer preget av dette fagets tradisjon (der oppdeling og forenkling er vesentlig). Nyere teorier nærmer seg henholdsvis sosiologi og psykologi, og bygger på et perspektiv der helheten ved kommunikasjonsprosessen står mer sentralt. Et slikt perspektiv er klart vanskeligere å kombinere med informatikkteori.

En av de viktigste konklusjonene fra arbeidet med hovedoppgaven er at emnet kommunikasjon i systemutvikling ikke kan behandles innenfor et tradisjonelt informatikkperspektiv. Det er nødvendig å trekke inn andre fag for å få med alle viktige aspekter ved kommunikasjonen.

Kombinasjonen av fag kan være vanskelig å få til hvis fagenes perspektiv er forskjellige. Ethvert fag bygger på noen grunnleggende antakelser om virkeligheten - og spesielt om den delen av virkeligheten faget tar for seg. Dette perspektivet kommer til uttrykk gjennom fagets tilnærming til virkeligheten. Perspektivet legger ramme for hvordan ny viten (innen faget) kan etableres, representeres, struktureres o.s.v.

Enhver som er opplært innen en fagtradisjon vil adoptere dette fagets perspektiv som en måte å forstå og forholde seg til virkeligheten. For en person med ett valgt (fag-)perspektiv på verden, vil det være vanskelig å *samtidig* danne et helhetlig verdensbilde bygget på et annet fags grunnlag, i forhold til et emneområde.

Kombinasjonen av fag må derfor være utvidelser av et fags emneområde ved å tilføre flere aspekter ved dette, basert på andre fags perspektiver på emneområdet. Gjennom slik fagkombinasjon kan etterhvert også fagenes perspektiv utvides.

Kommunikasjonen i systemutviklingsprosessen (som arbeidsprosess) er et emneområde der det er nødvendig å utvide informatikken ved å trekke inn andre fags perspektiv. I de ulike nivåene av kommunikasjonen vil det være forskjellige fag som er nyttige. Både analysen av menneskers mentale prosesser og av kommunikasjonsprosesser, er nyttige for informatikeres arbeid med å konstruere retningslinjer for kommunikasjonen i systemutviklingen.

Et av problemene med denne fagkombinasjonen er informatikkens fokus på konstruksjon - formalisering, forenkling og oppdeling, står sterkt i fagets perspektiv. Andre fagområder vil særlig være til nytte i arbeidet med å avgrense hva det ikke er nyttig og/eller mulig å formalisere og foreskrive. Retningslinjene for kommunikasjonen i systemutviklingen kan vanskelig være det som tradisjonelt oppfattes som en systemutviklingsmetode.

Det er begrenset i hvor stor grad en systemutviklingsprosess lar seg foreskrive og regulere: den har uforutsigbar karakter. Det samme gjelder for kommunikasjonen i systemutviklingsprosessen.

Konstruksjonen av retningslinjer for denne kommunikasjonen kan følgelig ikke være strenge forskrifter, men må bestå av løsere formulerte retningslinjer til hjelp og understøttelse av samhandlingen i systemutviklingsprosessen. I dette arbeidet kan andre fags perspektiv på kommunikasjon gi viktige bidrag til informatikken.

Noter

1. Denne «skolen» innen systemutviklingsteorien bygger på arbeider av Kristen Nygaard (Universitetet i Oslo) og Lars Mathiassen (Universitetet i Århus).
2. En systemutviklingsmetode kan karakteriseres ved sitt
   - anvendelsesområde (hva den skal brukes til, hvem den skal brukes av)
   - perspektiv (hvilket «verdensbilde» den har innebygget)
   - retningslinjer for arbeidet; d.v.s.
     - prinsipper for organiseringen av arbeidet
     - teknikker
     - verktøy.

EDB OG HUMANIORA

Større presisjon ved bruk av edb og kvantitative metoder

Intervju: Rune Johansen

Roald Skarsten ble i 1979 ansatt som første medarbeider ved Edb-seksjonen, Det historisk-filosofiske fakultet, Universitetet i Bergen. Det har vært Skarstens oppgave å lede Edb-seksjonen i disse årene. Han kan i dag konstatere at seksjonens tjenester ikke bare brukes av stadig flere. Antall fagområder som benytter tjenestene, øker også. Sin edb-messige bakgrunn har Skarsten fra bruk av edb i eget forskningsarbeid og fra NAVFs EDB-senter for humanistisk forskning.

-Kan du i få ord presentere situasjonen i dag ved edb-seksjonen?

Edb-seksjonen er et rent fakultetstiltak. Hovedformålet er å yte edb-tjenester til fakultetets ansatte og til hovedfagstudenter. Siden starten i 1979 har det skjedd en rivende utvikling, og seksjonen har nå en oppdragsmasse som overstiger dens kapasitet. Én ansatt klarer dessverre ikke å dekke behovet for konsulentassistanse. Heldigvis ser det ut til at sider ved personalsituasjonen bedres i 1984. Det er all grunn til å tro at vi fra neste år av vil bli tilgodesett med to halve stillinger. En slik styrking av personalet vil gi bedre muligheter for å holde tritt med brukernes behov for øket edb-assistanse.

Jeg vil imidlertid understreke at fakultetet alt har tilgodesett edb-seksjonen med mye teknisk utstyr. Det er gledelig at fakultetet er innstilt på å utvikle edb-seksjonen som et redskap i forsknings- og utdannelsesarbeidet.

-Hva er formålet med tjenesten?

Våre oppgaver kan deles i fire hovedgrupper:

1. Gi konsulentassistanse både til ansatte og hovedfagstudenter
2. Sørge for at fakultetet får det edb-utstyr som passer best til oppgavene og drift av utstyret
3. Drive kursvirksomhet
4. Foreta metodisk utviklingsarbeid og programmering.

Et overordnet mål med det metodiske utviklingsarbeidet er å utvikle generelle programsystemer. Dette kan vi gjøre fordi det er de samme problem som går igjen innen flere fagfelt. Problemene er gjerne forbundet med innsamling, bearbeiding, sortering og utvelgelse av data. I tillegg kommer ofte statistisk analyse. En oppgave som må løses med tanke på den statistiske analyse av data er å få utviklet en program-

Roald Skarsten

pakke for språk- og litteraturforskere på samme måte som samfunnsfagene har sin SPSS-pakke. Dessverre er vi henvist til denne SPSS-pakken i dag, som innenfor vårt felt har sine klare svakheter.

-Ligger det ikke en fare i selve bruken av generelle programpakker? Kan edb-konsulenten komme til å tilpasse prosjektet til de oppgaver programpakken er utviklet for å løse?

Teoretisk er det en fare. Det er klart at jeg som edb-konsulent stadig må ha i minnet at styring kan skje. Foreløpig tror jeg at dette ikke har skjedd. Sikkerhetsmekanismen ligger i selve situasjonen i dag og i at jeg selv har min faglige bakgrunn fra fakultetet. Brukerne velger selv om de vil bruke edb. Det er ingen skoletvang. Jeg har så å si ett ben i hver leir. Jeg tror at dette bidrar til en fornuftig bruk av edb i de prosjekter som egner seg for bruk av dette verktøyet. Kommunikasjonsproblemene med dem som vil ha hjelp, er derfor ubetydelige.

Min bakgrunn gjør at jeg forholdsvis raskt kan sette meg inn i det enkelte forskningsprosjekt og skille ut de deler som måtte egne seg til edb-behandling. Enkelte ganger fraråder jeg bruk av edb fordi jeg tror at man vil vinne både tid og dypere innsikt ved å gå frem på tradisjonell måte, gjennom bruk av innfølingsevne, skarpsinn og refleksjon.

-Hvordan er gangen i konsulentbistanden når du gir grønt signal for bruk av edb?

Jeg redegjør for de praktiske problem som vil oppstå, hvilke maskinressurser det kan regnes med og hvilken type assistanse edb-seksjonen kan gi. Det er meget viktig at den enkelte selv blir klar over hvor mye tid han/hun må regne med å bruke på edb-siden. La meg understreke at edb-seksjonen ikke driver etter «postordreprinsippet». Seksjonen er et konsulterende organ. De som kommer hit må selv være villige til å bruke noe av sin tid på edb-siden.

-Du nevnte at edb-seksjonen er et rent fakultetstiltak. Hvilke muligheter har du for å styre utviklingen ved edb-seksjonen?

Jeg fungerer som et rådgivende organ. Beslutningsmyndigheten ligger helt og holdent hos fakultetet. Ut fra mitt synspunkt ville det være ønskelig at de som skulle avgjøre hvilke tiltak som skal settes ut i livet, hadde mer kunnskap om edb. Utdanningssektoren er i dag inne i en utvikling som krever at det handles raskt med tanke på bruk av edb. Det historisk-filosofiske fakultet har ennå ikke meislet ut en konkret strategi for edb-sektoren. Jeg skulle ønske at fakultetet snarest mulig kom frem til konkrete direktiver. Det er mange tiltak jeg kunne tenke meg å sette i verk. Behovet på brukersiden vokser, men handlingsrammen som er gitt pr. i dag, gjør det vanskelig å dekke disse behovene.

-Hvilke tiltak kunne du eventuelt ha tenkt deg å gjennomføre?

Etterutdanningstiltak for universitetets tidligere studenter står høyt på min prioriteringsliste. Den vanlige veien for kunnskap og opplysning har vært fra universitet til videregående skole, ungdomsskole og barneskole. Jeg synes det nå er klare indikasjoner på at universitetet er i ferd med å komme på etterskudd i forhold til utviklingen. Skal universitetet fortsatt være først i kunnskaps- og opplysningskjeden, må nytenkning og tilpasning til samfunnets behov prioriteres.

Et altfor stort antall studenter innenfor Det historisk-filosofiske fakultet, er eller står i fare for å bli arbeidsledige. Fakultetet må komme på offensiven vis-à-vis samfunnet. Studentene må gis mer selvtillit som filologer. Dette kan gjøres på flere måter, men én måte er å gjøre det gjennom edb. Se bare på edb-bruken i behandling av språklig informasjon. Vi står overfor en utvikling som kanskje vil ende med at 90% av informasjonsbehandlingen vil gjelde språklige data og bare 10% numeriske. Kunnskap om edb vil ikke bare høyne studentenes faglige nivå, den vil også gjøre filologene langt mer attraktive som arbeidstakere. Det er en kjent sak at det private næringsliv og den offentlige sektor har behov for filologer. Men slik situasjonen er i dag, prioriteres søkere med edb-utdannelse på bekostning av den tradisjonelle filologutdannelsen.

En mulighet er at fremmedspråkinstituttene går sammen om en egen amanuensisstilling for edb og fremmedspråk. Derved har man et utgangspunkt for å dekke de behov skolen og samfunnet har i dag. Et

annet optimistisk mål er at alle forskere på fakultetet skal ha tilgang til en mikromaskin. Denne vil med tiden bli like selvfølgelig som skrivemaskinen og telefonen.

-Ved Det historisk-filosofiske fakultet er eller har det vært en ganske sterk skepsis mot å bruke edb i forskningsarbeidet. Skjer det en holdningsendring på dette området?

Man skal være skeptisk til bruk av ny teknologi. Bare gjennom nøye kjennskap til den nye teknologien og en gjennomgripende refleksjon over de konsekvenser bruken av den fører med seg, kan man, så langt det er mulig, sikre seg mot ufornuftig bruk av det nye.

Til sine tider grunner imidlertid skepsisen og argumentasjonen mot bruk av edb på ren og skjær uvitenhet. Heldigvis begynner stadig flere å kunne noe om edb. Man vet hva edb er, hva edb kan og ikke kan. Følelsen av å kontrollere verktøyet bidrar selvsagt til at skepsisen mot å ta det i bruk avtar.

Alt nå er det klart at den raskt økende bruk av edb også har åpnet forskernes øyne for at edb i visse sammenhenger kan være et nyttig hjelpemiddel. Som et stikkord her vil jeg nevne tekstbehandling. Utstyr for tekstbehandling er lett å betjene, og man får raskt uttellinger i form av høyere effektivitet og bedre kvalitet.

Universitetsfolk må sette seg på skolebenken igjen. Det er i ferd med å skje en revolusjon. Om ti år vil alle studentene her på universitetet være kvalifiserte edb-brukere. De vil ikke finne seg i å bli diktert hvordan de skal arbeide. De kommer til å forlange å få bruke edb for å nå de resultater de vil frem til. Det sier seg selv at det vil bli vansker dersom ikke veilederne følger opp på dette feltet.

-Men dette betyr vel også at selve veilederrollen endres?

Studentene får raskere tilgang til opplysninger og bakgrunnsmateriale for sine hovedfagsoppgaver. Ergo får de bedre tid til refleksjon og fordypelse. Dette fører igjen til at selve spørsmålsstillingen i sammenheng med deres arbeid blir grundigere, problemkomplekset mer omfattende og kravene til dem som skal rettlede større. Denne utviklingen er meget positiv. Det er jo nettopp dette forskerne vil.

Stadig flere innser at man her står overfor oppgaver som må løses hvis man vil dempe den kommunikasjonskløft som vil oppstå under den nåværende situasjon. Antall lærere som ønsker hjelp og kunnskap om edb er økende. På dette området skjer det store endringer. I fjor forsøkte vi på fakultetet å arrangere et edb-kurs i forbindelse med «Vestlandske Lærerstemne». Kurset måtte avlyses på grunn av manglende oppslutning. I år tilbød vi et kurs «Edb for filologer». Dette kurset ble overtegnet.

-Gjelder dette de «samme» personer som man henvendte seg til året før?

Ja, og dette tolker jeg som et tegn på at det i løpet av dette året er

skjedd en holdningsendring. Man viser vilje til å sette seg inn i det nye redskapet.

-Hva mener du om filologens plass i morgendagens samfunn?

Filologene representerer en innsikt i menneskelige forhold som det er uhyre viktig å ta vare på i automatiseringens tidsalder. Jeg tror at filologene vil være dem som er best skikket til å sikre en balansert utvikling i samfunnet når det gjelder bruk av edb. Men dette forutsetter at de selv kan noe om edb. Det plager meg at man i praksis utdanner filologer til arbeidsledighet.

Jeg tror at informasjonsteknologien vil føre til at bedriftene vil få større behov for filologer som kan mer enn bare språk. De bør også kunne et minimum av edb-bruk. Edb-kunnskap kan bli nøkkelen for mange filologer til arbeidsmarkedet. Filologer med edb-kunnskap vil raskt kunne tilpasse seg bedriftens behov og bli dyktige arbeidstakere. Alle kan ikke bli forskere eller lærere.

Dette faktum må fakultetet ta konsekvensene av og bevisst sørge for at den primære kunnskap den humanistiske fagkrets gir blir formidlet ut til alle grener av samfunnet. Det kan skje gjennom at fakultetet utdanner filologer som har en kombinasjon av faglig tyngde i tradisjonell forstand og samtidig sitter inne med kunnskap om edb, som gjør dem anvendelige i flere funksjoner. Filologene har flere kvalifikasjoner som trengs i vårt teknologiske samfunn:

1. De er skriveføre
2. De har gode forutsetninger for å fungere i en pedagogisk sammenheng utenfor skoleverket
3. De har i kraft av sin utdannelse kunnskap som kommer godt med når det er tale om kommunikasjon, f. eks. i en selger/kundesituasjon
4. Et viktig virkefelt ligger innenfor teknisk dokumentasjon og rene lærebøker i edb-faget.

-Du hevder at edb-kunnskap vil høyne det faglige nivå. Kan du gå nærmere inn på dette?

Jeg tenker her spesielt på kvantitative metoder i den filologiske forskning.

Bruken av edb vil tvinge frem en større grad av presisjon i forskningsprosjektene. I tillegg får man et mer bevisst forhold til prinsipper for utvelgelse av data. Selve datainnsamlingen og bearbeidelsen for analyse kortes ned. Dermed blir det frigjort mer tid til refleksjon og fordypelse i problemstillinger. De kvantitative metoder tilpasset edb, fører til bedre konsentrasjon om de mest interessante problemstillinger. En større grad av empiri vil ikke være å forakte i vår forskningstradisjon.

Kvantitative metoder er selvfølgelig ikke noe mål i seg selv. Men de er utmerkete midler til å få frem de data som man på tradisjonelt vis vil reflektere over.

Vi ser i dag at intervjuteknikken i forbindelse med meningsmålinger har fokusert på bruken av ledende spørsmål. Tilsvarende vil bruk av spørreskjema i vår forskningssammenheng måtte fokusere på verdispørsmål i forbindelse med et prosjekt. Dermed vil kvantitative metoder ikke tilsløre, men avsløre verdioppfatninger i forskningsprosjekter.

Presisjon i forskningsarbeidet fremtvinges ved bruk av edb og kvantitative metoder. Metoden åpner for større grad av bevisstgjøring hos utøverne når det gjelder selve forutsetningene for forskningen.

Tidligere har man snakket om skoleretninger ved universitetet og på et mer generelt grunnlag vært klar over forutsetningenes betydning. Edb vil bidra til at forutsetningenes betydning vil bli klarere i hvert enkelt prosjekt.

Jeg vil imidlertid understreke at det ikke er snakk om et enten/eller med tanke på bruk av edb. Det er snakk om et både/og. Hovedsaken er at edb i mange tilfelle kan frigjøre ressurser til bruk for analyse og refleksjon. Men edb kan bare brukes på avgrensete områder. Edb egner seg godt til å trekke ut fakta til grunnlag for en analyse. Videre er edb et kraftig hjelpemiddel når man skal bevege seg fra det generelle til det spesielle. Nye sammenstillinger av opplysninger er også en smal sak for edb. Jeg synes det er viktig at rask datatilgang gjør det mulig å teste hypoteser i større omfang enn før, innenfor samme tidsramme. Videre er det en klar forskningsmessig fordel at publikasjonsprosessen er blitt enklere. Andre kan kontrollere resultatene uten å måtte bruke like mye tid som den som samlet forskningsmaterialet.

RAPPORTER

Edb-utdanning for lærere

Nytt studietilbud ved UNIT

Eirik Lien

Ved universitetene i Oslo, Bergen og Tromsø er det utdanning i informatikk, informasjonsvitenskap, datafag - med tanke på å gi en edb-utdanning som er egnet for å utøve edb som profesjon. Ved noen distriktshøgskoler blir det gitt edb-utdanning med samme siktemål. Nå viser det seg at de færreste som har tatt disse studiene, arbeider i skolen. Det vil enten være idealister (og de er det vel ikke mange av i dag?) eller de som har gjort det så dårlig at de ikke får arbeid annet sted. Disse studiene har nemlig ikke som mål å utdanne kandidater for skoleverket, og har derfor ikke noen studieplan som gir en emnesammensetning tilpasset skolen.

I vår fikk imidlertid åtte lærerhøgskoler godkjent fagplan for en halvårsenhet i edb (et halvt års spesialutdanning). I tillegg har NKI-skolen i samarbeid med Statens spesiallærerhøgskole utviklet en fagplan for første halvårsenhet i datalære. Denne halvårsenheten settes i gang i høst og avsluttes til våren neste år, og gjennomføres som fjernundervisning.

Men hvor skal så den delen av lærerstanden som har sin utdannelse fra universitetene, få en skolerelevant edb-opplæring? Universitetet i Trondheim har lenge hatt en edb-utdanning på høyt nivå, nemlig siv.ing. med datafag fra NTH. Men disse har aldeles ikke gått ut i grunnskolen eller videregående skole etter endt utdanning! Dessuten har heller ikke det studiet vært noe alternativ for de studentene som hører til ved NLHT, fordi studiesystemene er vidt forskjellige ved de to delene av universitetet.

Derfor har vi følt behovet for et eget edb-studium - og fant at i tråd med tradisjonene ved denne institusjonen burde et studium som bl.a. rettet seg mot lærerutdanning være fornuftig å satse på. Så det er da blitt vår «nisje», som det nå er blitt så populært å finne. Sentrale myndigheter syntes tydeligvis det samme ved at de har gitt bevilgninger til det formålet. Studiet har vært forberedt gjennom en egen komité

som laget en skisse til studieplan, organisering av studiet og forslag til hvordan det skal bygges ut med stillinger og utstyr. Komitéen foreslo at NLHT burde satse på å finne en egen profil som ikke overlappet det en kunne finne andre steder her i landet, og nevnte datamaskinstøttet læring som et relevant fagområde. Hvordan det er prøvd innpasset, skal vi se på lenger nede.

Denne høsten er det tredje semesteret det drives ordinært datastudium ved NLHT. Til nå har det vært basert på leid hjelp, der rektor *Arvid Staupe* ved TIH og inspektør *Jan Wibe* ved Katedralskolen har vært faglig hovedansvarlige. To faste stillinger (dosenturer) knyttet til edb-studiet har vært utlyst, og søkerne er nå under vurdering. I heldigste fall vil vi ha personer i disse stillingene ved starten av høstsemesteret neste år. Over statsbudsjettet for 1984 er det dessuten foreslått opprettet en amanuensisstilling knyttet til edb-studiet.

Til å administrere studiet er det oppnevnt et eget fagstyre som er satt sammen slik at alle de tre fagavdelingene er representert. Det vil si at det er representanter for de naturvitenskapelige fagene, de samfunnsvitenskapelige fagene og de humanistiske fagene. Dette er gjort fordi vi ikke vil knytte edb-faget til én bestemt faggruppe (f.eks. de naturvitenskapelige fagene), men markere at dette er et fagområde på tvers av tradisjonelle faggrenser. Ikke minst er dette viktig med tanke på at den primære målgruppa for studiet er lærere. I skolen mener vi det er av stor betydning at elevene ikke oppfatter edb som bare et rent teknologisk fag, ved at det f.eks. bare er matematikklærere eller fysikklærere som underviser i edb. Søkningen til studiet så langt har hatt en overvekt av studenter med realfagsbakgrunn, men vi har sett en markert økning fra i fjor til i år av studenter med samfunnsvitenskapelig og filologisk fagkrets. Derfor er det slett ikke urealistisk at en norsklærer eller en historielærer like gjerne kan gi opplæring i edb.

Så langt er studiet satt sammen av åtte forskjellige emner, som til sammen gir 19 vekttall (20 vekttall er normen for ett års studium). Emnene er delvis «lånt» andre steder fra, delvis særegne for dette studiet. De emnene som tilbys og som har vært tilbudt i år, er

DA1 Grunnkurs i databehandling
DA2 Videregående kurs i databehandling
DA3 Edb i skolen
DA4 Edb og samfunn
DA5 Grafisk databehandling
DA6 Datamaskinens virkemåte
DA7 Datastøttet læring
DA8 Sanntids edb

Vi har benyttet oss av samarbeid med NTH i ett av kursene (DA4 Edb og samfunn) og med TIH i to av kursene (DA6 Datamaskinens virkemåte og DA8 Sanntids edb).

Det er neppe mulig å utvide dette tilbudet før det er ansatt noen i de faste stillingene.

en skal tidlig krøkes.....

Til undervisning i DA1, DA2, DA5 og DA7 brukes NLHTs egne AX-anlegg som nå i høst ble supplert med en egen studentmaskin fra UNIT. Det er lagt opp forbindelser til alle de tre delene av NLHT ,ade, Rosenborg og Dragvoll) og disse vil bli bygd ut og forbedret. I åd med den filosofien som er nevnt foran, nemlig at studiet skal være tilbud til alle fagkategoriene, er det viktig at studentene skal kunne ·ive øvinger ved den delen av universitetet hvor de naturlig hører emme.
For å gi et inntrykk av hvordan vi har forsøkt å dreie en del av ınene over på det behovet skolen har, skal vi her gå gjennom noen av m og kommentere dem.

A1 Grunnkurs i databehandling

ırset er bygd opp omkring to hovedkomponenter:
programmering
generelt om datamaskinen

og det er rektor Arvid Staupe ved TIH som har hatt det faglige svaret for dette kurset. Med den bakgrunnen han har fra planlegging edb-undervisning i skolen og vurdering av behov for edb-utstyr i oleverket, er han en helt ideell person for dette kurset. Han har også t idéer på den pedagogiske sida som kommer studentene til gode. Det bli kommentert nedenfor.
I vår sammenheng er det viktig å se på den førstnevnte komponten, som er hoveddelen av kurset. Vanligvis vil en programmeringsplæring gå nokså prompte løs på et bestemt programmeringsspråk og

beskrive «verden» gjennom det. Rent pedagogisk kan det by på problemer fordi det er uvant å tenke ut fra et kunstig språk (programmeringsspråket) som løsningene av de enkelte oppgavene skal innpasses i. En er i andre sammenhenger vant til å beskrive et problem gjennom sitt eget naturlige språk og løse det på en måte som er blitt naturlig på en eller annen måte. Å da møte et språk som er begrenset og har særegenheter i forhold til det en naturlig har vært vant til, kan være en bøyg for mange.

I dette kurset introduseres i stedet programmering gjennom en robot ved navn KAREL. Utgangspunktet er å få beskrevet en «verden» der KAREL kan eksistere og dermed bli fortrolig med det verdensbildet. Hans verden er nemlig uhyre enkel og det er temmelig begrenset hva han er i stand til å gjøre. Han «lever» i en stor by som er avgrenset mot vest og sør, og som har et gatesystem som er kvadratisk. I denne byen kan vi sette opp gjerder som hindrer ham i å gå en bestemt vei. Dessuten kan han enten sette ut, lokalisere («høre») eller plukke opp noen «lyddingser».

Ut fra små, enkle og fundamentale operasjoner kan eleven så få KAREL til å gjøre kompliserte bevegelser gjennom et eget instruksjonsspråk som han forstår.

Instruksjonsspråket er lagt nært opp til det som skal læres senere i kurset, nemlig PASCAL. Dermed får vi en myk overgang til noe som ellers kunne ha blitt komplisert for mange. Og det er ikke minst viktig når edb skal undervises i skolen, der en ikke nødvendigvis har elever som er i stand til å abstrahere i den grad som er nødvendig for å kunne bruke et programmeringsspråk.

DA3 Edb i skolen

Dette er også et emne som er spesielt bygd opp i dette studiet. Målet med det er å gi de som skal ut i skolen, en metodisk bakgrunn for å kunne undervise i edb. Her i byen har vi heldigvis en svært sentral person å trekke veksler på i den sammenheng, nemlig inspektør Jan Wibe, som med glede og entusiasme har gått inn for oppgaven. Kurset har vært holdt én gang, og vil bli gjentatt til vårsemesteret.

Han har ut fra egen erfaring prøvd å bygge inn de aspektene ved faget som er viktige i undervisninga. Han har selv på kurset undervist i emnene

Hva vi vil med faget i skolen
Gjennomgåing av de fagplanene som eksisterer for skolen
Integrert undervisning; hvordan edb kan inngå som hjelpemiddel i undervisning av andre fag. Her er det viktig å gi gode eksempler på disse bruksområdene, f.eks. med relevant demonstrasjonsmateriale. Foreløpig mangler vi slikt materiale, spesielt er vel norsk programvare savnet.
Undervisning i programmering; orientering om hvilke program-

meringsspråk som brukes i skolen (BASIC, PASCAL, LOGO), hvordan skal vi legge opp undervisninga, hvilke typer problemer skal vi løse

Samfunnsorienterte emner; hvilke emner skal vi ta opp, på hvilket årstrinn passer det naturlig å formidle disse emnene, hvem skal ta seg av denne undervisninga: edb-lærere eller samfunnsfaglærere

I tillegg til dette var det inviterte forelesere som tok for seg emner som f.eks.

Hvordan edb og edb-basert biblioteksystemer kan brukes til å finne relevant stoff i undervisninga

Hvilke krav til utstyr og dokumentasjon vi må sette, her gikk det inn demonstrasjon av både maskinvare og programmer

Datamaskinbasert undervisning, kan edb erstatte læreren, hvordan kan læreprogrammer lages («den elektroniske læreren»)

Hvordan kan historiske data som er overført til edb brukes i historieundervisning

Rapport fra lærere som har undervist i skolen i edb, med erfaringer fra ungdomsskole og videregående skole

Men den viktigste delen av dette kurset er vel det prosjektarbeidet studentene skal utføre. I grupper skal de vurdere konkret bruk av edb som hjelpemiddel i undervisningen av andre fag. Poenget her er å være så konkret som mulig med tanke på egen undervisning. Det hele summeres til slutt opp i en rapport.

Erfaringene er langt på vei positive, og denne første gjennomføringa gir et godt grunnlag å arbeide videre på. I et emne som dette må en prøve seg fram og bygge opp kurset ut fra de erfaringene en gjør etter hvert.

DA7 Datastøttet læring

Opphavet til dette emnet er et godt eksempel på hvordan et fagstudium som dette kan bygges ut, og at en i en slik startfase er åpen for nye idéer og tanker. Under gjennomføringa av DA3 i vårsemesteret luftet en av studentene en tanke han hadde fått da det ble forelest om datastøttet læring. Han kunne tenke seg som et gruppearbeid å vurdere en del ferdige læreprogrammer med tanke på hvor anvendelige de kan være i den konkrete skolehverdagen.

Jan Wibe tok ham på ordet og lanserte idéen for fagstyret, som ga Wibe grønt lys til å legge fram en skisse for et slikt emne slik at det kunne gjennomføres i høstsemesteret. Og nå er det i gang med ca. 15 entusiastiske studenter, i all hovedsak lærere med bred undervisningserfaring.

Målet for emnet er rent konkret å utvikle og vurdere programmer og undervisningsopplegg med datamaskiner som hjelpemiddel. I selve studiet skjer det ved at studentene i seminarform gjennomgår det pedagogiske grunnlaget for programmert læring (prof. *Bjørgen* ved

Psykologisk institutt hadde ansvaret for den delen), og arbeider med å utforme forskjellige læreprogram. Til det bruker de et eget «forfatterspråk» (CAS) som er utviklet for VAX-anleggene. Dette språket gjør det relativt enkelt å utforme et læreprogram, uten at en behøver å ha noen omfattende programmeringserfaring. Disse læreprogrammene blir igjen administrert gjennom et spesielt system.

Hele denne teknologien er noe fjern fra norsk undervisningstradisjon, men vi synes at en i utgangspunktet ikke skal si nei til ethvert nytt verktøy. Først må en skaffe seg kompetanse på det spesielle området, og det er nettopp det vi ønsker å oppnå med prosjektoppgavene. I tillegg håper vi å få noe erfaring i utvikling av læreprogrammer som vi kan utnytte i annen programvareutvikling.

Sluttord

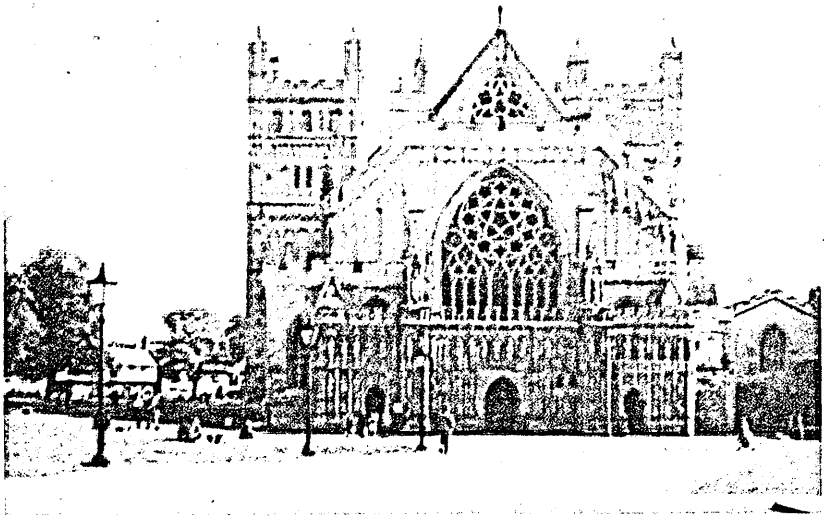
Det er viktig at lærere får edb-utdanning. Noen vil si at vi har kommet altfor seint i gang, men det kan være fordeler med å ikke være altfor rask også. Nå har vi passert en ny revolusjon ved at mikroprosessoren har fått skikkelig fotfeste, og maskinparken har innpasset den. Mikromaskinene har for fullt kommet inn i markedet, som er blitt mer stabilt enn det var bare for ett år siden. Programvaresida er i full utvikling, men vi har nå i noen år hatt «universelle» operativsystem på markedet (f.eks. CP/M). Vi kjenner også mer til hvilke konsekvenser bruken av edb har i samfunnet, lovverket har fått sine tilskudd som skal regulere bruken av edb (Personvernloven). Staten har opprettet et eget organ for å se til at bruken av edb følger de opptrukne linjene (Datatilsynet).

Alt dette gjør at vi nå har bedre mulighet for å kunne organisere et mer relevant opplæringstilbud enn vi kunne for bare få år tilbake. Det er et spennende felt, og det er tilfredsstillende for oss som har vært med på å trekke i gang dette studiet, at vi får positiv respons både hos studenter og myndigheter.

LEXeter '83

Lars Otto Grundt

I dagene 9.-12. september 1983 møtte ca. 250 delegater fra 40 land i Exeter (Sør-England) for å delta i LEX '83, Den internasjonale



Katedralen i Exeter.

kongress i leksikografi. Deltakerne var dels forskere og universitetsansatte, dels forlagsredaktører, dels fagoversettere. Dette brede frammøtet er et sikkert tegn på at leksikografi i dag er blitt til et selvstendig fagområde, med egne metoder og hjelpemidler. Som akademisk fag er leksikografi gjenstand for forskning og undervisning ved en rekke universiteter. I det praktiske liv er produksjon av ordbøker og terminologier blitt til en viktig inntektskilde for forlag og institusjoner og som sådan preget av internasjonal konkurranse.

5 avdelinger tjente som ramme for innleggene. 4 av dem svarte til bestemte ordbokstyper. Avdeling 5 tok hensyn til den viktige rolle edb i dag spiller i det leksikalske arbeide. Nedenfor vil jeg nevne en del av de viktigste innleggene som ble holdt i de forskjellige avdelingene.

Avdeling 1 drøftet problemer knyttet til generell og historisk leksikografi. I innledningsforedraget understreket professor *John Sinclair* at leksikografi mer og mer er blitt til gjenstand for selvstendig forskning og undervisning ved universitetene: Ordboksarbeidet stiller i dag så store krav når det gjelder å fremskaffe og tilrettelegge leksikalsk materiale at det ikke lenger nytter å steppe inn i en stilling som leksikograf straks etter å ha tatt eksamen i et språkfag. Selv om kravet om spesialutdannelse først og fremst stilles av forlag som konkurrerer på verdensmarkedet, ville det kanskje være ønskelig å sette i gang i Norge en regelmessig undervisning i leksikografi og terminologi.

Av andre foredragsholdere som drøftet generelle teoretiske problemer kan jeg av plasshensyn bare nevne professor *Gabriele Stein*:

Towards a new theory of lexicology og professor *Herbert E. Wiegand*: Structure and contents of a general theory of lexicography.

I avdeling 2 møtte deltakere som var interessert i tospråklige ordbøker. Professor *Ladislav Zgusta* klargjorde på en nyttig måte forskjellen i målsetting og metoder mellom de 3 viktigste typer leksika: ordbøker som skal tjene til forståelsen av tekster på fremmede språk (comprehension dictionaries), ordbøker som skal lette oversettelsen til et fremmedspråk (production dictionaries) og ordbøker som primært skal beskrive den kulturelle bakgrunn for ord og uttrykk i utdødde språk (latin, gresk, osv.). En funksjonell ordbokstypologi som den framlagt av Zgusta, gjør det mulig å bestemme for hver ordbokstype hvilke oppslagsord bør velges, hvilken form ordboksartiklene bør få, både når det gjelder uttale, oversettelse, definisjoner, eksempler m.m.

I tillegg til innledningsforedraget fikk vi også høre en rekke innlegg, bl.a. av *Viggo Pedersen* fra København: Prepositions in bilingual dictionaries og *Heikki Särkkä*: Improving the Finnish dictionary. *Mary Snell-Hornbys* foredrag om The bilingual dictionary - help or hindrance? var et vektig bidrag til teorien om tospråklige synonymordbøker.

Avdeling 3 var viet ordbøker beregnet på begynnere. I sitt innledningsforedrag gjorde *Anthony P. Cowie* således rede for ordbøker som brukes i engelskundervisningen for utlendinger. Andre foredragsholdere som *Louise Dagenais* og *Milford B. Kipfer* tok opp spesielle problemer innenfor emnet (definisjoner, ordning av oppslagsordene m.m.).

Avdeling 4 diskuterte først og fremst bruken av edb i ordboksproduksjonen. Innledningsforedraget var ved professor *Frank Knowles*. Han gav en bred orientering om emnet: edb er ikke bare et viktig hjelpemiddel ved innsamling, behandling og systematisering av informasjon, men overtar mer og mer den trykte ordboks funksjoner, bl.a. som oppslagsverk for oversettere, som hjelpemiddel ved utarbeidelse av frekvensordlister, osv. Denne generelle utredningen ble supplert av en rekke innlegg, bl.a. av *Nicoletta Calzolari et al.*: Computational tools for terminological analysis, *Gert Engel* og *Bodil N. Madsen*, København: From dictionary to data base og *Archibal Michiels* og *Jacques Noël*: Controlled defining vocabulary.

I avdeling 5 fikk deltakerne høre innlegg om terminologier og tekniske ordbøker. Professor *Juan Sagers* innledningsforedrag nevnte de viktigste problemer som terminologer og tekniske ordboksforfattere er stilt overfor både når det gjelder valg og definisjon av begreper, termenes form og internasjonal standardisering. Spesielt kom han inn på behovet for en klarere skilnad mellom allmennspråk og fagspråk.

Av andre innlegg vil jeg framheve *Ali Al-Kasimis* foredrag Principles of terminology standardisation: theory and practice, og *Rosemarie Gläsers* utredning om Terminology problems in linguistics, with special reference to neologisms. *Håvard Hjulstad* på sin side gjorde rede for Norsk termbank.

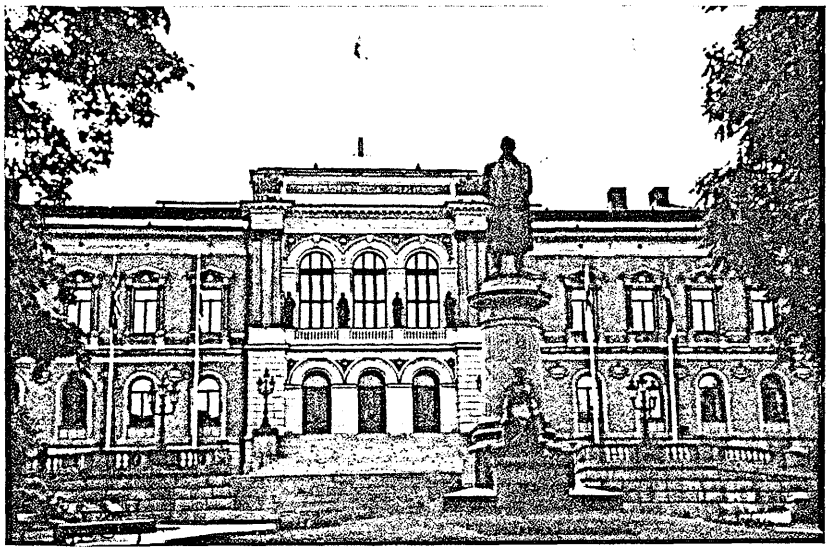
Dr. R.R.K. Hartmann, som var kongressens president, regner med at foredragene vil bli utgitt i løpet av den nærmeste framtid.

For øvrig må nevnes at EURALEX (European Association for Lexicography) ble dannet på møtet. Professor Gabriele Stein, fra Universitetet i Hamburg, ble valgt til forkvinne for styret. Adressen er: *EURALEX, The Language Centre, University of Exeter, Queen's Building, Exeter, Devon, EX4QH, England.* Dr. R.R.K. Hartmann er foreningens sekretær.

De nordiske datalingvistikkdager 1983

Uppsala universitet, 3. og 4. oktober

Jostein H. Hauge



Uppsala universitet.

De nordiske datalingvistikkdagene 1983 var det fjerde symposiet i denne serien. Symposiene er tenkt som et nordisk forum for presentasjon og diskusjon av datamaskinell metodikk innenfor språkvitenskap. Årets arrangement var lagt til Centrum för datorlingvistik ved Uppsala universitet hvor lederen, *Anna Sågvall Hein*, var ansvarlig for opplegget.

I innbydelsen til årets symposium ble det foreslått en tematisk konsentrasjon om emnene datamaskinell morfologisk og syntaktisk analyse (parsing). Responsen var positiv, slik at de fleste foredragene lå innenfor disse emner.

I forlengelsen av datalingvistikkdagene ble det arrangert et symposium om datamaskinstøttet leksikografi. Dette arrangementet, som blir referert annetsteds i bladet, var arrangert av Norsk Termbank og Nordisk institutt, Universitetet i Bergen. I alt hadde det meldt seg ca. 80 deltakere til de to arrangementer, herav 18 fra Norge.

Centrum för datorlingvistik vil gi ut foredragene i en enkel form.

Nedenfor vil vi omtale foredragenes tema for å antyde spennvidden i det nordiske arbeidet på feltet.

I tre foredrag ble arbeidet med språk og tekstanalyse ved Centrum för datorlingvistik behandlet. Anna Sågvall Hein gjorde greie for det språkanalysesystemet (parser för svenska) som benevnes Uppsala Chart Parser (SVE.UCP). Sentralt i analysemodellen er en prosessor som tildeler lingvistisk struktur til en ytring i et gitt språk i overensstemmelse med den lingvistiske kunnskap (grammatisk og leksikalsk) om det gitte språk.

Avgrensingen mellom prosessor og språkbeskrivelse blir dermed et sentralt spørsmål, bl.a. fordi en tvinges til å ta standpunkt til hva som er universelt og hva som er språkspesifikt. I foredraget ble det vist hvordan de grammatiske og leksikalske regler blir aktivisert, er bygd opp og virker.

Lars Ahrenberg tok i sitt foredrag opp analysesystemets grammatiske beskrivelser, som er basert på Martin Kays strukturbegrep. Han sammenlignet systemet med beskrivelsesmodellen innenfor leksikalsk-funksjonell grammatikk.

Lars Borin gjennomgikk hovedkravene til et lingvistisk basert system for tekstanalyse og viste hvordan et slikt system i dag er utviklet i Uppsala. Systemet består av to deler, et generelt databasesystem og en serie spesiallagede aksessmetoder for lett å kunne behandle tekst fra lingvistisk synsvinkel. Dette betyr at en kan f.eks.

- definere alfabet/skriftsystem
- lese inn tekster i et kompakt format
- produsere diverse typer ordlister
- endre informasjonen
- lemmatisere og homografseparere
- oppdatere leksikon o.a.

Fra Finland ble det i 4 foredrag orientert om arbeid innenfor automatisk morfologisk og syntaktisk analyse. I foredragene «Knowledge Engineering Applied to Morphological Analysis» (*Harri Jäppinen*) og «A Computational Model for Finnish Sentence Structures» (*Esa Nelimarkka*) ble det redegjort for et større prosjekt som støttes av det finske nasjonalfond for forskning og utvikling.

Arbeidet sikter mot å utvikle et portabelt system for spørsmål til databaser i naturlig finsk språk. Harri Jäppinen, Helsingfors universitet, gav i foredraget overbevisende inntrykk av at automatisk morfologisk analyse av finsk (som er et såkalt fleksjonsspråk) er langt fra trivielt. Det ble bl.a. vist gjennom meget komplekse strukturskjema for finsk morfologi og eksempler på hvordan ordene kan inngå i komplekse morfotaktiske mønstre. For analysen er det stilt opp som mål å ta med så lite morfologisk informasjon som mulig, men samtidig å ta vare på alle de morfologiske tvetydigheter som ordene kan ha. Selv om arbeidet sikter mot praktisk anvendelse i forbindelse med informasjonssøking, er den syntaktiske analysekomponenten bygget på almen lingvistisk basis, dels på kasusgrammatikk (semantikk) og dependensgrammatikk (funksjonell syntaks). Til nå er det spesifisert en delgrammatikk for finsk som for tiden blir kompilert i LISP ved bruk av en kompilator som bygges for formålet.

Kimmo Koskenniemi, Helsingfors universitet, redegjorde for et generelt system for morfologisk analyse og syntese. Det er generelt i den forstand at de samme språkuavhengige algoritmer og datamaskinprogram benyttes i analysen av mange språk. Systemet er i dag med gode resultater anvendt på finsk, engelsk, rumensk og japansk. Opplegget er kalt et to-nivå system ettersom det består av en leksikondel (som definerer ordstamme, fleksjonsmorfemer m.v.) og et parallelt sett med regler (som definerer de fonologiske trekk). Ved dette bryter systemet med strategien i mange andre systemer som benytter et sett med regler som følger etter hverandre.

Foredragsholderen mente systemet kunne finne anvendelse både i oversettelse, dialogsystemer, korreksjonsprogrammer og ved informasjonssøking.

Fred Karlsson, Helsingfors universitet, viste i foredraget «Tagging och parsing av finsk morfosyntax» hvordan en arbeider for å finne frem til metoder som effektivt kan analysere finsk vokabular. Gjennom eksempelmateriale ble det vist hvor morfologisk komplekst det finske språk er. F.eks. fins det opptil 9 posisjoner for bøyningsendelser i enkelte ordklasser og inntil 200 avledningssuffikser som kan knyttes til ordene. I et vokabular på 70.000 ord er 40.000 avledede ord i en eller annen forstand. Forsøk har vist at man kan redusere et maskinelt leksikon til 10.000 rotmorfemer dersom reglene for derivasjonsmorfemene kan spesifiseres og implementeres i et analysesystem. Siden 1979 har Karlsson nyttet programsystemet BETA som er utviklet av Benny Brodda, Stockholm. I dag er systemet utviklet så langt at 85% rett morfologisk og syntaktisk tagging oppnås i analyse av løpende tekst. Analysen er kun på ordnivå, men ved å trekke inn kontekst, regner man med å nå 93%.

Eva Ejerhed refererte fra arbeid ved Umeå universitet som sikter mot å utvikle et analyse-system (parser) for svensk og engelsk basert på en endelig tilstands-grammatikk (finite state grammar). Til i dag er det

utviklet grammatikk-fragmenter for både engelsk og svensk og en testversjon er implementert. I foredraget ble det gjort sammenligninger mellom systemene i Uppsala og Umeå når det gjaldt analysetid for testsetninger.

Gunnel Källgren fra Stockholms universitet tok opp spørsmålet om hvor langt en kan nå med det en kan kalle «heuristisk parsing», i foredraget også omtalt som «slarvig satslösning». Målet er å bestemme ordklasser, setningsdeler og setningsstrukturer ved hjelp av ulike typer kriterier, også slike som er usikre. Ved denne innfallsvinkel skiller arbeidet seg fra de klassiske analyseprosjekter der strengt formaliserte prosedyrer brukes. Angrepsmåten bør kunne gi svar på to spørsmål:

1) Hvor langt kan en komme med å angi en adekvat grammatisk beskrivelse etter denne metode?
2) Hvilke språkteoretiske implikasjoner har de resultater en får?

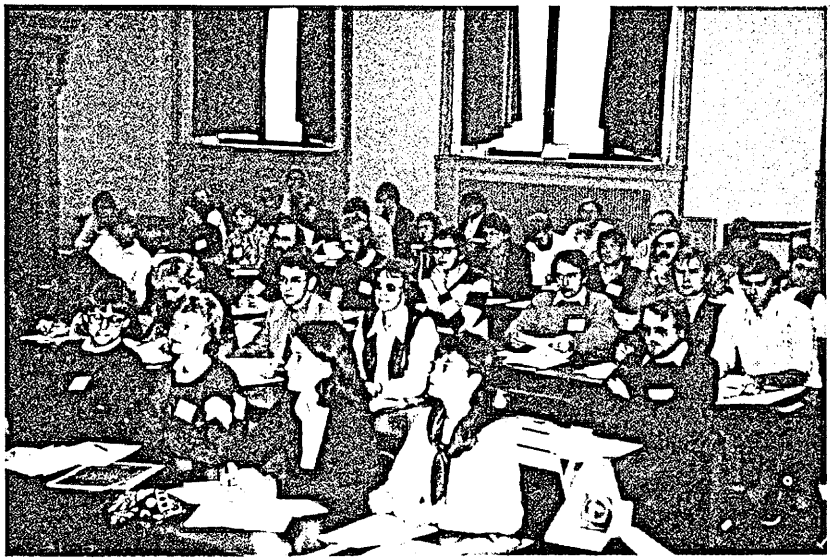
Gunnel Källgren mente at den tilnærmingsmåten som er valgt i prosjektet, har mye til felles med vår naturlige måte å analysere språk på.

I foredraget kom det frem at en kommer overraskende langt med å basere seg på «heuristisk parsing»: Godt over 90% av ordene blir rett klassifisert i løpende tekst.

Helge Dyvik, Universitetet i Bergen, og *Knut Hofland*, NAVFs EDB-senter for humanistisk forskning, redegjorde for bruk av leksikalsk-funksjonell grammatikk (LFG) ved maskinell språkanalyse i Bergen og om den datamaskinelle implementeringen i LISP. Arbeidet baserer seg på et samarbeidsprosjekt med forskere ved MIT og Xerox, Palo Alto. Til i dag er et fragment av norsk grammatikk beskrevet og implementert. En stor fordel med analysesystemet er at det er svært lingvistvennlig. De grammatiske regler kan skrives inn i form av regler innenfor LFG- metoden. Dette vil trolig gjøre det lettere enn ellers å få etablert et samarbeid med «tradisjonelle» lingvister om beskrivelsen av norsk grammatikk til bruk i analysesystemet. (Se for øvrig «Automatisk analyse av norsk», HD 1-83).

Fra dansk hold orienterte *Jan Erlandsen*, Københavns universitet, om de overveielser som ligger bak GESA-systemet for parsing. Et av de overordnede mål har vært å utvikle et system som gjør det enkelt for studenter og forskere å eksperimentere med lingvistiske beskrivelser og analysestrategier.

Utgangspunktet har bl.a. vært at det ofte er vanskelig for lingvister å arbeide med den type formalisme som datasystemet krever (f.eks. tabelldrevet parser). Alternativet som da byr seg frem, er å la språkbeskrivelsen utføres i en brukervennlig formalisme som så omskrives i en mer formell form før analysen tar til. Ulike problemer ved design av en slik oversetter ble drøftet. Løsningen som er valgt, bygger på kompilatorteknikk, dvs. teori for oversettere for programmeringsspråk (Pascal, FORTRAN osv.).



Deltakere på De nordiske datalingvistikkdager 1983.

Oversettelser fra japansk til dansk var emnet for *Arendse Bernth*, Københavns universitet. Ved Datalogisk institut er det utviklet metoder for å bruke logikkprogrammering som hjelpemiddel ved oversettelse. I prosjektet brukes programmeringsspråket PROLOG ved syntaksanalyse av japansk. Det bygges opp en predikatslogisk modell av teksten som så oversettes ved hjelp av en japansk-dansk maskinell ordbok.

Gjennom eksempler ble det vist at det er relativt enkelt å lage et analyseprogram i PROLOG når man først har spesifisert grammatikken. Problemet er større når det gjelder å etablere et «mellomspråk». Det er her et stort problem at en rekke av de semantiske distinksjoner som vi legger vekt på, ikke finnes i japansk. Slikt vanskeliggjør – forståelig nok – i betydelig grad en automatisk oversettelse.

Fra samme institutt orienterte *Gregers Koch* om et prosjekt som sikter mot å kunne oversette tekst i naturlig språk til logiske formler innenfor rammen av en semantisk teori.

Bente Maegaard, Københavns universitet, drøftet i sitt foredrag utviklingen av såkalte «regelformalismer». I oversettelsesprosjekter går en nå mer og mer over til å lage systemer der program og data holdes adskilt, dvs. at også grammatikken blir en del av datagrunnlaget. Dette fører imidlertid til at grammatikken må skrives i et bestemt format, dvs. regelformalisme. EUROTRA (EF's oversettelsesprosjekt) er svært omfattende også når det gjelder bruk av regelformalismer. Det er videre meningen at samme formalisme skal brukes i alle tre hovedkom-

ponenter i systemet, dvs. ved analyse, overføring og generering. Problemene knyttet til bruk av samme formalisme ved beskrivelser både på ord- og setningsnivå ble omtalt.

I det nåværende arbeid med å gi en systembeskrivelse for EUROTRA er to lingvister fra hvert land med i arbeidet. I den senere arbeidsfasen vil imidlertid inntil 15 prosjektdeltakere med lingvistbakgrunn delta fra hvert av de 7 land. EUROTRA blir dermed det desidert største og ambisiøse oversettelsesprosjekt til nå i Europa.

Temaet for foredraget til *Hanne Ruus*, Københavns universitet, var hvordan en best kan tilpasse den morfologiske analyse til bruk i en etterfølgende syntaktisk/semantisk analyse. Det viser seg f.eks. at dersom en karakteriserer alle ord i uttrykket «med så søde ord» med alle tilgjengelige morfologiske opplysninger, blir det ca. 50 alternative veier som et syntaktisk analyseprogram må undersøke. Foredragsholderen fremla metoder for å redusere omfanget og viste hvordan dette også vil virke inn på beskrivelsen av dansk morfologi i sin alminnelighet.

Et prosjekt for automatisk rotlemmatisering var emnet for *Tove Fjeldvig* og *Anne Golden*, Universitetet i Oslo. Formålet var å undersøke mulighetene for automatisk å gruppere ord med felles rot på tvers av ordklassene. Av ressursmessige grunner ble tanken om å bruke et maskinelt utviklet leksikon forlatt til fordel for arbeid med generelle regler for reduksjon av ordformer til rotform.

Resultatene fra et tekstmateriale på ca. 20.000 tekstord viser at ca. 97% ble gruppert riktig ved bruk av de reglene som var utviklet.

I neste omgang kan det bli aktuelt å videreføre arbeidet innenfor en gitt språksektor, eksempelvis juridisk språk. Resultatene fra en tidligere undersøkelse av lærebokspråk viser at spesielle fagrelaterte regler trolig ytterligere kunne forbedre analyseresultatet. (Se også Fjeldvig og Goldens artikkel på s. 22).

Sture Allén, Göteborgs universitet, viste i foredraget «Inte bara idiom» at forekomsten av kollokasjoner (dvs. tilbakevendende ordforbindelser) utgjør et konstituerende trekk i språktekster. Her hører både de ordinære idiomene («være på tapetet») hjemme, men også den store mengden av grammatiske konstruksjoner, f.eks. «med [partisipp] følelser» og metaspråklige faste vendinger (f.eks. «som det heter»).

Viten om disse språktrekk bør ifølge Allén utnyttes aktivt i oppbyggingen av systemer for grammatisk analyse.

«Utnyttande av ordklasser för författarbestämning» var temaet for foredraget til *Bengt Beckmann*, Uppsala universitet. Her ble det redegjort for en undersøkelse innenfor rammen av det svensk-norske prosjektet som har analysert forfatterspørsmålet til «Stille flyter Don». Tidligere har en bl.a. ved forfatterbestemmelsen studert ordklassefordeling i setningsbegynnelse og -slutt som konstituerende tekstdrag ved forfatterbestemmelsen. Beckmann har utvidet dette kriteriet til å gjelde ordklassekombinasjoner totalt og innenfor hele setninger. Resul-

Anna Sågvall Hein og Sture Allén.

tater av undersøkelsen ble fremlagt og kommentert. Han viste bl.a. at resultatene kan bli like pålitelige ut fra et sterkt redusert tekstmateriale.

Henrik Holmboe, Århus universitet, rapporterte om arbeidet med å lage en konkordans over de danske runeinnskrifter. I overensstemmelse med vanlig praksis blir innholdet gjengitt både tegn for tegn og i en transkribert norrøn form. Til hver innskrift blir det også satt på referanseopplysninger og opplysninger om alfabettype.

Bruk av lingvistikk innenfor rettsinformatikk var utgangspunktet for foredraget til *Benny Brodda*, Stockholms universitet. Fremtidssamfunnet vil etter Broddas mening bl.a. være kjennetegnet av enorme tekst- og kunnskapsdatabaser. Fremveksten av slike vil få de tradisjonelle informasjonssøkemetoder til å bryte sammen. Dette gjelder f.eks. systemer som baserer seg på booleske søkeprinsipper. Dette scenario reiser bl.a. krav om metodeutvikling der også lingvistisk fagkunnskap inngår. Det vil bl.a. være behov for systemer som kan foreta morfologisk analyse, utnytte leksikografisk informasjon og bruke syntaktiske metoder for å øke treffsikkerheten. Også strategier for automatisk indeksering og for å utnytte den informasjon som man har oppnådd i tidligere søkinger, bør videreutvikles. I sitt foredrag presenterte Brodda sitt konsept for et nytt informasjonssystem og metoder for bl.a. å utnytte morfologisk analyse i et informasjonssøkesystem.

Ved avslutningen av konferansen ble ulike spørsmål med tilknytning til de nordiske datalingvistikkdager drøftet. Det ble bl.a. vedtatt å ta

imot en invitasjon fra professor Fred Karlsson, Helsingfors universitet om å legge det neste arrangementet til Finland.

COMPILING har til nå vært meldingsblad for nordisk datalingvistikk. I Uppsala ble det vedtatt å nedlegge dette bladet. NAVFs EDB-senter for humanistisk forskning ble oppfordret til regelmessig å presentere datalingvistisk stoff i tidsskriftet Humanistiske Data (se egen melding). På neste konferanse vil spørsmålet bli tatt opp igjen.

Den nordiske samarbeidsgruppen vil videreføre sitt arbeid og vil f.o.m. høsten -83 også ha professor Fred Karlsson som fast medlem fra finsk side.

Baldur Jónsson, Islands universitet, takket på vegne av deltakerne arrangøren, Anna Sågvall Hein, for utmerket tilrettelagt og gjennomført arrangement.

Symposium om datamaskinstøttet leksikografi og terminologi

Uppsala, 5.-6. oktober 1983

Bjarne Norevik/AB

I tilknytning til De nordiske datalingvistikkdagene 1983 i Uppsala ble det arrangert et symposium om datamaskinstøttet leksikografi og terminologi. Dette arrangementet ble planlagt under Nordisk forskersymposium om datamatunderstøttet leksikografi på Sandbjerg i Danmark (august 1982). Norsk termbanks deltakere på dette symposiet tok på seg det faglige ansvaret for et nytt symposium, og det ble etablert et samarbeid med Centrum för datorlingvistik som sto som arrangør av datalingvistikkdagene 1983.

Det var et femtitall deltakere fra hele Norden som hadde møtt fram. Blant deltakerne fantes både leksikografer og terminologer.

Symposiets hovedmål var å sette søkelyset på format og metoder innenfor datamaskinstøttet leksikografi og terminologi. Det ble holdt 11 foredrag som mer eller mindre tok for seg problemer innenfor dette området.

Følgende forskere presenterte foredrag med disse titlene: *Sture Allén* (Språkdata - Göteborgs universitet): «Leksikalisk databas», *Boel Bøggild-Andersen* (Aarhus universitet): «Jysk ordliste - beskrivelse af et prosjekt», *Martin Gellerstam* (Göteborgs universitet): «Samhällsterminologi på invandrarspråk», *Håvard Hjulstad* (Norsk termbank - Universitetet i Bergen): «Databehandling av leksikografi og terminologi

- fins det ein fellesnemnar?», *Bodil N. Madsen* (Handelshøjskolen i København): «Fellesformat for nordiske termbanker», *Rune Midtvedt* (PDS - Universitetet i Bergen): «STRILEK-formatet - en presentasjon», *Peter Nordqvist* (Centralen för teknisk terminologi): «Termbanken Tepa - en lägesrapport», *Heribert Picht* (Handelshøjskolen i København): «Terminologiklassifikation», *Hanne Ruus* (København universitet): «Krav til leksikografisk programmel», *Sigurður Jonsson* (Háskóla Íslands): «Datorsystem för lagning och utgivning av ordlistor», *Kjell Åström* (Tekniska Nomenclaturcentralen): «Ska vi byta data med varandra? - om kraven på framtida språklig datautväxling».

Det viste seg at terminologene som alt i lang tid har brukt datamaskinen som hjelpemiddel, nå var opptatt av å harmonisere på nordisk nivå ved å forsøke å komme fram til et felles lagringsformat og et enhetlig klassifikasjonssystem. Det ble nedsatt arbeidsgrupper (i Nordtermregi) som konkret skal arbeide mot et slikt mål.

Leksikografene har også brukt datamaskinen som hjelpemiddel i arbeidet sitt, men på langt nær så gjennomført og med samme «velvillige instilling til edb» som terminologene. Som det kom fram under symposiet, er leksikografenes viktigste oppgave nå å presisere hvilken hjelp de vil ha av datamaskinen, altså hvilke krav de skal stille til leksikografisk programvare.

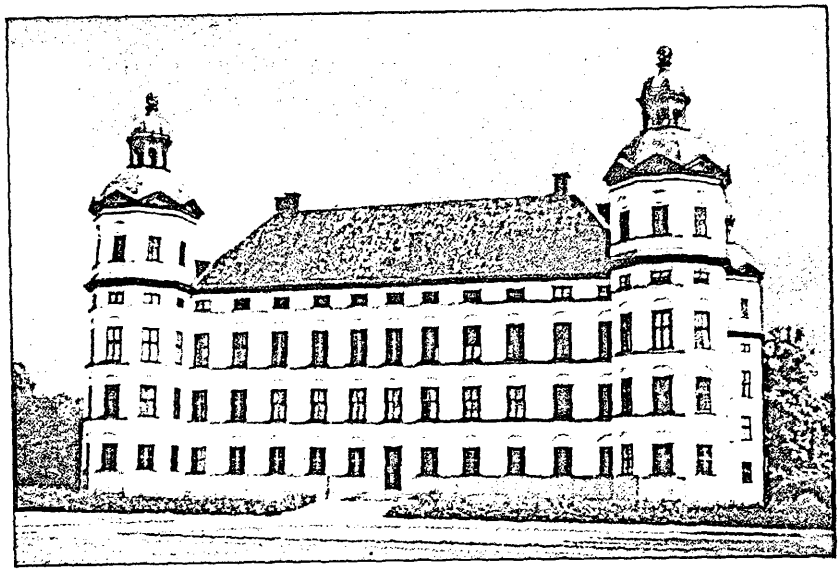
En rapport fra symposiet vil foreligge i desember 1983. Den vil koste kr. 20 og kan bestilles fra *Norsk termbank, UiB, Strømgt. 53, 5000 Bergen.*

Nordiska museet intensiverer edb-virksomheten

Jostein H. Hauge

Humanistiske Data besøkte i oktober Nordiska Museet i Stockholm for å sette seg nærmere inn i edb-virksomheten der. Besøket var en oppfølging av tidligere kontakt mellom NAVFs EDB-senter for humanistisk forskning og intendent Göran Bergengren ved museet. Bergengren har tidligere forelest ved kurs og konferanser som Senteret har arrangert om edb-bruk i museumsarbeid.

Siden 1967 har Bergengren arbeidet med edb-metoder i svensk museumsvesen. Han har hele tiden vært en sentral person i denne typen arbeid samtidig som han også har deltatt i internasjonal virksomhet innenfor organisasjonene International Committee of Museums (ICOM) og dens dokumentasjonsorgan CIDOC. For tiden er han med i samordningsgruppen SAMOREG som utreder en felles dokumenta-



Skokloster slott.

sjonsstandard ved svenske museer.

Startpunktet for det praktiske edb-arbeidet var den edb-registrering av inventar som ble foretatt da den svenske stat overtok det vakre Skokloster slott (se bildet). Senere ble samlingene ved Kgl. Livrustkamren (nå på Slottet) registrert på edb.

Arbeidet med databehandling av Nordiska museets øvrige materiale har pågått siden 1975. I den første tiden ble det benyttet et eksternt servicebyrå ved databehandlingen, men våren 1981 fikk museet installert en Hewlett Packard 1000 minimaskin.

Som ved mange andre store museer så det lenge ut til at edb-registreringen av de sentrale museumsdata ville bli en uoverkommelig oppgave.

I løpet av de par siste år er det imidlertid blitt betydelig fart i dette arbeidet ved at museet har fått anledning til å etablere en midlertidig registreringssentral finansiert av sysselsettingsmidler. Ved denne sentralen, som ble opprettet i 1982 i Älvsbyen i Norrbotten, var det i den første fasen en stab på 21 kontorutdannete medarbeidere.

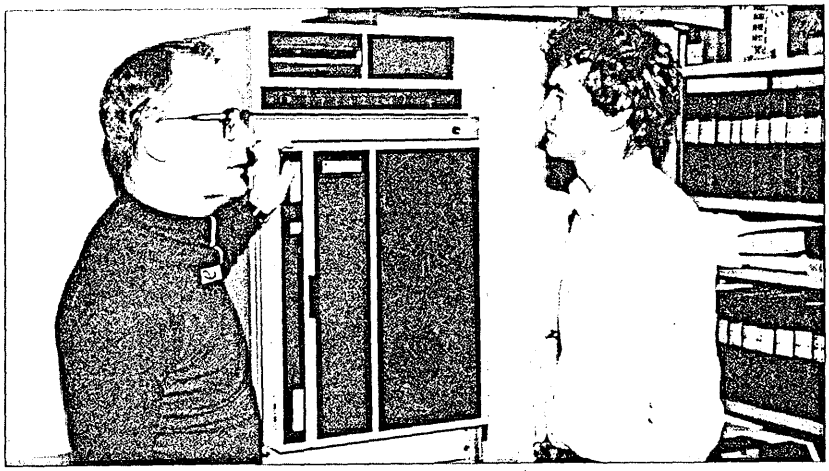
I løpet av 9 måneder ble det her overført aksesjonsprotokolldata i et antall av 140.000 nummer. Ved en videreføring av denne virksomheten i 1983 vil ytterligere 170.000 nummer bli tilrettelagt for databehandling. Sentralen har fått midler til å disponere en minidatamaskin og 10 terminaler.

Med utgangspunkt i dette materialet er det meningen senere å supplere registreringene med ytterligere gjenstandsopplysninger og magasininformasjon.

Nordiska museet i Stockholm.

Den store mengde tilrettelagte data gjør det også nødvendig å øke maskinkapasiteten på museet med bl.a. ekstra datalagringsutstyr.

Edb-virksomheten ved Kulturarvet i Falun ble presentert i forrige



Intendent ved Nordiska museet Göran Bergengren (t.v.) i samtale med Knut Hofland fra Senteret.

nummer av Humanistiske Data. I vår samtale med Bergengren ble det opplyst at denne institusjonen fra 1. juli i år er lagt inn under Nordiska museet.

I tillegg til disse museer arbeidet Armémuseet i Stockholm målbevisst for å integrere edb-rutiner i sitt museumsarbeid. Her har en også tatt opp arbeidet med å utvikle metoder for bildelagring kombinert med databehandling. Det ble opplyst at også flere lensmuseer i dag er i ferd med å orientere seg mot databehandlingsmetoder.

Vårt besøk ga et overbevisende inntrykk av at edb-arbeidet ved Nordiska museet er i planmessig fremdrift under Göran Bergengrens ledelse. Det er også tydelig at edb-arbeidet ved øvrige museer i Sverige i årene som kommer vil bli betydelig utbygget.

NAVF

Nytt fra RHF/NAVF

Prosjektinformasjonstjeneste

Det er nå fattet vedtak i Styret i NAVF om å opprette en NAVFs informasjonstjeneste for igangværende forskningsprosjekter for en prøveperiode på 5 år fra 01.01.1984. Tjenesten er i første omgang avgrenset til rådene for humanistisk (RHF) og samfunnsvitenskapelig (RSF) forskning samt forskning for samfunnsplanlegging (RFSP).

Informasjonstjenestens tekniske administrasjon skal etableres i tilknytning til NAVFs EDB-senter for humanistisk forskning i Bergen.

De tre faglige rådene har besluttet at de vil opprette egne fagtjenester som skal samarbeide nært med den tekniske enheten. For samfunnsvitenskap og samfunnsplanlegging vil det bli én felles tjeneste. Fagtjenestene og den tekniske enhet vil bli koordinert av et eget utvalg som også skal fungere som styre for den tekniske tjenesten. Fagtjenestene vil dessuten få egne styringsgrupper.

Den humanistiske fagtjenesten vil også bli lagt til Senteret slik at en her får et kontorfellesskap. Målsettingen for den er å etablere og drive en dokumentasjonstjeneste for igangværende RHF-finansiert forskning. Den skal arbeide for at denne forskning til enhver tid kan dokumenteres tilfredsstillende for ulike formål. Etter vedtak i rådet skal fagtjenesten også foreta totalkartlegging av forskning innen humanistiske fag eller utføre andre, f.eks. sektor- eller temaavgrensede kartlegginger, evt. som eksternt betalte oppdrag. Et overordnet mål for arbeidet er å stille resultatene til disposisjon for RHFs sekretariat og andre NAVF-organer. Dessuten vil det bli utviklet informasjonstjenester rettet for forskersamfunnet, offentlige myndigheter, media og allmenheten.

Fagtjenestens daglige virksomhet vil bli ledet av direktøren for Senteret.

Utredning om behov for vitenskapelig utstyr i humanistisk og samfunnsvitenskapelig forskning

Rådene for humanistisk og samfunnsvitenskapelig forskning i NAVF har besluttet at de i 1984 vil gjennomføre en utredning om behovet for vitenskapelig utstyr innen sine fagområder. Arbeidet vil bli utført i nært samarbeid med NAVFs Utredningsinstitutt. I 1983 er det alt i NAVF-regi foretatt en kartlegging for naturvitenskap og medisin.

Etterhvert som humaniora og samfunnsvitenskap også i sterkere grad tar i bruk nye tekniske hjelpemidler, vil det med all sannsynlighet bli aktuelt med mer vitenskapelig utstyr. Innenfor humaniora vil dette kanskje særlig gjelde områder som filmvitenskap, musikkvitenskap, arkeologi og språkvitenskap.

RHF har hittil ikke operert med noe eget, separat program for utstyr på sitt budsjett. Dette er blitt vurdert i tilknytning til bestemte prosjekter. I stor grad har det inntil nå vist seg at behovet har kunnet bli dekket ved utlån fra eksisterende utstyrspark ved NAVFs instrumenttjenester. En kan imidlertid ikke regne med at dette er en situasjon som vil vedvare. RHF har også grunn til å tro at det reelle behov for utstyr innen humanistiske fag ikke har manifestert seg i form av søknader til NAVF f.eks. En av årsakene til dette kan være at humanister er mindre bevisste enn forskere fra andre områder om hvilke muligheter som foreligger på feltet. RHF er derfor svært interessert i at det nå for første gang i Norge foretas en bred kartlegging også for sine fag.

Elisabeth Johnsen

MELDINGER

En historisk nyhet

En nyhet på statsbudsjettet for 1984 er at Registreringssentral for historiske data er finansiert utover den 3-årige prøveperioden. Stillingene er satt opp på overgangsstatus under Universitetet i Tromsø.

RHD har hittil konsentrert innsatsen om registrering og databehandling av folketellingene 1865 til 1900 på individnivå. Vi dekker nå Midt- og Nord-Troms pluss områder i Finnmark, Nordland og Sør-Norge. En annen nyhet er at folketellinga 1875 for Kristiania foreligger på mikrokort med sorterte registre.

Våre produkter omfatter maskinskrevne versjoner av originalkildene og alfabetiske registre til disse. Dessuten omarbeider vi den historiske befolkningsstatistikken ut fra dagens krav til klassifisering og områdeinndeling. Målet er et nasjonalt personregister for 1700- og 1800-tallet som forskere kan benytte bl.a. til erstatning for nyere registre som beskyttes av reglene om personvern. Registret bygges opp ved å skrive av de gamle folketellingene og kirkebøkene, og databehandle dem.

Det er mange argumenter for å opprettholde virksomheten ved registreringssentralen på sitt nåværende nivå. Hovedgrunnen er at vi bare har behandlet kildene for noen deler av landet. Og de brukergrupper vi forsyner med data, er ofte avhengig av materiale fra et bestemt geografisk område for at det skal være interessant å anvende.

Dette gjelder i særlig grad innenfor skoleverket, hvor våre listedata brukes i lokalsamfunnsundervisninga. Siktemålet er å få til integrerte opplegg med opplæring i edb-faget. Erfaring viser at slikt lokalt materiale motiverer elevene i sterk grad. Kombinasjonen med bruk av edb gir samfunnsfaget et mer praktisk tilsnitt.

Kildemateriale med lokal tilknytning er selvsagt også viktig for by- og bygdehistorikere og for slektshistorikere. Det samme gjelder innenfor arkivverket; det nytter ikke å tilby rekvirentene kilder fra en annen bygd fordi de er databehandlet. Samtidig er materialet i sin opprinnelige form svært tungvint å bruke. All blaingen er faktisk i ferd med å slite ut de gamle kirkebøkene og folketellingene. Dessuten sparer våre alfabetiske registre arkivpersonalet for mye arbeid.

Initiativet til registreringssentralen ble tatt av faghistorikere som ønsker å studere 1700- og 1800-tallssamfunnet med mikrohistoriske metoder. Dette vil si å ta utgangspunkt i data om enkeltindivider. Bare på den måten kan man få sikre kunnskaper om vår sosialhistorie. Men

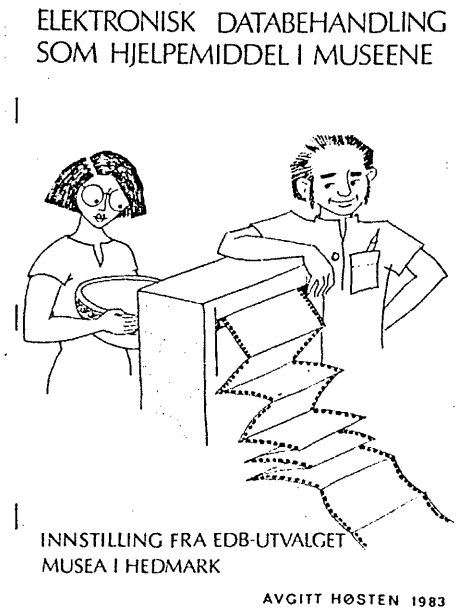
Registreringssentral for historiske data

også navnegranskere, pedagoger, sosialmedisinere o.a. viser stor interesse for våre kildeutgaver.

Datapolitisk representerer registreringssentralen et forsøk på å snu den utvikling at slike serviceoppgaver i stor grad blir lagt til de større byene. RHD har utviklet generelle teknikker for dataregistrering som i praktisk bruk har vist seg velegnet til desentralisering.

Gunnar Thorvaldsen

Innstilling fra EDB-utvalget for Musea i Hedmark



I 1982 oppnevnte Musea i Hedmark et edb-utvalg med følgende medlemmer: konservator Anne-Lise Svendsen og bibliotekar Guri Velure, Glomdalsmuseet, og amanuensis Øivind Vestheim og kontorsjef Egil M. Kristiansen fra Norsk Skogbruksmuseum. Utvalget har nylig avgitt innstillingen «Elektronisk databehandling som hjelpemiddel i museene.» Innstillingen gir en lettfattelig oversikt over arbeidsoppgaver som er egnet for edb, tilgjengelig edb-utstyr og forskjellige systemer som museene kan ta i bruk. Det blir dessuten pekt på typer av kostnader de ulike alternativene vil medføre. Utvalget understreker at museene må satse på fleksible systemer som kan tilpasses de enkeltes behov.

Utvalget foreslår at Musea i Hedmark nedsetter en prosjektgruppe med museumsfaglig og datafaglig ekspertise som skal utvikle et edb-opplegg museene kan samarbeide om. Etter utvalgets mening bør museumsorganisasjonen vurdere å innlede et samarbeid med Kommunedatasentralen for Øst-Norge, slik at KDØ kan ta ansvaret for den datafaglige delen av prosjektet. Prosjektgruppen må dessuten ha nær kontakt med NAVFs EDB-senter og NKKM's EDB-komite.

Senterrapport nr. 26

Stig Welinder: Paleodemography

Demografi, eller befolkningslære, er studiet av storleiken, samansetjinga og forandringa til ei befolkning. Paleodemografi, som ein vil gissa, vert brukt om demografi i forhistorisk tid. Det er dette emnet Stig Welinder tek opp i rapporten «Paleodemography». Han tek føre seg deskriptiv demografi, dvs. beskriving av ei definert befolkning, med hjelp av osteologiske undersøkingar (osteologi = læra om knoklane).

Rapporten inneheld ein presentasjon av programmet DEMO. Det skal hjelpa brukaren å rekonstruera storleiken til og strukturen av befolkninga til ein bustadplass. Programmet byggjer på kjønns- og alderssamansetjinga av folket som er vorte gravlagd på gravplassen til bustadplassen.

Viss ein har eit datamateriale ifrå ein gravplass, kan ein med hjelp av DEMO rekonstruera befolkningsstrukturen på gravplassen. Ein kan også få ut tabellar som fortel kva fordeling dei ulike aldersklassane hadde. Vidare kan ein sjå kva sjansane var for å døy innafor ein aldersklasse. Ein ser også, i dei same tabellane, kor mange år ein person hadde att å leva. Desse tabellane kan ein også få presentert grafisk. No er det ofte slik at gravplassane vart brukt over eit tidsrom som femner fleire generasjonar. Ved hjelp av programmet kan ein få eit estimat som fortel kor stor den faktiske befolkninga var.

Det er ein del uvisse knytta til slike analysar. Blant anna kan det vera vanskeleg å fastslå kva kjønn eller alder eit individ har. Desse problema drøfter Welinder i rapporten og han kastar lys over både problema og programmet DEMO med hjelp av tre døme ifrå gravplassar i Sverige.

A STAR is born

I eit tidlegare nummer av HD (nr. 1-83) skreiv *Stig Welinder* at programpakken STAR (STatistics in ARchaeology) var ferdig. No er pakken også klar til å verta implementert på andre maskinar enn UNIVAC-maskinen i Bergen. STAR vil etter alt å døma vera implementert i Oslo og Trondheim når dette vert lest. Dessutan vil implementeringa vera i gong i Tromsø (CYBER) og Stavanger (IBM hos Rogalandsdata).

Nærare opplysningar om kva pakken inneheld, kan ein lesa i rapporten som er nemnt over. Det er også laga ein informasjonsbrosjyre. Brosjyren presenterer nærare detaljar om STAR og gjev enkelte tekniske opplysningar. Dessutan vil ein finna opplysningar om prisar på magnetbandet og rapportane som er utarbeidde i tilknytning til pakken. Brosjyren kan ein få ved å venda seg til Senteret.

Videreføring av COMPILING

På de nordiske datalingvistikkdagene i Uppsala 3. og 4. oktober, ble det vedtatt å nedlegge meldingsbladet COMPILING. COMPILING har i en rekke år vært informasjonsorgan for nordisk datalingvistikk. Innholdet har bestått av meldinger, rapporter og artikler. Det har i den senere tid vært vanskelig å få til en regelmessig utgivelse, særlig på grunn av sparsom tilgang på relevant stoff.

På konferansen i Uppsala ble NAVFs EDB-senter for humanistisk forskning bedt om å overta COMPILINGs informasjonsoppgaver. Tanken var at Humanistiske Data jevnlig skulle inneholde stoff av interesse for nordisk datalingvistikk. Muligheten for å lage temanummer om emnet, ble også drøftet.

Senteret har påtatt seg denne oppgaven og vi ønsker nå å oppfordre leserne til å bidra med stoff. Særlig ønsker vi regelmessig å kunne gi meldinger om opplærings- og konferansevirksomhet i de nordiske land og presentere forskning på feltet.

Vi har alt bedt den nordiske samarbeidsgruppen for datalingvistikk om å bistå oss ved at medlemmene for de enkelte nordiske land fungerer som nasjonale kontakter for informasjon på feltet. Medlemmene i den nordiske samarbeidsgruppen er:

Fred Karlsson, Helsingsfors universitet
Bente Maegaard, Københavns universitet
Sture Allén, Göteborgs universitet

Kolbjørn Heggstad, Universitetet i Bergen
Baldur Jónsson, Islands universitet

De som er interessert i nærmere kontakt om saken kan henvende seg til vår informasjonskonsulent Kristin Natvig.

CONDUIT

Den amerikanske organisasjonen CONDUIT har som formål å fremme bruken av datamaskinstøttet undervisning på universitetsnivå. For å oppnå dette målet forsyner CONDUIT universitetslærere med idéer, undervisningsmateriale og informasjon om datamaskinstøttet undervisning. Dessuten assisterer organisasjonen i produksjonen og, fremfor alt, distribusjonen av datamaskinbasert undervisningsmateriale. Ferdige programmer og pakker blir underlagt en nøye utprøving og vurdering, og de som blir godkjent, bearbeides for enkel implementering på ulike typer maskiner.

CONDUIT utgir tidsskriftet «Pipeline» to ganger i året. Bladet inneholder artikler om datamaskinstøttet undervisning og beskrivelser/vurderinger av tilgjengelig programvare. Vårnummeret '83 har som tema «CAI in English Composition.»

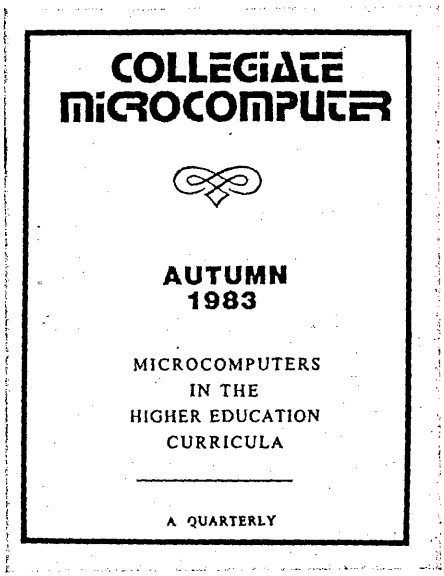
CONDUITs adresse: *P.O. Box 388, Iowa City, Iowa, 52244, USA.*

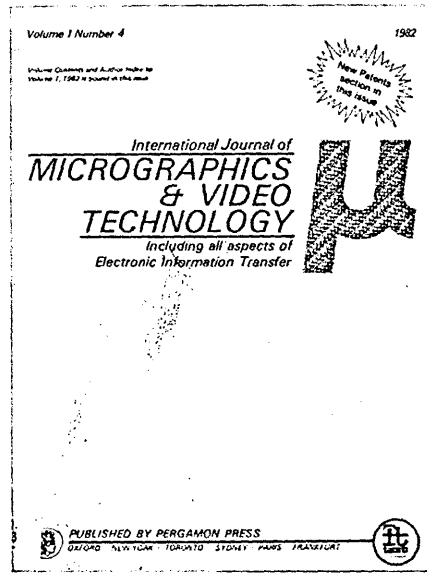
KORT OG GODT

Under denne overskriften skal Humanistiske data bringe korte nyhetsmeldinger. Redaksjonen tar gjerne imot tips fra leserne.

- Nytt fra HF-data, Universitetet i Oslo: For perioden 1.8.83-1.8.84 er *Åsbjørn Brændeland* ansatt som vikar for Ivar Fonnes. Brændeland melder at hele 150 studenter deltar i kurset «Edb for humanister» i høst - en økning på 300% på ett år.
- Innstillingen om et dokumentasjonssenter for kulturpolitikk og kulturforskning i Stavanger (presentert i HD 1-83) er utkommet i bokform: K. Bjander, E. Fossåskaret, T. Karlstein og L. Rosenlund: *Kulturpolitikk og kulturforskning.* Universitetsforlaget 1983. 92 s.
- Nytt tidsskrift f.o.m. 1984: *Journal of Educational Computing Research*, utgitt av Baywood Publishing Company, Inc., 120 Marine St., P.O. Box D, Farmingdale, New York, 11735, USA.

AKTUELLE TIDSSKRIFTER





Collegiate Microcomputer

Collegiate Microcomputer er et forum for utveksling av idéer om mikromaskinens rolle i universitetene. Bladets artikler omhandler bruken av mikromaskiner i undervisning, forskning, bibliotek og kontorer, og i planlegging og utvikling. Ellers består stoffet av bl.a. vurderinger av maskiner og programvare, beskrivelser av kurs, presentasjoner av resultater av forskning og eksperimenter (også på studentnivå) og anmeldelser av produkter, tjenester og litteratur.

Tidsskriftet utgis kvartalsvis og et årsabonnement koster $ 36 ($ 60 sendt med luftpost). Adresse: *Collegiate Microcomputer, Rose-Hulman Institute of Technology, Terre Haute, Indiana, 47803, USA.*

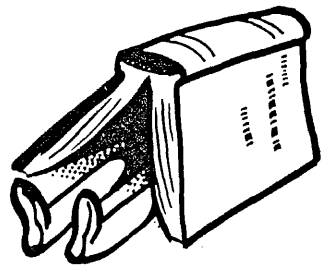
International Journal of Micrographics & Video Technology

Dette tidsskriftet rapporterer om utviklingen i elektronisk informasjonsoverføring innen mikropublisering, elektroniske tidsskrifter, videoteknologi m.m. Bruken av databaser ved hjelp av disse midlene er også inkludert. Søkelyset settes på utvikling og implementering av integrerte systemer i mikropublisering og ulike typer institusjoner.

I tillegg til artikler om dette feltet, har bladet følgende faste innslag: nyheter innen forskning og produktutvikling, bokanmeldelser og meldinger om møter, seminarer o.l.

Bladet utkommer med fire nummer i året. Et abonnement koster $ 55 for institusjoner og $25 for enkeltpersoner, og kan tegnes hos: *Pergamnon Press., Headington Hill Hall, Oxford OX3 0BW, UK.*

Nytt i biblioteket



Daltveit, Memund m.fl.: STAR. A program package for archaeological use. I. Introduction and STAR manual. Bergen, 1983. 73 s.

Berg, Karin m.fl.: STAR. II. Student textbook and Star examples. Bergen, 1983. 108 s.

Welinder, Stig: STAR. III. Archaeology for statisticians. Bergen, 1983. 137 s.

Daltveit, Memund m.fl.: STAR. IV. STAR algorithms. Bergen, 1983. 26 s.

Fogelvik, S. m.fl.: Man, landscape and society. An information system. Stockholm, GOTAB, 1981. 84 s.

Fogelvik, Stefan og Sperlings, Sven: Using individually based historical data for research in the humanities and the social sciences. Stockholm, uten år. Stensil. 17 s.
Frängsmyr, Tore: Humanioras egenart. En rapport. Oslo: Universitetsforlaget, 1983. 113 s.
Golden, Anne og Hvenekilde, Anne: Rapport fra prosjektet LÆREBOKSPRÅK. Oslo, 1983. 178 s.
Holmboe, Henrik (red.): SOM 3. Sprog og mennesker. Nordisk forskerkursus om datamatunderstøttet leksikografi, august 1982. Århus, 1983. 156 s.
Actes du Congrès international informatique et sciences humaines. Liège, 18-21 Novembre 1981. Liège, 1983. 932 s.

KONFERANSER

Cranfield

Internasjonal konferanse om datamaskinell oversettelse

En internasjonal konferanse om metoder og teknikker knyttet til datamaskinell oversettelse skal avholdes i Cranfield, England, 13.-15. februar 1984. Konferansen arrangeres av Cranfield Institute of Technology i samarbeid med the Natural Language Translation Specialist Group of the British Computer Society.

Konferansen vil legge vekt på analyseteknikker og programmering, illustrert ved demonstrasjoner av operative oversettelsessystemer.

Ytterligere opplysninger: *Douglas Clarke, Department of Mathematics, Cranfield Institute of Technology, Cranfield, Bedford MK43 0Al, England.*

Eleventh International ALLC Conference

Association for Literary and Linguistic Computing arrangerer sin 11. konferanse 2.-6. april 1984 ved Université Catholique de Louvrain, Belgia. Hovedtemaene for konferansen blir teori, metoder, problemer og anvendelse i tilknytning til edb-bruk i litteratur- og språkforskning.

Påmeldingsfrist for konferansen er 31. januar. Konferansesekretariatets adresse: *Dr. Jacqueline Hamesse, Institut Supérieur de Philosophie, Chemin d'Aristote,1, B-1348 Louvrain-la-Neuve, Belgium.*

Fogelvik, Stefan og Sperlings, Sven: Using individually based historical data for research in the humanities and the social sciences. Stockholm, uten år. Stensil. 17 s.
Frängsmyr, Tore: Humanioras egenart. En rapport. Oslo: Universitetsforlaget, 1983. 113 s.
Golden, Anne og Hvenekilde, Anne: Rapport fra prosjektet LÆREBOKSPRÅK. Oslo, 1983. 178 s.
Holmboe, Henrik (red.): SOM 3. Sprog og mennesker. Nordisk forskerkursus om datamatunderstøttet leksikografi, august 1982. Århus, 1983. 156 s.
Actes du Congrès international informatique et sciences humaines. Liège, 18-21 Novembre 1981. Liège, 1983. 932 s.

KONFERANSER

Cranfield

Internasjonal konferanse om datamaskinell oversettelse

En internasjonal konferanse om metoder og teknikker knyttet til datamaskinell oversettelse skal avholdes i Cranfield, England, 13.-15. februar 1984. Konferansen arrangeres av Cranfield Institute of Technology i samarbeid med the Natural Language Translation Specialist Group of the British Computer Society.

Konferansen vil legge vekt på analyseteknikker og programmering, illustrert ved demonstrasjoner av operative oversettelsessystemer.

Ytterligere opplysninger: *Douglas Clarke, Department of Mathematics, Cranfield Institute of Technology, Cranfield, Bedford MK43 0Al, England.*

Eleventh International ALLC Conference

Association for Literary and Linguistic Computing arrangerer sin 11. konferanse 2.-6. april 1984 ved Université Catholique de Louvrain, Belgia. Hovedtemaene for konferansen blir teori, metoder, problemer og anvendelse i tilknytning til edb-bruk i litteratur- og språkforskning.

Påmeldingsfrist for konferansen er 31. januar. Konferansesekretariatets adresse: *Dr. Jacqueline Hamesse, Institut Supérieur de Philosophie, Chemin d'Aristote,1, B-1348 Louvrain-la-Neuve, Belgium.*

COLING 84

COLING 84 - foredragsinnbydelse

Den 10. internasjonale konferansen om datalingvistikk - COLING 84 - skal finne sted 2.-6. juli 1984 ved Stanford University, California. Sammendrag av foredrag må leveres senest 9.1.84 til: *Professor Yorick Wilks, COLING 84, University of Essex, Colchester CO4 3SQ, Essex, England.*

Flere opplysninger om konferansen kan skaffes fra: *Dr. Martin Kay, Xerox PARC, 3333 Coyote Hill Road, Palo Alto, California 94304, USA.*

Second International Conference on Automatic Processing of Art History Data and Documents

Konferansen skal finne sted 24.-27. september i Pisa. Formålet er å bidra til utviklingen av standarder for formidling av informasjon innen kunsthistorie. Følgende emner vil bli tatt opp: tesauri og leksikon, ikonografi, bibliografi, biografi, kataloger, arkeologi, og dokumenter og kilder. Andre emner som vil bli diskutert, er integrerte informasjonssystemer og vitenskapelig edb-anvendelse innen kunsthistorie.

Sammendrag av foredrag må leveres innen 31. desember til: *Laura Corti, Scuola Normale Superiore, Piazza dei Cavalieri 7, 56100 Pisa, Italy.*

SUMMARY

Videoplateteknikk
Videodisc technology
Executive Officer Elin Solstrand at the Centre gives an introduction to the principles of videodisc technology and its applications. She also refers to projects abroad that make use of this form of storage.

The capacity, cost, durability, and quality of different kinds of videodiscs is discussed. The terms analog and digital storage are explained as well as the technique of optical laser storage. A survey is given of methods of duplication and the research that has beeen carried out on the production of erasable discs. Finally, Solstrand outlines various areas of application – in education, publishing, archives, libraries, and data bases.

Automatisk rotlemmatisering
Automatic root lemmatization
One of the projects currently taking place at the Institute for Computers and Law, University of Oslo, deals with various linguistic aspects of information retrieval from texts. So far work has been carried out by Research Fellow Tove Fjeldvig and M.A. Anne Golden on the development of a method for grouping words with joint roots irrespective of part of speech. Automatic root lemmatization can facilitate the analysis of search arguments in natural language, and help create a better basis for searching in texts.

Fjeldvig and Golden have developed a set of rules for root lemmatization incorporated in a program system, which they describe in detail. They also give examples of the method and discuss the different uses it can be put to if further developed.

Edb og språknormering
ADP and language planning
In this article Senior Executive Officer Aagot Landfald gives a survey of the ways in which the Norwegian Language Council utilizes printed lists based on machine stored language material. This material, which is used by the Council in its work on language planning, has been prepared by the Division of Computational Linguistics, University of Bergen, the Norwegian Text Archive, and the Centre.

Landfald describes in detail the different kinds of coded lists, and gives examples of both their use in language analyses and these analyses' consequences for language planning. Lists derived from an official vocabulary of Norwegian and New Norwegian are used in connection with regulating orthography, whereas information on the actual use of words is obtained from lists based on newspaper material. The Council has also initiated the transformation of a large body of fiction to machine readable form in order to

investigate if and how various orthographical reforms influence authors' language usage.

3. Nordisk forum for edb-bibliotekarer: Automatisk indeksering
3rd Nordic forum for ADP librarians: Automatic indexing

The third Nordic forum for ADP librarians took place at Norefjell in September. The seminar's main theme was automatic indexing and other methods of improving computer assisted searching in bibliographic data. Senior Computing Officer Øystein Reigem at the Centre gives a summary of the speech made by Professor Gerald Salton, Cornell University, USA, on these subjects.

The basis of automatic indexing is the whole text of a short document or an abstract of a long one. A program automatically extracts index terms from this material, which are then used by a search program. A strategy for extraction includes deleting function words from the text, reducing the remaining words to roots, and deciding the value of each root as an index term. Frequent and highly frequent occurences can be improved by being automatically combined into thesuarus classes and phrases, respectively. This process results in a far richer selection of index terms than manual indexing can provide.

Salton also presented strategies for the improvement of searching techniques. One method is based on a whole scale of more or less «softened up» operators ranging from solely Boolean searching to vector searching. Another way of improving searching is the addition of «relevance feedback» to the system.

The seminar also included a discussion of problems connected with introducing automatic indexing in Norwegian libraries.

Systemutvikling: Informatikkens grense mot de «myke» fagene
System development: The border between Informatics and «soft» disciplines

The point of departure for this article by Research Fellow Tone Bratteteig, University of Oslo, is her MA thesis in Informatics: «Communication in System Development». First she gives a description of the discipline of Informatics, concentrating on the perspective of reality it mediates. System development methods usually view organizations as harmonious, and therefore contribute to the myth of «neutral» technology and «objective» expertise.

In Bratteteig's opinion, system development must be viewed as a work process, encompassing a number of other processes - one of these being communication between people. In order to specify what communication in system development is and should be, Bratteteig turned to theories expounded in socio- and psycholinguistics, of which she gives a summary. She discovered that the traditional definition of communication in these disciplines is the same as the one found in Informatics: communication equals the mere exchange of information. The components of this process are analyzed separately, simplified and formalized.

However, a holistic view of the process of communication is apparent in newer theories within socio- and psycholinguistics. Bratteteig's conclusion is that this approach is of great importance when constructing plans for communication in system development. Other disciplines can also be useful in this work, which cannot be guided by strict rules.

Større presisjon ved bruk av edb og kvantitative metoder
Greater precision with the aid of ADP and quantitative methods
Rune Johansen at the Centre interviews Roald Skarsten, who has led the Faculty of Art's ADP section at the University of Bergen since 1979. The section's services, which increasingly are in demand, include assistance to personnel and graduate students at the Faculty, arranging courses, and developing programs and systems.

A future service Skarsten would like to give high priority to is postgraduate courses in ADP for humanists. The introduction of ADP in schools and the widespread use of ADP in the private and public sectors makes knowledge of this field crucial to teachers and students alike. Humanities students need training in ADP in order to secure jobs. Above all, philologists have an important task in ensuring a balanced technological development in our society.

In Skarsten's view, the use of quantitative methods adapted to ADP in humanistic research leads to a higher degree of precision and more time for concentrating on interesting problems. Not all types of research are suited to computational methods, but in general the use of ADP simplifies the process of analysis, reflection, and verification.

Edb-utdanning for lærere
ADP for teachers
Computing Officer Eirik Lien reports on a course in ADP for teachers at the University of Trondheim. Being interdisciplinary, it avoids giving the impression that ADP is merely a technological matter.

The subjects on offer are basic and advanced data processing, ADP in education, ADP and society, graphic data processing, and real-time ADP.

«Basic data processing» teaches programming and general computer knowledge. «ADP in education» deals with, among other things, integrated instruction, and the teaching of programming and social aspects of ADP. The goal of «Computer assisted instruction» is to help teachers develop and evaluate programs and other computer based educational aids. All three subjects are based on both lectures and project work.

LEXeter '83
Professor Lars Otto Grundt, University of Bergen, reports on LEX '83, the international conference on lexicography. The conference took place in Exeter, England in September, and was attended by 250 delegates from 40 countries.

The theme of one of the five sessions was the use of ADP in the production of dictionaries. The introductory speech was made by Professor Frank Knowles. Prof. Knowles pointed out that ADP not only is an important aid in collecting, treating, and systematizing information – ADP-based products are actually in the process of taking over the functions of printed dictionaries.

Other speeches dealt with data bases and terminological analysis via ADP. All of the papers presented at the conference will be published soon.

De nordiske datalingvistikkdagene 1983
The 1983 Nordic symposium on computational linguistics
This symposium – the fourth of its kind – was arranged in Uppsala in October for 80 participants. The main theme this year was computational morphological

and syntactical analysis. Director Jostein H. Hauge at the Centre gives a summary of the speeches made at the symposium.

Speakers from Norway, Sweden, and Finland gave accounts of projects aiming at automatic morphological and syntactical analysis. It was noted that Finnish researchers have made great advances in this field, especially considering the complicated morphological system of the Finnish language.

Projects in automatic translation were also presented, including a status report on the development of the EEC's EUROTRA system. Another subject spoken on was the need for linguistic knowledge in general text retrieval systems of the future.

Proceedings from the conference, containing a number of articles in English, are due to be published soon.

Symposium om datamaskinstøttet leksikografi og terminologi
Symposium on computer assisted lexicography and terminology

This symposium, reported on by Senior Executive Officer Bjarne Norevik at the Norwegian Term Bank, was arranged in connection with the Nordic conference on computational linguistics held in Uppsala, Sweden in October. The symposium focussed on formats and methods in computer assisted lexicography and terminology. 11 papers were presented to 50 participants from the Nordic countries.

Terminologists are currently concerned with developing a joint Nordic storage format and a uniform classification system. Project groups with the aim of achieving this goal were appointed at the symposium.

The symposium also made clear that the most important task of lexicographers is to define their requirements concerning lexicographical software.

Nordiska museet intensiverer edb-virksomheten
The Nordic Museum intensifies its ADP activities

In October Humanistiske Data visited Curator Göran Bergengren at the Nordic Museum in Stockholm in order to study the museum's ADP activities. Bergengren is a member of a co-ordination group which is at work on setting up joint standards for documentation at Swedish museums.

Data processing of the museum's materials was started in 1975. In the past two years this work has been speeded up thanks to a temporary registration centre employing 21 people which was established in Älvsbyen in 1982.

The Army Museum in Stockholm has integrated ADP routines in their work too, and is in the process of developing methods for storing pictures in combination with ADP. Several county museums in Sweden are also taking ADP methods into consideration.

Nytt fra RHF/NAVF
News from the Council for Research in the Humanities

The Norwegian Research Council for Science and the Humanities is to establish an information service for ongoing research projects for a trial period of five years from January 1st. To begin with the service will be adapted to the requirements of the Councils for Research in the Humanities, Social Science Research, and Research for Societal Planning. The Centre will be responsible for technical administration and for the Humanities division of the service. The purpose of this division is to ensure that humanistic research financed by the

Council can be satisfactorily documented. It will also give overviews of Humanistic research in general. Results of this work will be put at the disposal of the Council, and conveyed to researchers, public authorities, the media, and the general public.

In 1984 the Councils for Research in the Humanities and for Social Science Research will make a study of the need for scientific equipment in their respective fields. It is assumed that requirements for technical aids within the Humanities will be on the increase, especially in Film Science, Music Science, Archaeology, and Linguistics. So far humanists have not applied to the Council for funding for this kind of equipment, presumably because they lack knowledge about the existing possibilities of use in their disciplines.

Meldinger

News

The Norwegian Historical Data Archives in Troms has received funds for further activities after its three-year trial period expires on January 1st. The Archives will continue to enter and process censuses and clerical records, with the aim of producing a national register of all persons who lived in the 18th and 19th centuries. Printed lists based on materials held by the Archives are used by historians, archives personnel, and teachers of local history.

In 1982 the museums in Hedmark County appointed a committee which has issued a report on ADP as an aid to museum work. The report gives an overview of tasks suitable for data processing, available hardware, and different systems that can be implemented. In the committee's view a project group should develop a joint ADP system that is flexible enough to suit each museum's particular needs.

The Centre has published a new report in its series: *Paleodemography* by Stig Welinder. This report contains a presentation of the program DEMO, which can be used in reconstructing the size and structure of settlement site populations. Welinder also discusses problems involved in this type of analysis, giving examples from three grave-fields in Sweden. The report is in English and costs Nkr 55.

The archaeological program package STAR (see HD 1-83) is in the process of being implemented in Oslo, Trondheim, Tromsø, and Stavanger.

At the 1983 Nordic symposium for computational linguistics a decision was made to discontinue publication of the newssheet COMPILING, due to difficulties in obtaining sufficient relevant material on a regular basis. For many years COMPILING has been a forum for information on Nordic computational linguistics, containing news, reports, and articles. The Centre is to take over COMPILING's functions by regularly incorporating information in this field in Humanistiske Data.

The purpose of the American organization CONDUIT is to promote the use of instructional computing at the collegiate level. CONDUIT assists in the use, production, and distribution of instructional computing materials, and publishes the journal «Pipeline». CONDUIT's address: P.O. Box 388, Iowa City, Iowa, 52244, USA.

Journals presented in this issue are: «Collegiate Microcomputer», which is a forum «for the exchange of ideas about the roles of microcomputers in all areas of college and university life», and «International Journal of Micrographics

and Video Technology», which reports on developments in electronic information transfer.

Forthcoming conferences:

International Conference on the Methodology and Techniques of Machine Translation - Cranfield, England, 13-15 February 1984. More information from: Douglas Clarke, Department of Mathematics, Cranfield Institute of Technology, Cranfield, Bedford, MK43 0A1, England.

Eleventh International ALLC Conference - Université de Louvrain, Belgium, 2-6 April 1984. For more information contact Dr. Jacqueline Hamesse, Institut Supérieur de Philosophie, Chemin d'Aristote,1, B-1348 Louvrain-la-Neuve, Belgium.

COLING 84 - Stanford, California, 2-6 July 1984. Information from: Dr. Martin Kay, Xerox PARC, 3333 Coyote Hill Road, Palo Alto, California 94304, USA.

Second International Conference on Automatic Processing of Art History Data and Documents - Pisa, 24-27 September 1984. Call for papers. Write to: Laura Corti, Scuola Normale Superiore, Piazza dei Cavalieri 7, 56100 Pisa, Italy.

Forts. fra 2. omslagsside.

RAPPORT nr. 14. *NOVA*STATUS HÅNDBOK*
Del 1: Søking. Brukerveiledning. 3. opptrykk februar 1983. ISBN 82-7283-011-6 Pris kr. 20.
Del 2: Fil-beskrivelser. Systemdokumentasjon. Utsolgt.
Del 3: Generering og oppdatering av databaser. Pris kr. 20.

RAPPORT nr. 15. *Ivar Fonnes: Tekstsøking på tegnnivå.* Januar 1980. ISBN 82-7283-012-4 Utsolgt.

RAPPORT nr. 16. *Årsmelding 1979.* NAVFs EDB-senter for humanistisk forskning. ISBN 82-7283-013-2 Gratis.

RAPPORT nr. 17. *Svein Lie: Automatisk syntaktisk analyse.* Del 1. Grammatikken. Desember 1980. ISBN 82-7283-014-0 Pris kr. 30.

RAPPORT nr. 18. *Datateknologi og humanistisk forskning.* Bidrag til en NAVF-utredning. Desember 1980. ISBN 82-7283-015-9 Pris kr. 30.

RAPPORT nr. 19. *Statistiske metoder på arkeologisk materiale.* Rapport fra et seminar på Bryggens museum, Bergen 24.-26. november 1980. Mars 1981. ISBN 82-7283-017-5 Pris kr. 35.

RAPPORT nr. 20. *EDB-prosjekter i humanistiske fag 1980.* Juni 1981. 2. opptrykk oktober 1981. ISBN 82-7283-018-3 Pris kr. 45.

RAPPORT nr. 21. *Rune Johansen: Bruk av EDB i teatervitenskapelig forskning.* Mai 1981. ISBN 82-7283-019-1 Pris kr. 35.

RAPPORT nr. 22. *Årsmelding 1980.* NAVFs EDB-senter for humanistisk forskning. ISBN 82-7283-020-5 Gratis.

RAPPORT nr. 23. *Stig Welinder: A program package for archaeological use.* 1981. ISBN 82-7283-021-3 Pris kr. 45.

RAPPORT nr. 24. *Rapport fra seminar om bruk av edb innen teater og teatervitenskap.* Januar 1982. ISBN 82-7283-026-4 Pris kr. 50.

RAPPORT nr. 25. *Ole Lauvskar: Diskriminantanalyse i SPSS.* Desember 1982. ISBN 82-7283-028-0 Pris kr. 55.

RAPPORT nr. 26. *Stig Welinder: Paleodemography.* Oslo 1982. ISBN 82-7283-030-2 Pris kr. 55.

RAPPORT nr. 27. *Årsmelding 1981.* NAVFs EDB-senter for humanistisk forskning. ISBN 82-7283-029-9 Gratis.

RAPPORT NR. 28. *Årsmelding 1982.* NAVFs EDB-senter for humanistisk forskning. ISBN 82-7284 31-0. Gratis

RAPPORT NR. 29, 30, 31, 32: *Stig Welinder et al.: STAR I-IV* A program package for archaeological use. Bergen 1983. Samlet pris kr. 180. (Rapportene kan også kjøpes enkeltvis).

NR. 29 STAR I Introduction and Star manual. ISBN 82-7283-033-7 pris kr. 50.

NR. 30 STAR II Student textbook and STAR examples. ISBN 82-7283-034-5 pris kr. 60.

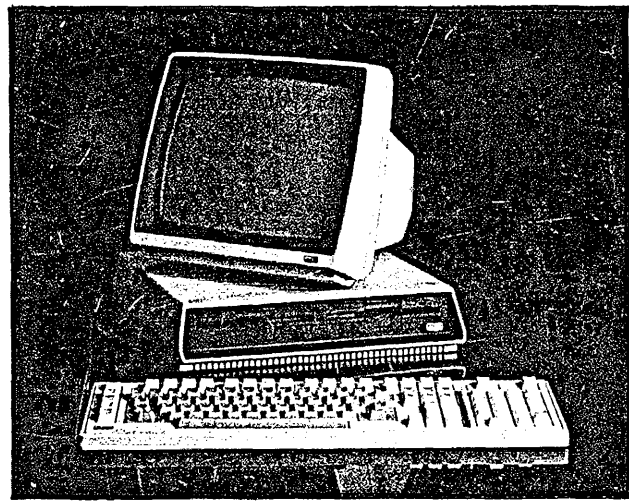
NR. 31 STAR III Archaeology for statisticians. ISBN 82-7283-035-3 pris kr. 60.

NR. 32 STAR IV STAR algorithms. ISBN 82-7283-036-1 pris kr. 30.

Vurderar du innkjøp av EDB-utstyr?

COS a/s

strative computer systemer
STRAUME
)33 14 40



ifor ikkje ta ein prat med ACOS?
kan hjelpa deg gjennom edb-jungelen, og saman kan vi finna ei løysing som ver nett for deg.
COS har stort utval i maskin- og programvare, og kan levera løysingar som er ckande for dei fleste.

re datamaskiner 1-50 arbeidsplassar.
TOS og DISCOVERY.

å datamaskiner (1 arbeidsplass).
TOS, EPSON, HEWLETT PACKARD og NEC.

rtable datamaskiner.
SON, KAYPRO og PIED PIPER.

kre av våre maskinleverandørar:
FONOR, STORM SYSTEM, SCANVEST-RING, HEWLETT PACKARD, NDBERG DATA, DATAHUSET, NORSK MARKONI, TECKNITRON, SON DATASENTER.

kre av våre programleverandørar:
AK, COMPUTAS, A/S EDB, A/S SYSTEMUTVIKLING, MIKROFORM, B-TEAM, TEKNISK PROGRAMUTVIKLING, GG-DATA, TEKNISK TA, ING. GRØNER, ZIOLOCO, MIKROKONSULT, LANDBRUKETS TASENTRAL.

Vi har alt fleire hundre brukarar på Vestlandet – blir du ein av dei?

DU HAR OPPGÅVA – VI HAR LØYSINGA.