humanistiske data

Utgitt av NAVF.s EDB-senter
for humanistisk forskning
Bergen

The Norwegian Computing Centre
for the Humanities

NORGES ALMENVITENSKAPELIGE FORSKNINGSRÅD

Artikler
Konferanserapporter
Meldinger
Summary

NR.
1
1980

HUMANISTISKE DATA is published by the Norwegian Computing Centre for the Humanities. — The Editor is: *Jostein H. Hauge*, Director of the Centre. *Issues are free. Contributions are welcome.*

HUMANISTISKE DATA blir utgitt av NAVF.s EDB-senter for humanistisk forskning i Bergen. Senterets leder, *Jostein H. Hauge,* har det redaksjonelle ansvar for bladet. De som ønsker å få bladet tilsendt, kan bestille det ved henvendelse til senterets adresse: Villavei 10, Boks 53, 5014 Bergen-Universitetet.

Innlegg kan sendes til samme adresse.

Merk ny adresse: Harald Hårfagresgt. 31, boks 53, 5014 Bergen-Universitetet.

INNHOLD

MEDARBEIDERE I DETTE NUMMER:

Ove Magnus Bore, kulturvernkonsulent, Fylkeskonservatoren i Hordaland, Bergen
Håvard Hjulstad, konsulent, Norsk termbank, Universitetet i Bergen
Knut Hofland, konsulent, NAVFs EDB-senter for humanistisk forskning, Bergen
Rune Johansen, forskningsstip. (NAVF), Teatervit. inst., Universitetet i Bergen
Stig Johansson, dosent, Britisk institutt, Universitetet i Oslo
Eirik Lien, konsulent, EDB-tjenesten for HF, NLH, Universitetet i Trondheim
Roald Skarsten, konsulent, EDB-seksjonen v/ HF, Universitetet i Bergen
Gunnar Thorvaldsen, vik. amanuensis, Inst. for samf.vit., Universitetet i Tromsø
Per Vestbøstad, konsulent, Norsk tekstarkiv, Universitetet i Bergen
Stig Welinder, dosent, Oldsaksamlingen, Universitetet i Oslo

Redaksjonen avsluttet 15.7.1980

SEFRAK. Prøveprosjekt for databehandling av kulturminneregisteret.

Ove Magnus Bore

BAKGRUNN FOR PROSJEKTET

"Registrering av faste kulturminner i Norge" er et landsomfattende prosjekt som tar sikte på å registrere alle faste kulturminner i Norge fra tiden etter reformasjonen og fram til 1900. Registreringen administreres av et sekretariat under Miljøverndepartementet, "Sekretariatet for registrering av faste kulturminner i Norge" (SEFRAK). Organiseringen av arbeidet i det enkelte fylke er tillagt fylkeskonservatorene. Etter enkelte prøveregistreringer startet prosjektet for fullt i 1975. I 1979 var arbeidet igang i 154 kommuner.

Registreringen omfatter i hovedtrekk fire arbeidsoperasjoner: utfylling av skjema, oppmåling og tegning av grunnplan, kartfesting og fotografering. Det er utarbeidet to typer skjema, ett for registrering av hus og ett for registrering av andre faste kulturminner. Hvert skjema inneholder 44 rubrikker. Disse kan deles inn i følgende hovedavsnitt: arkiv- og kartreferanser, lokalisering, funksjon, miljø, byggemåte, alder, vedlikehold, sikring og andre opplysninger. For hvert kulturminne fylles det ut ett skjema.

Det har vært en forutsetning under planleggingen av prosjektet at de store datamengdene man her vil stå overfor kun kan behandles tilfredsstillende med EDB. Utarbeidelsen av registreringsskjemaene ble derfor gjort i samarbeid med ekspertise innen databehandling. Resultatet er et skjema der den vesentligste del av opplysningene kodes numerisk. En mindre del av opplysningene framkommer ved avkrysning av alternative svar, mens en relativt stor del av skjemaet er åpen for fritekst. Friteksten er oftest knyttet til den kodete teksten som en utdypning eller presisering av kodeinnholdet, eller den fungerer som en ren tilleggsopplysning. For fastsetting av koder er det utarbeidet et kodesystem med tilhørende kodeliste.

Da prosjektet ble satt igang, var databehandlingen av materialet og rutinene for denne hverken utprøvd eller fastlagt på noe vis. I

1975-77 gjennomførte Sekretariatet et prøveprosjekt utført ved NIBR ved bruk av databehandlingssystemet NIMS. Prosjektet framstilte fem forskjellige typer dataprodukt: arkivkort, katalog, statistikk-tabell, oversiktskart og områdeanalyse. Prosjektet ga et tilfredsstillende resultat for de tre siste produktenes vedkommende. Når det gjalt utarbeidelse av arkivkort og katalogprodukter stod det fremdeles endel problemer igjen å løse.

Hordaland er et av de fylker i landet der progresjonen i kulturminne registreringen har vært størst. Etterhvert som materialet hopet seg opp, ble det et stadig mer tvingende behov for å få det over i en form som kunne tjene de forskjellige brukerkategoriene, spesielt innenfor den enkelte kommuneadministrasjon. NAVFs EDB-senter og Fylkeskonservatoren har i en årrekke hatt et samarbeide i forbindels med databehandling av gjenstandsregistreringer og museumskatalogiseringer (KGR-materiale). På grunnlag av positive erfaringer fra dette samarbeidet tok Fylkeskonservatoren i 1978 initiativ overfor EDB-senteret for å få igang et prøveprosjekt for databehandling av kulturminneregisteret. Senteret på sin side sa seg interesert i prosjektet, og det finansielle ble ordnet med en kostnadsfordeling mellom Fylkeskonservatoren, EDB-senteret og SEFRAK.

I sin behandling av saken la Sekretariatet vekt på at prosjektet i størst mulig grad skulle følge opp NIMS-prosjektet og minst mulig gjenta dette. Følgende målsetting ble dermed satt for prosjektet:

1) Utprøving av datarutiner (Dataregistrering, programmering, utskriving)
2) Utprøving av dataprodukt og presentasjonsformer (Arkivkort i klartekst, listeprodukt)
3) Utprøving av databehandling av kulturminnematerialet gjennom fulltekstsøkesystemet NOVA*STATUS
4) Vurdering av tidsfaktor og kostnader

Som grunnlagsmateriale for prosjektet valgte man registreringsmaterialet fra Fitjar kommune som var ferdigregistrert i 1977 - 763 objekt. I tillegg tok man med materialet fra en registreringskrets i Øygarden kommune - 198 objekt. Alle objektene var bygninger.

Leder for prosjektet var fylkeskonservator Nils Georg Brekke. Ellers deltok i prosjektgruppen kulturvernkonsulent Ove Magnus Bore ved Fylkeskonservatorens kontor, EDB-konsulent Sigbjørn Århus og driftsassistent Per Vestbøstad fra NAVFs EDB-senter.

Prosjektet ble avsluttet med et møte 30/10-79 der resultatet ble lagt fram for styret for Sekretariatet og Miljøverndepartementet. Materialet fra Fitjar er nå ute til bruk i kommunen. Man tar sikte på å holde et nytt møte i løpet av høsten 1980 der prosjektet bl.a. vil bli presentert for pressen.

Prosjektet er beskrevet i rapport nr. 8 i NAVFs EDB-senters rapportserie (SEFRAK. Rapport frå prøveprosjekt for databehandling av kulturminneregisteret).

UTPRØVING AV DATARUTINER

Punching av materialet ble utført på Fasit 6120 papirbåndpunch. Samtidig som dataene ble registrert på hullbånd, ble det med skrivemaskin fylt ut et arkivkort. Kortet var konstruert etter mønster av registreringsskjema med samme rubrikkinndeling. Det var videre laget slik at det tilfredsstilte tekniske krav for optisk lesing. Man anså dataregistrering via optisk lesing som den mest realistiske registreringsmetode for endel fylker de nærmeste år. Punchekortet vil således kunne bidra til at overføring av materialet til maskinlesbar form kan starte opp parallelt en rekke steder i landet uten særlig investeringer i teknisk utstyr. Valget av registreringsmetode hang ellers sammen med den årelange erfaringen man satt inne med i bruk av papirbåndpunch. Metoden var allerede brukt en rekke år i forbindelse med KGR-materialet. Rutinene var således godt innkjørt, og man hadde tilgang på personell som var godt innlært i bruk av systemet. Den registreringsmetode som etter all sannsynlighet vil bli brukt på kulturminnematerialet ved et framtidig landsomfattende driftsopplegg er punching direkte på terminal. Metoden er tilfredsstillende utprøvd i andre sammenhenger og vil antagelig passe kulturminnematerialet godt. Man fant imidlertid ingen grunn til å komplisere gjennomføringen av prosjektet ved å ta i bruk systemet. Utprøving av dette ble således utsatt til neste fase i utviklingen av driftsrutinene.

Ved overføring til platelager benyttet man samme program som ble laget for innlesing av KGR-materialet. Korrekturutskrifter ble gjort med et standard UNIVAC-program, DATA. Dataene ble siden overført og lagret på magnetbånd. Mesteparten av korrekturen foregikk interaktivt, og til retting benyttet man et annet standard UNIVAC-program, MED.

Det videre programmeringsarbeid ble utført innenfor programsystemet NOVA*STATUS som inneholder de sorterings- og redigeringsfunksjoner man trenger for å lage sorterte kataloger og selektive utskrifter. Følgende program ble laget:

1) STATUS FORM
   Programmet organiserer materialet i den form man ønsker det skal ha på dataskjermen ved interaktiv søking. Programmet setter ful stendig fortekst på hver rubrikk og rubriserer innholdet i faste kolonner.

2) SEF KORT
   Programmet formaterer data for hvert objekt slik at de passer inn på det ferdigtrykte EDB-kortet som ble utarbeidet i prosjektet. Programmet fjerner fortekstene ettersom de er påtrykt kortet. Videre deschiffrerer det alle numeriske koder.

3) PAPIR KORT
   Programmet formaterer data i de sorterte katalogene etter 110 kolonners bredde for utskrift på vanlige EDB-lister. Også dette programmet tolker de numeriske kodene.

I programarbeidet la man spesiell vekt på å systematisere data på en måte som ga de ferdige produkter høy grad av leselighet. Dette ble bl.a. gjort ved gjennomgående rubrisering, ved å skille fortekstene fra teksten med bruk av dobbel skrift og ved å skille klar tekst og fritekst ved bruk av henholdsvis store og små bokstaver.

UTPRØVING AV DATAPRODUKT OG PRESENTASJONSFORMER

Kulturminneregistreringen tar sikte på å dekke tre bruksområder: Offentlig fysisk-økonomisk planlegging, forsknings- og opplysnings-

arbeid og antikvarisk vernearbeid. Disse bruksområdene omfatter et meget stort spekter av brukere og bruksmotiver, og dataproduktene bør derfor ha en form som gjør den egnet til bruk over et bredt spekter.

Arkivkortet som ble fylt under punching forelå nærmest som et biprodukt av dataregistreringen. Kortet har plass for foto og grunnplan og vil til en viss grad kunne fungere tilfredsstillende i et register. Kortet inneholder imidlertid kun kodete opplysninger i tillegg til eventuell fritekst, og setter således krav til bruk av kodebok for å kunne nyttes fullt ut. Opplysningene er ikke korrekturlest og rettet. All oppdatering må gjøres manuelt og kortet kan ikke reproduseres.

Disse problemene ble løst gjennom konstruksjon av et EDB-kort, - et arkivkort som fylles ut datamaskinelt. Kortet ble laget etter mønster av registreringsskjema og punchekort med samme rubrikkinndeling og trykk på begge sider. Kortet inneholder både klartekst og fritekst. D.v.s. at alle numeriske koder er deschiffrerte. Kortet blir skrevet ut i en fase der data er korrekturlest og rettet. Det kan produseres i x-antall eksemplarer og oppdatering skjer maskinelt bl.a. ved sammenkobling med andre relevante registre. Det har videre den fordel at det ved hver utskrift kan sorteres etter de ønskete kriterier.

Denne utskriftsformen på dobbelsidig rubrisert arkivkort var tidligere ikke utprøvd på EDB-senteret og kan således karakteriseres som banebrytende.

I tillegg til EDB-kortet ble det produsert en rekke katologer. Katalogene ble skrevet ut både som fulle utskrifter og som referansekataloger. Rubrikktitler, koder, klartekst og fritekst er skilt fra hverandre med de virkemidler som er nevnt ovenfor. I tillegg er de emnemessige hoveddelene skilt med horisontale linjer. Objektene er skilt fra hverandre ved at hvert objekt går over to sider uansett tekstlengde. Det ble i prosjektet skrevet ut to fulle utskrifter - en topografisk katalog (sortert etter matrikkelnr.) og en typologisk katalog.

Referansekatalogene inneholder et mindre utvalg av data og skal i første rekke tjene som nøkkel til hovedarkivet (enten i form av kort arkiv eller EDB-katalog). Det ble skrevet ut tre slike kataloger - en kronologisk, en etter arkivnummer og en etter fotonummer.

Til slutt ble det laget et eksempel på en kombinert presentasjonsfor Dette ble gjort i form av en katalog der data og foto og våningshus i Fitjar sentrum ble stilt sammen. Man ville med dette vise hvordar man på en svært enkel måte kan sette sammen deler av et materiale ti et produkt med høy informasjonsverdi.

UTPRØVING AV DATABEHANDLING AV KULTURMINNEMATERIALET GJENNOM FULL-TEKSTSYSTEMET NOVA*STATUS

De arkivkort og katalogkort som er nevnt ovenfor ble alle produsert gjennom programsystemet NOVA*STATUS. De samme resultatene kan imidlertid også oppnås ved en rekke andre programsystemer. Hovedårsaken til at man valgte å behandle materialet med et fulltekstsøkesystem, var de muligheter systemet gir for søking i fritekst ved interaktiv databehandling. Det er en forutsetning for å nytte materialet fullt ut at den datamengden som ligger i tilleggsopplysningene er tilgjengelig for søking.

Under arbeidet med kulturminnematerialet vil man vekselvis ha behov for å søke i tallkoder og i fritekst. Man vil også ha behov for å søke i kombinasjoner av tallkode og tekst. Det vil videre i en rekke brukssituasjoner stilles krav til rask besvarelse av en rekke alternative spørsmål. De prøver som ble gjennomført i prosjektet, viste at anvendelsen av et fulltekstsøkesystem fungerte meget tilfredsstillende til dette bruk. Fleksibiliteten i systemet gir brukeren mulighet til raskt å klarlegge trekk og nyanser i et meget stort materiale.

I et stort kulturminneregister vil man finne utallige språklige nyanser og rene stavefeil. Registreringen foregår på begge målforme og samme ord kan ha en rekke forskjellige skrivemåter. På grunn av mangel på en fast fagterminologi vil man også finne en rekke begrepsvariasjoner. De mulighetene som ligger i NOVA*STATUS til

å søke på trunkerte ord løser noen av de problemene som ligger i dette. Man kan komme enda et skritt nærmere en løsning ved å legge inn en synonymordliste med et relevant begrepsapparat.

I arbeid med et kulturminneregister vil det ofte være ønskelig med forskjellige statistiske produkt som f.eks. krysstabeller sortert etter funksjon og tid. Ettersom slike produkter var framstilt med et tilfredsstillende resultat i det første prøveprosjektet, var det ikke aktuelt å gjenta dette i det nye prosjektet. NOVA*STATUS er ikke bygd opp med tanke på statistikk. Det er imidlertid klart at det er en relativt enkel operasjon å bygge på et program som gir en del aktuelle statistiske produkter.

VURDERING AV TIDSFAKTOR OG KOSTNADER

Prosjektet ble gjennomført med en kostnadsramme på kr. 33.000. De forskjellige utgiftene hadde følgende prosentvise fordeling:

Overføringskostnader	46%
Datakontroll	17%
Programmering	23%
Rekvisita	14%

I tillegg til disse kostnadsførte operasjonene gikk en del av arbeidet inn i de daglige driftsbudsjett ved EDB-senteret og Fylkeskonservatorens kontor.

Ved overføringen ble det punchet 5 objekt pr. time. Ved punching direkte på terminal vil tempoet kunne økes, muligens fordobles. Dersom en legger inn kontroller, vil tempoet igjen synke noe. Man vil da imidlertid minske behovet for korrekturlesing og retting. Andre usikre momenter for tid- og kostnadsberegninger kan nevnes. I prosjektet har man f.eks. dratt nytte av universitetets rabattordninger for bruk av dataanlegget. Man har ellers hatt tilgang på billig arbeidskraft gjennom bruk av studenter og sivilarbeidere i enkelte arbeidsoperasjoner. Videre må en del av arbeidet sees på som et utviklingsarbeid og følgelig karakteriseres som en engangsoperasjon. Dette gjelder særlig utvikling av program og korttyper. Til slutt

kan nevnes at utgifter til rekvisita er unormalt høye p.g.a. små opplag av trykking av kort.

På denne bakgrunnen er det klart at prosjektet gir lite grunnlag for eksakte kostnadsberegninger for en framtidig drift. Disse kan først gjøres når databehandling av kulturminneregisteret gjennomføres etter de driftsrutiner departementet legger opp til. Prosjektet forteller likevel nok om tidsfaktoren i de enkelte operasjonene til at man kan gjøre visse grove forhåndskalkyler i en oppstartingsfase.

KONKLUSJON

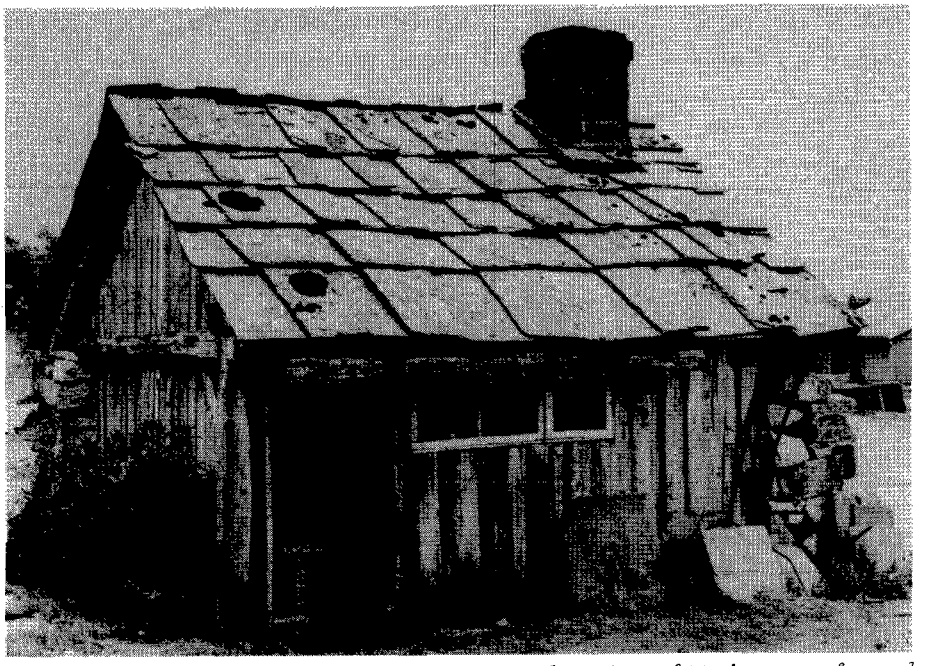
I den organisasjonsmodell for kulturminneregisteret som Miljøvern-departementet etter anbefaling fra Sekretariatet går inn for, tar man sikte på at dataregistreringen skal foregå ved fylkeskartkontorene men den videre databehandling og lagring av materialet skal legges til de interkommunale datasentralene. Man tar sikte på å koble registeret sammen med GAB-registeret (Grunn-, adresse- og eiendomsregisteret). Sammenkoblingsformen er ennå ikke avgjort.

Prøveprosjektet har på en tilfredstillende måte løst en rekke av de problem som gjensto å løse før ordinær drift kunne settes i gang. Prosjektet har utarbeidet relevant utskriftsprodukt og de spesialprogram som behøves for å skrive ut disse. Det har vist hvordan produktene kan gies full informasjonsverdi ved å oversette kodete opplysninger i klartekst. Og det har vist hvordan materialet kan brukes interaktivt og hvordan bruken av et fulltekstsøkesystem gjør det mulig å benytte materialet fullt ut ved å søke i fritekst.

Departementet har i 1980 tatt initiativ til et tredje prøveprosjekt for databehandling av kulturminneregisteret. Det er nedsatt en arbeidsgruppe med representanter fra Fylkeskartkontoret i Hordaland, Sekretariatet for registrering av faste kulturminner i Norge, Historisk museum i Bergen og Fylkeskonservatoren i Hordaland. Prosjektet bygger videre på det foregående og tar sikte på å gjennomprøve driftrutinene etter departementets modell. Materialet vil bli punchet på Fylkeskartkontoret direkte på terminal. Det vil så bli overført og behandlet videre på Kommunedata Vestlandets dataanlegg. I denne

utprøvingen ligger bl.a. overføring av de programmer som er utarbeidet i prøveprosjektet fra Univac til IBM-anlegg. NOVA*STATUS kan ikke brukes på IBM-maskiner. IBM har imidlertid et eget programsystem STAIRS som i hovedtrekk er et parallelt system.

Det nye prøveprosjektet vil videre ta opp en utprøving av databehandling av arkeologiske registreringer og en samkjøring av disse med kulturminneregisteret. Man tar også sikte på å prøve ut databehandling av den del av kulturminneregisteret som karakteriseres som "andre faste kulturminner". En del av en kommune i Hordaland hvor det er gjort både registrering av hus og andre faste kulturminner og registrering av fornminner vil bli valgt ut til formålet. EDB-senteret vil ikke delta aktivt i prosjektet, men har sagt seg villig til å yte den konsulenthjelp som måtte være nødvendig for gjennomføringen.



En omfattende registrering av bygninger har foregått i mange år under ledelse av Sekretariatet for registrering av faste kulturminner. Eldhus, Landa, Fitjar kommune.

Bruken av EDB i teatervitenskapelig forskning.

Rune Johansen

Innen norsk teatervitenskapelig forskning er scenografien et "forsømt" område. En av årsakene til dette er at det scenografiske materiale er spredt i forskjellige institusjoner, og det har heller ikke blitt katalogisert med det mål for øyet å lage en total oversikt. Derfor vil en forsker, som gir seg i kast med denne siden av teaterproduksjonen, bruke uforholdsmessig lang tid på å finne frem til et adekvat analysemateriale. I undervisningssituasjonen fører dette til at man må bruke bilder fra utenlandsk teater for å forklare enkelte sider ved den norske teateraktiviteten.

Når jeg, som en del av mitt prosjekt, har valgt å bruke EDB-behandling av materialet er det fordi jeg ønsker å dekke to hovedbehov: 1) Lage et anvendelig kartotek over det materialet jeg selv skal forske i. 2) Danne basis for et offentlig arkiv til bruk for forskere eller institusjoner som måtte ha interesse av norsk scenografi.

Materialinnsamlingen, avfotograferingen og katalogiseringen av materialet er et tidkrevende arbeide, men jeg har den fordelen at jeg, i mine studier, har arbeidet med norsk scenografi. I denne prosessen har jeg allerede registrert hovedtyngden av materialet. Dermed blir ikke lokaliseringen av billedstoffet tidkrevende, og jeg kan istedet konsentrere meg om avfotografering og katalogisering.

For ytterligere å redusere tidsbruken i arbeidet med dette "biproduktet", har jeg valgt å begrense meg til oppsetninger på Nationaltheatret i perioden ca. 1908 - 1935. Teaterhistorisk sett er denne perioden viktig, fordi konflikten mellom tradisjon og "modernisme" innen teatret kommer klart til syne. Et annet moment er at National theatrets dekorasjonsvesen i perioden 1899 - 1935 ble ivaretatt av to personer: Jens Wang 1899 - 1917 og Oliver Neerland 1917 - 1935. Dette gir muligheter til å studere hvordan de to malerne arbeidet innen den samme institusjonen, og hvordan norsk dekorasjonskunst utviklet seg i en periode på 35 år.

Henrik Ibsen,
"Kongsemnerne"
10.10.1900 4. akt.
Scene Oslo Kongsgård

Da billedarkivet skulle tilrettelegges for offentlig bruk, ble dataregistreringen et problem. Systemet måtte gjøres enkelt å bruke og å videreutvikle. Samtidig skulle det dekke mine egne behov. Resultatet ble et innregistreringsskjema delt inn i 16 informasjonskategorier. Fjorten av disse henviser direkte til bildet, f.eks. spillested type scenografi, stiltype, dramaforfatter og scenograf. De to øvrige dekker kilder og korrespondanse. Teoretisk sett kan alt scenografisk materiale (også andre typer bilder av interesse) legges inn i dette arkivet. Sett ut fra en forskningsmessig situasjon ligger de største mulighetene i de to sistnevnte informasjonsgrupper. Disse kan utvides i et ubegrenset omfang. I EDB-programmet er det bygget inn muligheter til å arbeide interaktivt i kilde- og korrespondansematerialet.

Ved bruken av EDB kan man få kontroll og oversikt over et materiale av en størrelsesorden man hittil ikke har kunnet gi seg i kast med. Når datamassen er punchet inn i maskinen tar det bare sekunder å finne frem til det dokumentet man ønsker. Innregistreringen av data er et møysommelig og tidkrevende arbeide, men maskinens hurtige databearbeidelser tjener raskt inn tiden man har brukt til punching. Fordi man kommer raskt til opplysningene får forskeren mer tid til selve forskningsarbeidet.

Programmet gir muligheter for å ordne hver av de fjorten nevnte

informasjonsgruppene enten alfabetisk eller kronologisk. Derfor kan man nærme seg materialet fra flere innfallsvinkler enn den scenografiske.

Et annet ønske i forbindelse med bruken av EDB i teatervitenskapeli forskningsområder av kvantitativt tilsnitt er forskernes plikt til å tilrettelegge det gjennomgåtte materialet til bruk for andre som kan dra nytte av det. Et generelt trekk i dagens situasjon er at forskeren arbeider med sitt "opus magnum" og har liten tanke for å gjøre sitt kildemateriale tilgjengelig for andre. Dette er en sløsing med ressurser. Andre forskere må spille mye tid på å finne frem til et adekvat materiale selv om de arbeider innen samme tema som sin "forgjenger". Denne prosessen har sine positive sider, men har lite å gjøre med selve forskningen. Man kan her innvende at litteraturlisten og noteapparatet dekker det behov jeg har påpekt. Til en viss grad er dette riktig, men hva så med det materialet man ikke har benyttet i avhandlingen? Det reiser seg et spørsmål av sosialt og kollegialt tilsnitt når forskeren legger henvisningene til det ubenyttede materialet i sitt skrivebord, med tanke på å benytte det til senere publikasjoner.

Et annet moment ved min bruk av EDB er å gi et konkret materiale so kan danne grunnlaget for en dyptloddende diskusjon om bruken av EDB til tilretteleggelsen av instituttets eget teaterarkiv, og hvordan systematiseringen av et slikt materiale skal koordineres på landsbasis.

Er tilrettelegging av primærkilder meriterende arbeid?

Eirik Lien

All forskning bygger på data i en eller annen form. Data kan være representert som bokstaver på et papir, som lydsvingninger i lufta, som kvikksølvhøyde på et termometer, som tanker hos en forsker eller som magnetiserte felter på et magnetbånd. I mange tilfeller vil det være behov for å overføre data fra ett medium til et annet - enten for å sende en kopi til et annet sted eller for lettere å bearbeide den (kartotekkort, f.eks.). En slik operasjon vil alltid kreve arbeid og som regel tolking. I hvilken grad dataene skal tolkes, avhenger av tilstanden til lagringsmediet (utvisket blekk, lakuner), kunnskap om det dataene står for ("kjenne koden"), de ønsker en har for representasjon av dataene på det nye lagringsmediet og om dataene skal systematiseres eller få en annen kode før overføring. I tillegg må en kjenne den teknikken som brukes for å kunne få dataene registrert på det nye mediet.

Humanistiske forskere har ofte arbeidet med primærkilder som sitt datagrunnlag, eller med kopier eller avskrifter av disse. Primærkilder er av og til av praktiske årsaker skrevet av og gitt ut som bok. En annen metode er å ta fotostatkopi direkte av primærkilden. Begge disse metodene krever en god del arbeid, spesielt den første hvor det kan bli snakk om f.eks. å tyde eller tolke utydelig skrift, eller velge den riktige tolkningen av flere mulige. Dessuten vil en ofte i en trykt utgave stille opp opplysninger på en mer oversiktlig måte. Alt dette sparer senere brukere for mye arbeid, ved at dataene er blitt lettere tilgjengelig.

Mange kilder blir brukt av få, men en del sentrale kilder har stort bruksområde, både ved at mange bruker dem og at flere fagfelt kan dra nytte av de opplysningene som ligger der. Slike kilder kan f.eks. være skatte- og matrikkellister, folketellinger, diplomer og ordbøker.

I et prosjekt vil en god del av tida gå med til å systematisere dataene for lettere å finne grunnlag for de sammenhengene og opplysningene en er på jakt etter. Ut fra samme systematisering kan en

trekke mange forskjellige slutninger. Dette fører til at sentrale kilder kan bli systematisert av flere forskjellige personer på samm måte eller svært like måter.

Systematisering er som regel en tidkrevende oppgave hvis den gjøres for hånd. I de siste åra er EDB for alvor lansert som et hjelpemiddel også for humanistiske forskere.

Det en imidlertid ganske snart oppdager, er at det er uhyre tidkrevende å gjøre dataene EDB-tilgjengelig. Det er først en nitid registrering av dataene ut fra et forelegg ("punching") og etter de en langsommelig prosess med korrigering. Velsignet er derfor den forsker som bare kan ta en kopi av et allerede EDB-tilgjengelig materiale. For er dataene først EDB-tilgjengelig, er det som regel en relativt grei prosess å systematisere dem slik at de er tjenlig for det aktuelle prosjektet.

Etter hvert er det laget EDB-versjoner av en del sentrale kilder, ei del er under registrering nå og en god del er planlagt. De som alle rede fins er f.eks. folketellinga fra 1801, folketellinga i 1875 for Christiania, Tiendepenningskatten 1520/21, en god del norsk fiksjons prosa og sakprosa, noen aviser og tidsskrifter og noen trontaledebatter - bare for å nevne noe. Bak disse registreringene ligger det mange årsverk.

Ut fra beskrivelsen i innledninga er disse registreringene egentlig ei kopiering. Men det er ikke bare ei mekanisk kopiering - i overføringa ligger det også tolkning og systematisering som krever fagl innsikt. M.a.o. hadde det for mange av disse kopieringene ikke vært mulig å få en EDB-versjon uten sterk medvirkning av personer som er faglig kvalifisert både i EDB og ett eller flere humanistiske fag.

Men hva er så vitsen med en slik nitid overføring, tolking, systematisering og korrigering hvis dataene ikke blir brukt? Av fare for at interessante problemstillinger skal bli "brukt opp" av andre forskere før registratoren sjøl er kommet i gang med den egentlige delen av prosjektet, nemlig bearbeiding av dataene, er han svært til bakeholden med å la andre få adgang til å kopiere EDB-versjonen.

I vårt akademiske miljø hvor kvalifikasjoner måles ut fra hva som er produsert av vitenskapelig arbeid, må en forsker som har lagt ned mye arbeid i å gjøre data EDB-tilgjengelig, reservere bruken av dem for seg sjøl. Poenget her er at registreringsarbeidet ikke regnes som meriterende ved søknad til vitenskapelige stillinger. Dermed får vi den paradoksale situasjonen at de årsverk som registrator har lagt ned i arbeidet, ikke kommer andre enn han sjøl til gode og da bare hvis han går videre med arbeidet sitt. Det letter ikke arbeidet for andre enn registratoren til tross for at dataene er på et medium som er svært lett å kopiere. På denne måten får en ingen ekstra-gevinst av det kolossale arbeidet som er nedlagt.

Den eneste måten å unngå en slik situasjon på, er å anerkjenne også dataregistrering av denne typen som det det er - nemlig forskning. Innsatsen kan måles både ut fra faglig humanistiske forskningskri-terier og ut fra EDB-baserte kriterier.

En "frigiing" av EDB-versjoner av sentrale kilder vil lette bruken for andre, og registrator blir kreditert for arbeidet. Dermed blir det også mulig å ha kopier av disse kildene ved f.eks. EDB-tjenestene ved de enkelte universitetene. Sjøl om materialet blir tilgjengelig for andre, bør det likevel være mulig for registrator å reservere visse bruksområder for et gitt tidsrom. EDB-tjenestene kan sørge for at dataene ikke blir brukt ut over en slik reservasjon.

Ved å anerkjenne kvalifisert dataregistrering som vitenskapelig arbeid er det mulig å få full nytte av dataene. Da slipper andre å tigge og be om å få kopi av EDB-versjonen - og får en dem, er den kanskje skrelt ned til et absolutt minimum av opplysninger. Og det må jo være godt for en som har lagt ned års arbeid i et produkt, å vite at andre har glede og nytte av det.

A computer program package for archaeological use.

Stig Welinder

The program package presented here is designed for the statistical and quantitative analysis of archaeological data. In the first place it is intended to be used in teaching and in various kinds of demonstrations of quantitative methods and computer technique. Its usefulness in actual research is restricted.

The package consists of programs based on the SPSS package (Nie et al. 1975), programs borrowed from other scholars, and programs written especially for the package by the author in collaboration with Ivar Fonnes, computer supervisor at the Faculty of Arts, Unive: sity of Oslo. The package contains programs for the following methods:
Descriptive statistics including bivariate analysis
Principle components analysis
Factor analysis
Cluster analysis
2 methods of seriation

The ambition of the program package is that every analysis of a corpus of archaeological source material shall end in a graphic representation of the data and their inherent structure.

In its present form the package consists of 19 separate FORTRAN pro· gram files. From the start with an indata file organised as a tabl(in fixed format according to the requirements of the SPSS package, the files of package are to be used in the sequence of Figure 1. The output from one file is automatically used as input for a succeeding file. The package contains the following programs:

A. Data defining programs

DATA1 transforms various kinds of data to present absence matrices. It is one example of many possible transformations.

HELP1 creates the file to be used when defining an SPSS system file. This program as well as the other programs based upon the SPSS package are meant to faciliate the use of the latter for archaeological purposes.

HELP3 transforms an input data file to the input format of the program JOZEF

B. Descriptive statistics. Uni- and bivariare analyses.

CONDES.SPS a readymade file to be used when running the SPSS subprogram CONDESCRIPTIVE

FREQ creates the file to be used when running the SPSS subprogram FREQUENCIES

SCATT creates the file to be used when running the SPSS subprogram SCATTERGRAM

CROSS creates the file to be used when running the SPSS subprogram CROSSTABS

The subprograms CONDESCRIPTIVE and SCATTERGRAM are used for interval and ratio data, and the subprograms FREQUENCIES and CROSSTABS for nominal and ordinal data.

C. Multivariate analyses.

PRINC creates the file to be used when running the SPSS subprogram FACTOR using principal factoring without iterations and no rotation, i.e. calculating principal components (Doran et al. 1975, pp. 190-197)

HELP2 prepares the output from the SPSS subprogram FACTOR for the program GRAPH1

GRAPH1 makes a diagram of the grouping of the input data according to principal components (cf. Doran et al. 1975, Fig. 9.8 (a))

FACTOR creates the file to be made when running the SPSS subprogram FACTOR using principal factoring without iterations and VARIMAX rotation.

JOZEF seriates a presence/absence matrix according to the method described by Saers 1978 (the actual program has kindly be submitted by J. Saers)

GRAPH2 makes a graphical representation of the output from the program JOZEF

CORR1 calculates correlation coefficients from presence/absen data, simple matching coefficients (Doran et al. 1975, pp. 140-141) and Jaccard coefficients (Doran et al. 197 p. 141)

CORR2 calculates robinson similarity scores (Doran et al. 197 pp. 139, 272-273)

CLUST makes a cluster analysis from the correlation matrices calculated by the programs CORR1 and CORR2. The UPGW method by Sokal et al. 1973 is used

GRAPH3 makes a graphical representation of the output from the CLUST program

GELF seriates the correlation matrices calculated by the programs CORR1 and CORR2 according to the metod by Gelfand 1971

GRAPH4 makes a graphical representation of the output from the GELF program

The program package is presently available at the DEC-10 computer of the University of Oslo. All of the package can easily be transformed to any computer center using the SPSS package. Parts of it can also be transformed to other centers (the program JOZEF cannot be submitted without permission from the author of that program. duplicated manual for the use of the package in Oslo is available.

Example:

Four types of decoration (cord, lines, twisted cord, others) are distributed among potsherds at 10 Middle Swedish Early Neolithic sites according to the following percentages

Vallby		5		5	89
Hjulberga	1:A	9	2	61	28
Hjulberga	1:B	19	3	50	28
Hjulberga	1:C	38		42	21
Hjulberga	2:A	53			47
Hjulberga	2:B	36	7	3	55
Hjulberga	2:C	32	24	12	32
Brokvarn		42	1	20	37
Østra Vrå		18	11	32	39
Mogetorp		33	31		36

The program CORR2 calculates Robinson similarity scores according to the following:

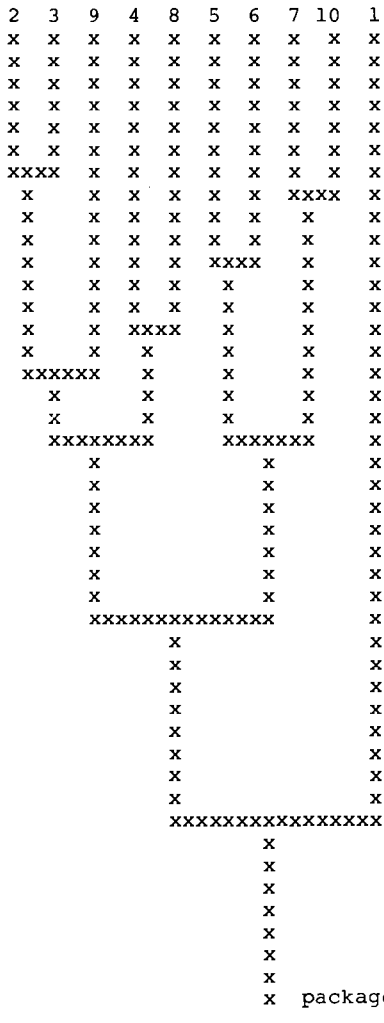
200	77	77	62	105	126	85	95	99	83
77	200	178	143	74	83	102	116	142	78
77	178	200	163	94	105	124	136	162	100
62	143	163	200	117	118	129	157	141	107
105	74	94	117	200	165	128	158	114	138
126	83	105	118	165	200	147	153	133	151
85	102	124	129	128	147	200	154	146	176
95	116	136	157	158	153	154	200	152	140
99	142	162	141	114	133	146	152	200	130
83	78	100	107	138	151	176	140	130	200

The program GELF calculates an optimal series based likeness between the sites:

1 6 5 9 8 7 10 3 2 4

The program CLUST performs a cluster analysis, which is graphically represented by the program GRAPH2:

Results of the program GRAPH2.



REFERENCES:

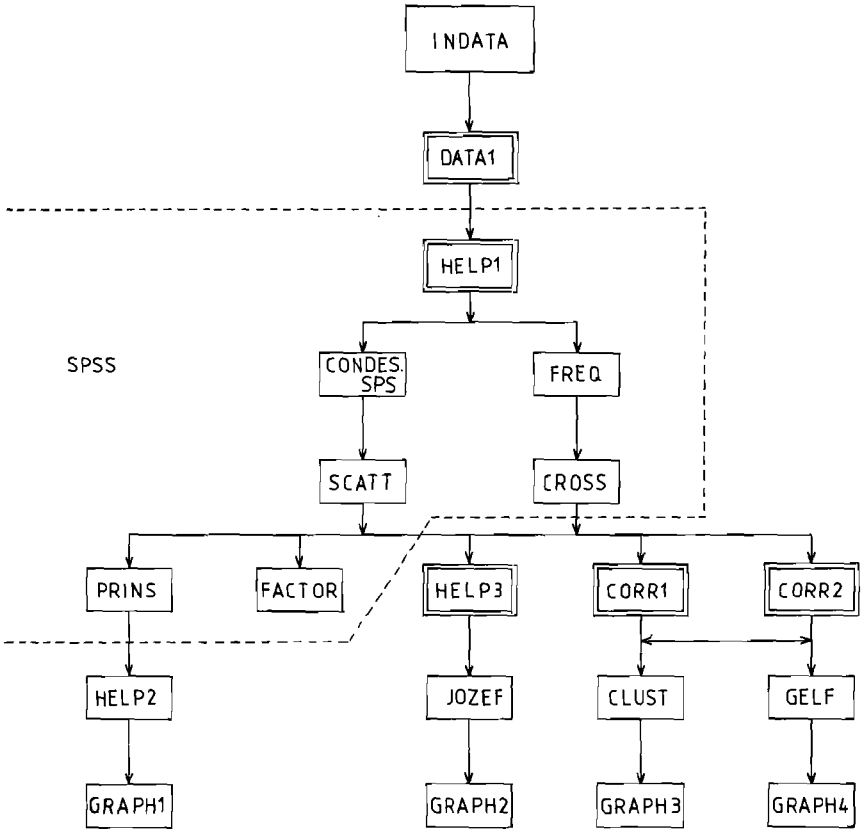
Doran, J.E. & Hodson, F.R. 1975. Mathematics and computers in archaeology. - Edinburgh.

Nie, N.H., Hull, C.H., Jenkins., J.G., Steinbrenner, K. & Bent, D.H. 1975. SPSS. Statistical package for the social sciences (2nd ed.) - New York.

Saers, J. 1978. Birka graves by computer. - Norwegian archaeological review 11:2. Oslo

Sokal, R. & Sneath, P.H.A. 1973. Numerical taxonomy. - San Fransis London

Fig. 1. The recommended order of use of the programs (files) of the program package. The programs in double frame can directly read the indata file. The line of short dashes delimits the programs based on the SPSS package.



Emigrantforskning — Historie på individnivå.

Gunnar Thorvaldsen

Registreringssentral for historiske data (RHD) vil,om den realiseres på permanent basis, gi utvidede forskningsmuligheter på en rekke felter. Ett slikt felt er emigrantforskning ut fra mikrohistoriske metoder. Registrering av den typen det tas sikte på ved RHD, kan sies å være den eneste praktiske mulighet for denne typen forskning, om man ønsker å dekke opp flere og noe omfattende geografiske områder innenfor akseptable forskningsøkonomiske rammer.

Dette notatet søker i all korthet å belyse bruk av individdata i emigrasjonsstudier. Etter en historiografisk innledning summeres kildesituasjonen. I denne sammenheng framheves fordelene ved databehandling av individorientert kildemateriale, slik den er tenkt utført ved Registreringssentral for historiske data.

HISTIORIOGRAFI: FRA MAKRO- TIL MIKROHISTORIE

Inntil 1960-åra var norsk (og internasjonal) utvandringshistorie preget av framstillinger med nasjonen som analyseenhet. Man beskrev de norske settlementene, overfarten og de forlatte samfunn. Fokus lå i forløpet av emigrasjonsbølgene med årsaksforklaringer knyttet an til strukturelle beskrivelser av mottaker- og avsenderland, hvor befolkningsoverskuddet gjerne ses som den fremste bakenforliggende årsaksfaktor.

Hovedkilden til denne type migrasjonshistorie er de offentlige beretninger, brevmateriale, muntlige kilder, aviser, reiseskildringer og statistikk. Stort sett er statistikken bare utnyttet i den aggregerte form den fikk ved den opprinnelige opptelling. Grunnlagsmaterialet, emigrantlister hvor hver emigrant er nevnt med navn, har i noen utstrekning vært benyttet til å publisere fortegnelser over emigranter fra bestemte lokalsamfunn. Ellers har førende emigranter fått stor oppmerksomhet, mens den "menige" emigrant stort sett forblir anonym.

Etter at sosialhistorien i 1960-åra slo igjennom for alvor, har også emigrasjonsforskerne konsentrert seg mer om studier hvor kunnskaper om mange enkeltindivider danner grunnlaget for kollektive biografier. Vi kan ikke nøye oss med kunnskaper om de samfunn emigrantene forlot og kom til. Vår forståelse av emigrasjonsbeslutningene og bakgrunnen for dem blir vesentlig større hvis vi kan rekonstruere utflytternes liv. Kunnskap på makronivå om at de fleste emigranter kom fra landsbygda, leder på mikroplanet naturlig over i spørsmålet om sammensetningen av emigrantflokken. Dominerte f.eks. bønder eller husmenn, kvinner eller menn, unge eller eldre? Hadde mange flyttet før (etappevandring)? Gjennom studier på individnivå kan vår kunnskap bli sikrere, mer detaljert - og mer menneskelig.

KILDESITUASJONEN

Hvor finner man så opplysninger om vanlige mennesker i emigrasjonens halvsekel? Før 1867 bare i kirkebøkenes utflyttingslister som dessverre er notorisk mangelfulle. Mange utflyttere er helt utelatt, og det står lite om de nevnte. Fra -67 førte politiet protokoller i utskipningshavnene. Disse er ganske fullstendige selv om noen emigrerte over utenlandske havner (for å slippe militærtjeneste), mens andre stakk av fra skip i Amerika. Et større problem er disse kildenes angivelse av siste bosted isedetfor fødested. De som f.eks. tjente til billetten ved transittopphold i hovedstaden, blir da vanskelige å spore tilbake til fødestedet. Dessuten løy mange på alderen for å få barnebillett, og yrket er ofte omtrentlig oppgitt. På denne bakgrunn er det farlig å trekke slutninger om emigrantenes sosiale og geografiske opphav ut fra emigrasjonsprotokollene alene.

Disse kildemessige vanskene kan man komme rundt ved å utnytte andre individdata. Opplysningene i emigrasjonslistene er som regel tilstrekkelige til at man kan finne utvandrerne igjen i andre nominative kilder som folketellinger og kirkebøker. Der fins sikrere og mer fullstendige fakta om sosial og geografisk mobilitet, og om den sosiale sammenheng emigranten sto i da han tok utvandringsbeslutningen.

Dette minner oss om at individstudier ikke må lede dithen at vi glemmer helheten; d.v.s. de samfunn, små og store, som emigrantene

forlot. Full forståelse kan vi bare nå gjennom såkalt kontekstuell analyse. Med dette menes å kombinere opplysninger fra forskjellige nivåer: Det individuelle, det nasjonale og nivåene imellom. Tar man utgangspunkt i individdata, kan man i prinsippet aggregere til alle de nivåer kildene spesifiserer. Både gård og grend, både krets og kommune kan være analyseenhet.

REGISTRERINGSSENTRALEN: EDB-VERSJONER AV INDIVIDDATA.

Hvilken nytte kan emigrantforskere i en slik sammenheng ha av individdata fra Registreringssentral for historiske data? Planen for første 3-årsperiode er å gjøre folketellingene 1865-1910 samt kirkebøkene på 1800-tallet maskinleselige for utvalgte norske regioner. Etter foreliggende planer skal Nord- og Midt-Troms, Stjørdalen og Selbu, Sunnhordland, Lillehammerområdet og deler av Østfold dekkes. Disse regionene hadde i 1865 ca. 90.000 innbyggere, og skulle gi grunnlag for langt mer omfattende studier av emigrasjon med et mikrohistorisk utgangspunkt enn hva som hittil har vært mulig. I alt blir det da mulig å kombinere opplysninger fra ca. 1.4. mill. individenheter ved hjelp av EDB. På sikt er det aktuelt å dekke større områder og flere kildetyper; bl.a. emigrasjonsprotokollene. Emigrasjonsforskningen kan imidlertid nyttiggjøre seg data allerede fra starten. Studier på individnivå blir i alle fall utført for relativt begrensede regioner.

Vi vil i tre punkter hevde det tilsynelatende paradoks at slik forskning kan ha større nytte av at folketellinger og kirkebøker blir databehandlet enn om emigrasjonslistene blir maskinleselig. For det ene er de sistnevnte kildene et godt utgangspunkt i sin nåværende form. Til dels foreligger manuelt utarbeidede alfabetiske registre.

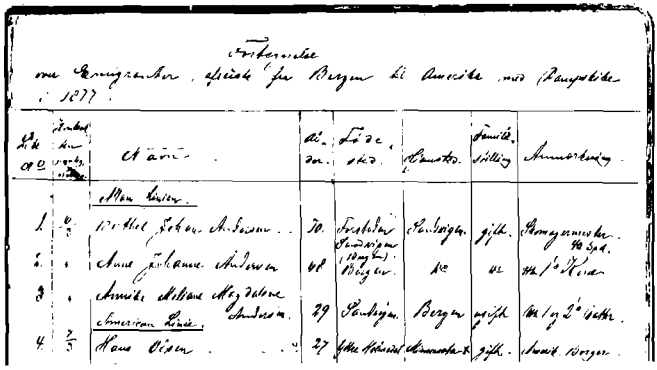
For det andre er det som nevnt i folketellinger og kirkebøker at man finner fyldige opplysninger om emigrantenes bakgrunn. Når man skal kombinere data fra flere nominative kilder med individet som sammenknyttingspunkt, blir arbeidet fort uoverkommelig med tradisjonelle, manuelle metoder. Gjenfinningsarbeidet går langt lettere når man kan finne emigrantene igjen i folketellinger og kirkebøker som er alfabetisert med EDB. Som eksempel kan vi trekke fram studier av

etappeutvandring over Kristiania havn. Man hadde mistanke om at emigranter med siste bosted i hovedstaden i mange tilfeller var innflyttere dit. Dette kunne undersøkes ved å finne dem igjen i forrige folketelling hvor fødested er oppgitt. Nå er 1875-tellinga for Kristianias 75.000 innbyggere databehandlet. Ved søking i de alfabetiserte datalistene blir det mulig å finne nåler i denne høystakken.

For det tredje kan EDB-teknikken lette de statistiske studier av de forlatte samfunn. Datidas trykte statistikk er ofte mangelfull, og som regel satt opp etter andre kriterier enn de som springer ut av dagens problemstillinger. Vi har nå programvare som muliggjør lettvint koding og aggregering med individdata som utgangspunkt. Dermed kan opplysningene standardiseres til tallkoder for de ulike egenskaper. Disse telles opp maskinelt for de nivåer forskeren har bruk for i sine analyser.

Til sist kan vi nevne de spesielle muligheter som foreligger i folketellinga av 1910. Her skulle alle tilbakevandrede norskamerikanere noteres særskilt. Da kan vi få tak i data om hva som hendte noen av utvandrerne "over there".

Men de som kom tilbake var neppe typiske for massen av emigranter, dem vi kan finne i amerikanske arkiver. Det burde være klart hvilke muligheter som foreligger ved å koble sammen norske og amerikanske data om emigrantene. En slik sammenkobling er i praksis utenkelig uten bruk av EDB.



Bildet viser et utsnitt av Emigrantprotokoll nr. 1, 1874-1885, Bergen politikammerarkiv (i Statsarkivet i Bergen).

Tiendpengeskatten 1520/21 i EDB-versjon.

Eirik Lien

BAKGRUNN

Høsten 1975 startet to prosjekt ved Universitetet i Trondheim et sam arbeid med EDB-tjenesten for humanistiske fag. Det ene, "Jordeiendomsforhold og godseiere i Trøndelag. Fra Aslak Bolt til Landkommisjonen", var et arbeid Audun Dybdahl hadde fått NAVF-stipend til. Det andre var "Norsk personnamnleksikon", ledet av ei prosjektgruppe ved Nordisk institutt og da med Terje Aarset som NAVF-ansatt vit.ass

Begge disse prosjektene var i startfasen "på jakt" etter aktuelle data, og med EDB-tjenesten som ekteskapsformidler fant de sin felles kjærlighet i Tiendpengeskatten. Det ektefødte barnet måtte da bli en EDB-versjon av skattelista. I årene framover så stadig nye deler av EDB-versjonen dagens lys etter varierende drektighetsperioder. NAVFs EDB-senter og EDB-tjenesten var hele tida påpasselige og ivrig fødselshjelpere.

Hvorfor kastet vi oss så ut i dette arbeidskrevende og tidkrevende prosjektet? For det første fordi skattelista er en sentral kilde, for det andre er den aktuell for flere fagområder (her: både personnavnforskning og jordeiendomshistorie), for det tredje var den ikke EDB-tilgjengelig fra før og for det fjerde kjente Audun Dybdahl kilden svært godt og ville kunne gi verdifulle bidrag til systematisering og eventuelle tolkningsproblemer. Den er også lett tilgjenge lig ved at den er publisert i "Norske Regnskaber og Jordebøger fra det 16. Aarhundrede".

KILDEN

Skattelista inneholder de enkelte skattebetalernes navn, og innbetalingenes størrelse og betalingsmiddel er angitt. Hver person er det mulig å plassere geografisk fordi det er separate lister for hvert lokale administrative område (sogn, tingsted, skiprede o.l.) og dess uten er gårdsnavn svært ofte også påført. Lista slik den er til-

gjengelig, dekker storparten av Vestlandet fra og med Jæren og nordover, hele det nåværende Trøndelag og Nord-Norge. Øst- og Sørlandet mangler altså - og det er selvsagt et betydelig minus.

I områdene Fosen, Nordmøre, Romsdal, Sunnmøre, Nordfjord, Sunnfjord og Trødelag er mantallet satt opp på forhånd, og de innbetalte skattesummene føyd til senere. D.v.s. at de som av en eller annen grunn ikke betalte også er å finne i lista. I de øvrige områdene som bare dekkes av rene regnskaper, er bare betalende personer kommet med. Navn og innbetalt skatt er ført opp samtidig. Disse listene er derfor ikke så fullstendige sett fra et demografisk synspunkt som listene fra de førstnevnte områdene. I enkelte tilfeller er det også på forhånd påført hvor mye de var taksert til og ved innbetaling hva de betalte dette beløpet i.

En del typiske innførsler i skattelista ser slik ut i den trykte kildeutgaven:

Mandtall i Fozelen
Statzbygndhen

Endrit pa Grastad iij mark iii ß satisfecit
Niels pa Ingedall ij mark satisfecit
Oluff ibidem ij mark satisfecit
Niels ibidem ij lod sølff oc vj lod sølff ffor iorde gotz oc ij lodh sølff ffor barne penningh satisfecit
Pæder pa Brøskyffthe jx lod j quintin søllf xj mark vj ß oc iij mark ffor barne penningh xij ß satisfecit
Jon pa Graningh v lod sølff oc j lod sølff x ß ffor iorde gotz sat.
Ragnild ibidem vj mark sat.
Jon pa Fenstad i mark iiij ß sat.
Arne ibidem ij lod solff sat.
Oluf pa Bwde vj ß sat.

REGISTRERINGSBESKRIVELSE

Vi ønsket i utgangspunktet å lage en EDB-versjon som lå så nært opp til kilden som mulig, slik at vi i registreringsfasen tolket dataene

minst mulig. Den eneste tilleggsinformasjonen vi førte inn, var å lage en systematisk, hierarkisk oppbygd tallkode som entydig identifiserte hver innførsel. Vi valgte å la hver innførsel ("person") være enheten (dataposten) i datasettet og ikke hver enkelt persons verdiinnbetaling. I eksempelet ovenfor er altså Endrit pa Grastads to innbetalinger på 2,5 mark og 3 skilling knyttet til samme identitetskode.

Dette fører til at vi bygde opp hver datapost med fire hovedfelt:

1. identitetskode
2. navnefelt
3. innbetalt skatt
4. eventuell tilvisningskode til samme person et annet sted i datasettet.

De tre første feltene vil alltid være til stede.

1. identitetskoden er bygd opp som et 7-sifret tall:
   siffer 1 - 2 identifiserer lenet
   " 3 - 4 identifiserer de lokale administrative områder i lenet
   " 5 - 7 er en fortløpende nummerering av hver datapost innafor det lokale området

2. navnet registreres slik det står, men med den forskjellen at vi angir gårdsnavnet der kilden ved ibidem viser til foregående person

3. innbetalt skatt registreres slik det står med verditall og verdibetegnelse, men det er innført en to-bokstav kode for betegnelse i alt 54 forskjellige betegnelser.

4. eventuell tilvisningskode er samme persons identifikasjonskode(r andre steder i datasettet.

Vi mener selv at vi har oppnådd en rimelig grad av fleksibilitet ved denne fremgangsmåten, og sluppet å gjøre vold mot dataene. I alt er

10477 dataposter registrert.

De 10 innførslene som er vist ovenfor ser slik ut i registrert form:

```
0101001 ENDRIT PA GRASTAD/2.5MA/3SK/
0101002 NIELS PA INGEDALL/2MA/
0101003 OLUFF PA INGEDALL/2MA/
0101004 NIELS PA INGEDALL/2LS/JG6LS/BP2LS/
0101005 PED<ER> PA BRØSKYFFTHE/8.5LS/0.5QS/10.5MA/6SK/BP3MA/BP12SK/
0101006 JO<N> PA GRA<N>INGH/5LS/JG0.5LS/JG10SK/
0101007 RAGNILD PA GRA<N>INGH/5.5MA/
0101008 JO<N> PA FENSTAD/1MA/4SK/
0101009 ARNE PA FENSTAD/2LS/
0101010 OLUFF PA BWDE/6SK/
```

PROBLEMER

Det har vært to typer problemer forbundet med overføringa av kilden til EDB-form; det ene å velge mellom to eller flere mulige tolkninger i kilden, det andre å finne igjen de personene i restanselister/tilleggslister som er nevnt i tidligere lister. Det første problemet er særlig knyttet til å føre innbetalingene til riktig person i enkelte lister. Noen steder er innbetalingene ført over flere linjer og navnet på skatteyteren er nevnt på bare ei av disse linjene. Vi mener vi har funnet den riktige løsningen i de fleste tilfellene.

Å identifisere enkeltpersonene i tilleggslistene viste seg mer å være en tålmodighetsprøve enn et direkte problem. Bare i de tilfellene hvor navnet er svært alminnelig, av typen Oluff Pedersen, kan det være vanskelig å avgjøre om det er samme person i tilleggslista som i hovedlista. Dette dreier seg i høyden om 12 - 15 tilfeller.

PROGRAMUTVIKLING

I tilknytning til de to prosjektene er det laget en del spesialprogram, for personnavnprosjektet fins det program for å hente fram personnavna, alfabetisere og telle dem opp og rangere dem.

For jordeiendomsprosjektet er det program for opptelling og summering

av innbetalt skatt på ulike geografiske nivå, både totalt innbetalt skatt og skatt betalt for jordegods. Det er også utarbeidet ei omregningsliste som konverterer alle innbetalinger til felles verdienhet, nemlig mark.

TILGJENGELIGHET

Kopi av datasettet kan en få ved å henvende seg til EDB-tjenesten for humanistiske fag ved Universitetet i Trondheim. Det er også utarbeidet en kodebeskrivelse som gir de nødvendige opplysningene om oppbygging av datapostene og systematisering av dem.

Dataene er i prinsippet fritt tilgjengelig, men det kan i enkelte tilfeller bli satt sperre mot spesiell bruk i et visst tidsrom. Dette kan eventuelt avtales når de(n) interesserte mottar kopien.

Ludvig m, tysk namn, sms. av ghty. hlūt 'vidgjeten, berømt' og wīg 'kamp, strid; ei kjempe' (svarar til nord → Vig-). I bruk sidan 1300-t., skrive Lodvik o.l. i mellomalderen. Populært namn kring 1900. Noko brukt i mellomkrigstida, sjeldnare etter 1945. Helgennamn og fr. kongenamn, m.a. Ludvig (fr. Louis) 14. ("solkongen", 1638-1715). Ludvig Holberg, kjend da. diktar (1684-1754). I skrift både Ludvig og Ludvik i no. Den fr. forma Louis var ein del nytta først i vårt hå. Namnet Lodve kjem av norr Hlǫðvér, som er eit eldre lån av namnet L.

Litt.: Janzén i NK VII s. 134 m. tilv. Lindquist 1924.

Magne m, nord. namn, norr. Magni, avl. av appellativet magn n 'makt, styrke'. Ein av sønene til guden Tor ber namnet M. Det er kjent som døypenamn frå ca. 1100 og har vore allment sidan 1500-t. På 1700-t. var det vanleg i Hord og fanst og nokre gonger i SogFj. M. fekk eit oppsving med den nord. nemnerenessansen (innl. s. 00) og var eit populært namn i mellomkrigstida og fram til 1950-åra, først i denne perioden særleg i Hedm, Oppl og Rog. Vanleg etter 1970.

Smakebiter fra manus til Norsk personnamnleksikon (red. Reidar Djupedal og Ola Stemshaug), en navnebok som er under arbeid ved Nordisk institutt i Trondheim. Den vil trolig være i handelen til jul 1981. Navnelistene for Tiendpengeskatten 1520 har vært en verdifull kilde for leksikonet.

Some thoughts on the use of computers in linguistic research.

Stig Johansson

THE LINGUIST AND THE COMPUTER

In the introduction to a book on automatic language processing published about 15 years ago (Hays 1966) it is said that "Perhaps by 1975 computational linguistics will be known to everyman....." At the moment of writing, 1980, computational linguistics is not even known to all practising linguists, and it is used only by a minority of them in their research. This could partly be due to lack of training opportunities. More probably, there is a deeper explanation. Linguists in general have not become convinced of the usefulness of computers in their research and have therefore not bothered to learn.

The rather limited interest in using computers is probably due mainly to a misunderstanding. "Computational linguistics" is regarded as a special type of linguistics concerned with pedestrian tasks such as the making of word lists, concordances and the like. Something would be gained if we could abolish this term and think instead of computer-aided linguistics.[1] "Computer-aided" linguists are concerned with the same sorts of questions as all other linguists, only by using a computer they think they can do a better and more efficient job.

In this paper I shall concentrate on the uses of machine-readable corpora of modern English texts, as an example of the possibilities-- and limitations-- of computer-aided linguistic research.

CORPUS, INTROSPECTION AND EXPERIMENT IN LINGUISTIC RESEARCH

Part of the explanation why many linguists have not bothered to use computers is the view, widely held during the last two decades, that the proper data for linguistic research are not texts and observations of language use but rather native speakers' intuitions, especially the introspective judgements of the investigator. Actual language use has been regarded as contaminated by various aspects of

"performance" (Chomsky 1965). Moreover there has been a focus on highly specific problems, where observations of language would often provide insufficient evidence. The emphasis has not been on describing language use but on the testing of particular aspects of linguistic theory.

The climate of research is now changing (and has been for some time) More and more linguists are becoming interested in a systematic study of "performance". The validity of introspection as the principal source of data has been challenged. Complementary methods of data collection are being increasingly used, in particular text corpora (e.g. Francis 1964, Bergenholtz and Schaeder 1979) and elicitation experiments (e.g. Quirk and Svartvik 1966, Quirk and Greenbaum 1970, Greenbaum 1977).

Figure 1 outlines the relationship between corpus, introspection and experiment in linguistic research. A careful linguist does not approach a corpus or conduct elicitation experiments without some basis in introspection, nor does he use introspection alone. A corpus is used to test introspective hypotheses (e.g. as regards the factors influencing the choice of the s-genitive vs. the of-construction in English, the type of complementation with a certain group of verbs, the uses of a particular word order, etc.). Just as important, corpus studies generate new hypotheses, to be tested against a larger material or in elicitation experiments. Similarly, elicitation experiments, though normally arranged to test specific hypotheses, may yield results which stimulate new hypotheses to be tested against a corpus in further elicitation experiments. Good linguistic research arises from a fruitful combination of previous work, introspection, corpus and experiment, or, more generally, from tradition, imagination and observation.

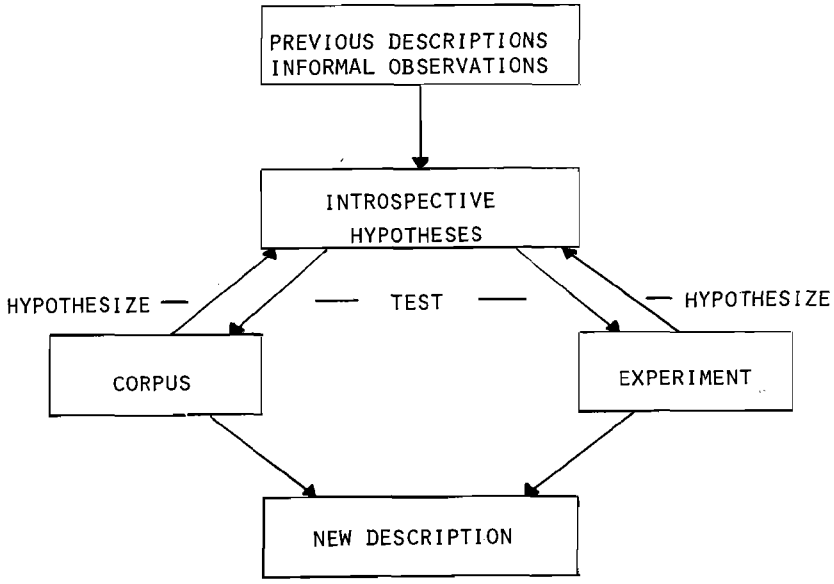
COMPUTER CORPORA

Once it is recognized that introspection alone is insufficient, the computer emerges as a possible important tool which can be used to organize the data for linguistic research, whether they are to be found in already existing texts (spoken or written) or orginate from

elicitation experiments. The attention is here, however, limited to machine-readable corpora.

A machine-readable corpus contains a number of running texts available for computer processing. To qualify as a corpus in the strict sense, it should be a body of texts put together in a principled way rather than a more or less random collection of material. Typical examples are the Brown Corpus of American English (Francis 1964, Kučera and Francis 1967, Francis 1979), the Lancaster-Oslo/Bergen Corpus of British English (Johansson et al. 1978), the London-Lund Corpus of spoken British English (Quirk and Svartvik 1979), and the Leuven Drama Corpus (Geens 1978). All these corpora have been, or are being, extensively used in research on present-day English; see especially the Brown Corpus bibliography in ICAME NEWS 2 (1979).

Figure 1. The place of a corpus in linguistic investigation[2]



The special advantage of a computer corpus is that, once it has been compiled, it "is a reservoir of linguistic usage in a form... that makes it relatively easy to extract exhaustively all available specimens of a given word or describable grammatical item" (Kučera and Francis 1967: v). If one is, for example, interested in studying the use of shall vs. will in British and American English, the s-genitive vs. the of-construction, some vs. any, etc. the relevant examples can be retrieved immediately.[3] A linguist working without a computer would have to spend months just collecting the material.

A further advantage of the computer corpus is that it can be made available to other researchers. This means that exactly the same material can be examined from many different points of view, whereas the data otherwise normally varies depending upon the investigator and the topic of investigation. Ideally, the corpus should be subjected to total accountability (cf. Quirk and Svartvik 1979:204), i.e. analysis of all relevant aspects, a goal which can hardly be reached without the aid of a computer. The possibilities of examining an indefinite number of aspects of machine-readable corpora will undoubtedly lead to many new discoveries. I am here thinking of the hypothesis-generating use of text corpora taken up before. Linguists who use a computer and attempt to take account of all the material provided will be forced to see what they might otherwise overlook. In conclusion, by putting a vast amount of material readily at their disposal, the computer will perhaps make linguists adopt a more realistic approach to data than was frequently found during the last two decades.

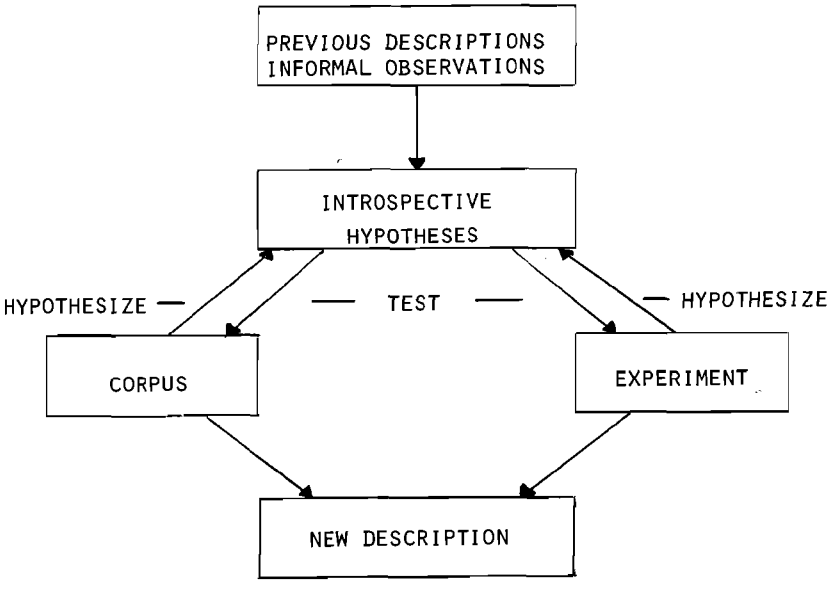
INTERNATIONAL CO-OPERATION

The question of general availability of material was brought up in passing above. A number of corpora of modern English texts are currently available in various countries. There is at the moment close co-operation between universities in England, the United States, Holland, Belgium and the Scandinavian countries. An International Computer Archive of Modern English (ICAME) has been set up at the Norwegian Computing Centre for the Humanities at Bergen. The object of ICAME, which publishes the newsletter ICAME NEWS, is to further

elicitation experiments. The attention is here, however, limited to machine-readable corpora.

A machine-readable corpus contains a number of running texts available for computer processing. To qualify as a corpus in the strict sense, it should be a body of texts put together in a principled way rather than a more or less random collection of material. Typical examples are the Brown Corpus of American English (Francis 1964, Kučera and Francis 1967, Francis 1979), the Lancaster-Oslo/Bergen Corpus of British English (Johansson et al. 1978), the London-Lund Corpus of spoken British English (Quirk and Svartvik 1979), and the Leuven Drama Corpus (Geens 1978). All these corpora have been, or are being, extensively used in research on present-day English; see especially the Brown Corpus bibliography in ICAME NEWS 2 (1979).

Figure 1. The place of a corpus in linguistic investigation[2]



The special advantage of a computer corpus is that, once it has beer compiled, it "is a reservoir of linguistic usage in a form... that makes it relatively easy to extract exhaustively all available speci mens of a given word or describable grammatical item" (Kučera and Francis 1967: v). If one is, for example, interested in studying the use of shall vs. will in British and American English, the s-genitive vs. the of-construction, some vs. any, etc. the relevant examples can be retrieved immediately.[3] A linguist working without a computer would have to spend months just collecting the material.

A further advantage of the computer corpus is that it can be made available to other researchers. This means that exactly the same material can be examined from many different points of view, whereas the data otherwise normally varies depending upon the investigator and the topic of investigation. Ideally, the corpus should be subjected to total accountability (cf. Quirk and Svartvik 1979:204), i.e. analysis of all relevant aspects, a goal which can hardly be reached without the aid of a computer. The possibilities of examini an indefinite number of aspects of machine-readable corpora will undoubtedly lead to many new discoveries. I am here thinking of the hypothesis-generating use of text corpora taken up before. Linguists who use a computer and attempt to take account of all the material provided will be forced to see what they might otherwise overlook. In conclusion, by putting a vast amount of material readi at their disposal, the computer will perhaps make linguists adopt a more realistic approach to data than was frequently found during the last two decades.

INTERNATIONAL CO-OPERATION

The question of general availability of material was brought up in passing above. A number of corpora of modern English texts are currently available in various countries. There is at the moment close co-operation between universities in England, the United State Holland, Belgium and the Scandinavian countries. An International Computer Archive of Modern English (ICAME) has been set up at the Norwegian Computing Centre for the Humanities at Bergen. The object of ICAME, which publishes the newsletter ICAME NEWS, is to further

international co-operation in the field. If successful, such co-operation provides greater economy and better utilization of existing resources as well as opportunities for mutual exchange of ideas and research results.

COMPUTER CORPORA: PROBLEMS AND PROSPECTS

International co-operation opens up almost unlimited possibilities. The problem for the corpus worker will certainly not be a dearth of material. This is especially so, as machine-readable texts can frequently nowadays be obtained from printing houses.[4] The linguist can therefore start his studies without being delayed by the process of preparing texts in machine-readable form.

A special problem in connection with the dissemination of machine-readable texts is how to avoid infringement of copyright. ICAME has strict rules for the distribution of texts; each recipient must sign a written agreement not to reproduce or redistribute material, to use it for research purposes only, etc. With a strict code of ethics of this kind, it should be possible to reach agreement with publishers and printing houses as regards the use of copyrighted material for the purposes of linguistic research.[5]

The wealth of data will perhaps soon pose some problems of its own. Linguists must guard themselves against becoming data compilers rather than analysers of data. They must also make sure that they assemble material in a principled way rather than amass readily available texts. This brings us to the additional difficulty that many types of texts will probably never be made available to the linguist in machine-readable form, e.g. all kinds of spoken and unprinted written material, which must then still go through the laborious process of punching and several stages of proof-reading.[6]

So far, we have only dealt with the use of unanalyzed texts. Many types of investigation, however, require analysed texts, such as the grammatically tagged Brown Corpus or the prosodically coded London-Lund Corpus (see ICAME NEWS 3, 1979). At present, there is only limited access to such material. The efforts of corpus compilers

should therefore concentrate on the preparation of research facilities of this kind. For such work to be possible on a large scale, there is a need to develop techniques of automatic language processing.

This is not the place to survey the different systems of automatic language processing that have been developed so far. Suffice it to say that automatic language processing puts the highest demands both on the linguist and the computer specialist. It seems unreasonable that one individual should take on both roles, which can easily lead to amateurishness in both fields. There is a need for co-operation between linguists with a good insight into the problems of linguistics and for computer specialists with expert knowledge of computational techniques, both of whom must know enough of each other's field to ensure successful communication.

In conclusion, it should be pointed out that it is not only the product--a corpus of analysed text--which is important in such co-operative projects. The process of automatic analysis itself will help reveal fundamental characteristics of language.

THE LINGUIST AND THE COMPUTER: CONCLUDING REMARKS

This paper has mainly considered the use of machine-readable corpora in linguistic research. Needless to say, there are many other areas where the computer can be an important tool, too numerous to be taken up here; see e.g. Allén and Thavenius (1970), Svartvik (1970) Allén (1972, 1975), Brodda (1977).

How can then linguists in general be made aware of the possibilities of the computer? A fundamental requirement is that the "computer-aided" linguist masters his own field and does not become the slave of the computer. If he can show that, by using a computer, he can contribute to a better solution of central linguistic problems and can present his results in an understandable way, there is no doubt that more and more practising linguists will follow suit, though the time may still be far away when "computational linguistics" is known to everyman.

NOTES

1. Svartvik (1970) distinguishes between "computer-aided linguistics, where the machine is called in for temporary and partial assistance" and "computer-based linguistics, which will be reserved for those relationships more equal than that between master and slave".

2. Though presented in a parallel way in Figure 1, corpus and elicitation differ in that corpus studies belong more typically to a hypothesis-generating stage, elicitation experiments more to a hypothesis-testing stage of the investigation.

3. Examples of such work are two "hovedfag" theses currently in progress at the University of Oslo, on shall and will in British and American English (Inger Krogvig) and the s-genitive vs. the of- construction (Mette-Cathrine Sørheim).

4. The Swedish "logoteque", compiled at the University of Gothenburg under the direction of Sture Allén, contains almost exclusively material of this kind. Similar English material is being collected by researchers in England and the US.

5. Allén has agreements with Swedish publishers which permit him to include copyright material in the Gothenburg "logoteque" (cf. note 4). It has proved more difficult to obtain such agreements with English publishers.

6. The possibilities of using OCR will, however, considerably simplify the collection of corpora of printed as well as unprinted written material. An extensive collection of many types of material, printed and unprinted, is the Gill Corpus (to be described in a forthcoming issue of ICAME NEWS), which claims to be "representative of text produced in braille". (Dr. J.M.Gill, Warwick Research Unit for the Blind, personal communication).

REFERENCES

Allén, S. 1972. "Ordstrømmen som datafløde". Humanistisk forskning 2. 19-22.

Allén, S. 1975. Språklig databehandling och särspråklig forskning. Department of Computational Linguistics, University of Gothenburg.

Allén, S. and Thavenius, J. eds. 1970. Språklig databehandling: datamaskinen i språk- och litteraturforskning. Lund: Student-litteratur.

Bergenholtz, H. and Schaeder, B., eds. 1979. Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora. Königstein/Ts.: Scriptor.

Brodda, B. 1977. "Något om datalingvistiken och framtiden". COMPILING, April 1977.

Chomsky, N. 1965. Aspects of the Theory of Syntax. Cambridge, Mass.: M.I.T. Press.

Francis, W.N. 1964. Manual of Information to Accompany a Standard Sample of Present-Day Edited American English, for Use with Digital Computers. Department of Linguistics, Brown University

Francis, W.N. 1979. "Problems of Assembling and Computerizing Large Corpora". In Bergenholtz and Schaeder (1979), 110-123.

Geens, D. 1978. "On Measurement of Lexical Differences by Means of Frequency". In G. Altmann, ed., Glottometrika 1. Bochum: Studienverlag Dr. N. Brockmeyer. 46-72.

Greenbaum, S. ed. 1977. Acceptability in Language. The Hague: Mouton.

Greenbaum, S. and Quirk, R. 1970. Elicitation Experiments in English: Linguistic Studies in Use and Attitude. London: Longm

Hays, D.G, ed. 1966. Readings in Automatic Language Processing. New York: American Elsevier.

ICAME NEWS. Newsletter of the International Computer Archive of Modern English. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S., Leech, G.N. and Goodluck, H. 1978. Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers. Department of English University of Oslo.

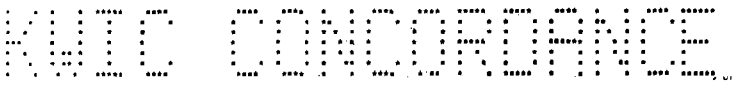
Kučera, H. and Francis, W.N. 1967. Computational Analysis of Present

Day American English. Providence, R. I.: Brown University Press.

Quirk, R. and Svartvik, J. 1966. Investigating Linguistic Acceptability. The Hague: Mouton.

Quirk, R. and Svartvik, J. 1979. "A Corpus of Modern English". In Bergenholtz and Schaeder (1979). 204-218.

Svartvik, J. 1970. "Computational Linguistics Comes of Age". Times Literary Supplement July 23, 1970.




```
W21 150   9      Of these **170 million worth, or about 60 per \0 cent., were temperate foodstuffs: **130 million worth, out of \0
W15  75   3      accurate. If a rate of interest of 3 3/4 per \0 cent:-** which is about equivalent to the Public Works Loan  \0
G07  20   1      the saying of \0R. Joshua \0b. Levi (early 3rd \0 Cent.) who said that three things were done by a human court \0
          WORDS=392, A=72, B=34, D=5, E=34, F=9, G=30,                  H=94, J=95, K=4, L=3, M=3, SAMPLES=101, A=16,
          B=10, D=4, E=11, F=4, G=16, H=14, J=23, K=2,                  L=1, M=2, CAT=11
BROWN     WORDS=161, A=32, B=19, C=9, D=6, E=22, F=27,                  G=1, J=30, K=2, L=1, M=3, SAMPLES=55, A=11, B=7,
          C=3, D=2, E=2, F=9, G=1, J=17, K=1, L=1, M=1,                 CAT=11
C15 215   3 Silenus (his arm thrown blissfully over a barrel), a centaur, executed in unglazed red earthenware, turning to

G08  87  10 far distance.' Years later, on 14 July 1911, the Pop Centenary Dinner was held at Eton. Curzon went, but Ritchie
H25  50   4      experiments may lead by the time we reach our Centenary in 19667 **Changes in Australia 1967** The wind
A13 255   6      A COLOURFUL exhibition commemorating the centenary of a remarkable event opens today at the Victoria
          WORDS=3, A=1, G=1, H=1, SAMPLES=3, A=1, G=1,                  H=1, CAT=3
BROWN     003

                                                                  Centennial
BROWN     WORDS=6, C=1, E=1, G=1, H=2, K=1, SAMPLES=6,                  C=1, E=1, G=1, H=2, K=1, CAT=5

                                                                  center
BROWN     WORDS=229, A=40, B=10, C=9, D=12, E=31, F=18,                 G=26, H=7, J=41, K=8, L=4, M=5, N=10, P=7, R=1,
          SAMPLES=116, A=17, B=7, C=6, D=5, E=14, F=10,                 G=15, H=5, J=14, K=5, L=4, M=1, N=8, P=4, R=1,
          CAT=15

                                                                  center's

BROWN     WORDS=2, A=1, J=1, SAMPLES=2, A=1, J=1, CAT=2
J65 134   8 <<=2*5>   THE RITUAL art of the synagogue naturally centered on the Scroll of the Pentateuch or (Segher Torah),
BROWN     WORDS=16, A=1, F=2, G=4, H=1, J=6, L=1, R=1,                  SAMPLES=13, A=1, F=2, G=4, H=1, J=3, L=1, R=1,
          CAT=7

                                                                  centerfield

BROWN     A11

                                                                  Centering
BROWN     WORDS=5, E=1, F=1, G=1, J=2, SAMPLES=4, E=1,                  F=1, G=1, J=1, CAT=4

                                                                  centerline
BROWN     E16

                                                                  centers
BROWN     WORDS=52, A=6, C=1, D=2, E=3, F=6, G=9, H=3,                  J=20, M=1, R=1, SAMPLES=32, A=4, C=1, D=2, E=3,
          F=4, G=4, H=3, J=9, M=1, R=1, CAT=10
B03 171   7    for the sake of $0,000 is there any reason why the centigrade countries should not change to fahrenheit? Can it
J08 198   3      to an external meter calibrated to read degrees Centigrade, covering a temperature range from 0 to 500 in
J19 102  10    between the two groups of three objects. The mean Centigrade temperature of a set of objects will be
B03 157   1    allowing these Bills a fair run. **By Degrees** Centigrade \0v. Fahrenheit. The fight is on. The challenger
B05 184   3  is to give fahrenheit the knock out. The backers of centigrade would have got off to a better start if they had
          WORDS=5, B=1, J=2, SAMPLES=3, B=1, J=2, CAT=2
BROWN     J72
B03 174   2  scale is demonstrably better than the other?  The centigraders may be in for a stiffer fight than they think.

                                                                  centimeter

BROWN     WORDS=3, J=3, SAMPLES=1, J=1, CAT=1

                                                                  centimeters
BROWN     WORDS=8, E=8, SAMPLES=1, E=1, CAT=1
J07  96   3  15 or 18 fins per inch (3.9 to 5.9 or 7.1 fins per centimetre) depending upon the type of corrugation.

J15 206   1 to the fact that caesium sources are usually several centimetres long it is necessary to define the term

J68  54   7  mercy on us', a vernacular Kyrie) are more or less centos of pleinsong motives. Moreover the earliest preserved

A05  17   1  This would apply also in the command structure and central administrative organisation.  \0Mr. Watkinson,
B19   3   7      years at school?>  \0RD Amory is to head the Central Advisory Council for Education during its
```

Knut Hofland
Stig Johansson

LOB CORPUS
KWIC CONCORDANCE

The Norwegian Computing Centre for the Humanities (NAVF)
P. O. Box 53
N-5014 Bergen-Universitet
NORWAY

Norsk termbank.

Håvard Hjulstad

Norsk termbank (NOTE) er no i ferd med å bli skipa som eigen institusjon med Norsk språkråd, Rådet for teknisk terminologi, Universitete i Bergen og Universitetsforlaget som stiftarar. Til no har termbank eksistert nokre år som eit prosjekt ved Nordisk institutt, Prosjekt for datamaskinell språkbehandling (PDS) ved Universitetet i Bergen. Det er meininga at termbanken skal dele husrom med PDS i framtida òg.

KVA ER EIN TERMBANK?

Ei lang rekkje organ og føretak arbeider med terminologiske problem. Dei ulike fagspråka får stadig bruk for nye nemningar på nye ting og prosessar. Og i mange fagområde kjenner dei etter kvart eit sterkare behov for å samordne terminologien.

Alt dette arbeidet fører til store mengder data som tidligare stort sett berre har vore tilgjengelege i manuell form på lister og i sete arkiv. Når eit manuelt arkiv veks, blir oversynet fort borte. Berr datamaskinell behandling kan gi ein oversynet attende.

I ein termbank er eit svært viktig siktemål å gi eit slikt oversyn. Ein termbank er ei maskinleseleg "ordbok", men også meir enn det: På eit standardisert format får ein rom til terminologisamlingar som kan sjå heilt ulike ut i eventuelle trykte utgåver. Men inne i maskina er dei (på lag) like: Ein kan sortere dei saman; ein kan dra ut delar av dei etter gjevne kriterium.

Dei tre "løyndommane" med ein termbank er da:

1. Han kan reint organisatorisk verke samordnande på terminologiarbeidet. Han kan stå som ein samlingsstad for slikt arbeid, utan at alt terminologiarbeidet treng vere geografisk knytt til termbanken.

2. Ein nyttar datamaskin og tek i bruk den store søkje- og sortering

kapasiteten hennar. Med eit databasesystem (dokumentsøkjesystem) nyttar ein direkte tilgang og får med det ei svært låg tilgangstid.

3. Termbanken har eit internt format på data som gjer det lett å samordne dei ymse prosjekta, samstundes som ein står fritt til å velje språk, kor omfattande definisjonar ein skal ha, kor mykje ein skal ta med av klassifikasjon osb.

NORSK TERMBANK I DAG.

Sidan mars 1979 har det vore éi fulltidsstilling ved Norsk termbank. Dessutan har vi hatt noko timelønt hjelp. Før den tid vart det dataregistrert ein del terminologi i PDS-regi.

I det siste året har vi utvikla det interne formatet og konvertert data til dette formatet. Vi har òg utvikla eit programsystem for formatet og eit kodesystem for alle spesialteikna ein har bruk for. Elles har vi vidareført gamle prosjekt og sett i gang nye. Vi har sett litt på databasesystem for NOTE, men ikkje noko av data ligg enno føre for søking med direkte tilgang.

Format.

Innanfor kvar termpost tek ein med alle opplysningane om eit omgrep. Kvar opplysning står i eit felt for seg med ein feltkode framfor. Feltkoden seier mest mogleg eintydig kva for opplysning det er, stundom òg korleis denne opplysninga knyter seg til andre felt innan-

Feltkoden gjer det mogeleg å knyte t.d. grammatiske opplysningar og sorteringsform til dei ulike termfelta. Det går òg an å knyte saman ulike termpostar med opplysningar om hierarki ("overterm", "underterm", "sideterm", "motsett term"). Sjølve feltkodane er heller kompliserte, men den vanlege brukaren kan få dei ut i "oppløyst" form.

Program.

Det er skrive eit femtital program til å behandle dei sekvensielle NOTE-filane. Ein kan dele dei i grupper:

1. Somme felt kan lagast meir eller mindre automatisk. Det er fleire program for feltgenerering, somme av dei med dialog og høve til manuelle inngrep.

2. Nokre program hjelper til med å finne feil i data og kodesetjing.

3. Data kan sorterast på kva felt som helst (på termar i eit gitt språk, på klassifikasjonskode osb.). Ein kan dra ut referanselister med tilvising frå eitt felt til eit anna (t.d. frå framandspråklege termar til norske).

4. Ein kan plukke ut termpostar etter gjevne kriterium, eventuelt med OG- og ELLER-operand.

5. Vidare kan ein dra ut framlengs og baklengs alfabetiserte lister over termar, og lage konkordansar over innhaldet i gjevne felt. Det siste er nyttig til å kontrollere termbruken i definisjonane.

6. Det er program som lagar utskrifter i "leseleg" form, og program som klargjer data for fotosetjing.

Når data skal leggjast opp for direkte tilgang, må det utarbeidast nokre program for det. Stort sett vil ein likevel nytte standardsystem, i alle fall i første omgang databasesystemet UNIDAS 1100.

Data.

Når det gjeld fagspreiing, er ikkje innhaldet i termbanken så imponerande enno. Vi har i maskinleseleg form dei prosjekta som Rådet for teknisk terminologi er i gang med (om lag 10 prosjekt) og nokre få av dei eldre RTT-ordbøkene. Dessutan er Norsk dataordbok lagra ved termbanken. Marknadsføringsterminologien, som Norsk språkråd arbeider med, er enno ikkje konvertert til NOTE-formatet, men han er maskinleseleg. Som eit samarbeidsprosjekt med dei norske standardiseringsorgana er vi no i gang med å dataregistrere terminologien i alle gjeldande Norsk Standard.

Datamengda er likevel stort alt om ein ser på linjetal eller blokktal i maskina. Til saman har vi om lag 160 000 linjer tekst inne;

ein stad mellom 15 000 og 20 000 termpostar.

NORSK TERMBANK I MORGON.

Etter alt å døme vil arbeidsmengda og datamengda ved termbanken auke stadig raskare i tida framover. Det er mange prosjekt som ventar på å bli tekne opp. Termbanken vil truleg måtte vekse seg større både i talet på tilsette og i datamaskinkapasitet.

Utpå hausten 1980 vonar vi å ha ein prøvedatabase i drift. Da vil det vere mogleg for registrerte brukarar å "slå opp" i data frå terminal. Etter kvart som datanetta blir utbygde, blir denne bruken mest uavhengig av kor i landet (eller verda) ein er. Også ein del av arbeidet på termbanksida kan ein da tenkje seg blir desentralisert.

Dei store datamengdene som må vere tilgjengelege for søking til kvar tid, gjer at ein må ha tilgang til stor masselagerkapasitet. Den kapasiteten som no er tilgjengeleg for vanlege brukarar ved datamaskina ved Universitetet i Bergen, vil ikkje strekkje til i framtida.

På neste side er det gitt eit par døme på korleis termpostane er sett opp.

DØME PÅ DATA

```
NB = bokmål
NN = nynorsk
DE = tysk
EN = engelsk
FR = fransk
TT = "prosjekt-
     opplysning"

001-009 = hovudterm
011-019 = synonym
301-309 = definisjon

..a = grammatisk
      opplysning
..c = sorteringsform
..i = "status"
```

```
NB001   bukserør
NB001a  s n
NB001i  7
NN001   bukserør
DE001   Hosenrohr <(n)>
DE001c  hosenrohr
EN001   Y-pipe
EN001c  ypipe
EN011   bifurcation
FR001   tuyau <(m)> culotte
FR001c  tuyau culotte
FR011   tube <(m)> en Y
FR011c  tube en y
NB301   Y-formet rørforgreningsdel.
NB301i  7
NB501   Se grenrør.
TT811   04
TT901   aaa5 / RTT - Vannturbiner
TT911   800128
TT921   Tei
=
NB001   delesirkeldiameter
NN001   delesirkeldiameter
DE001   Teilkreisdurchmesser <(m)>
DE001c  teilkreisdurchmesser
EN001   pitch diameter
FR001   diam$2etre <(m)> primitif
FR001c  diametre primitif
NB301   Den diameter en *deling refererer seg til.
TT811   04
TT901   aaa5 / RTT - Vannturbiner
=
```

DØME PÅ "BRUKARFORM"

```
**************************************************************  LNR: 00001
                                              PROSJ: aaa5         DEL: 04
NB-TERM:  7   bukserør
NN-TERM:      bukserør

DE-TERM:      Hosenrohr <(n)>
EN-TERM:      Y-pipe
EN-SYNO:      bifurcation
FR-TERM:      tuyau <(m)> culotte
FR-SYNO:      tube <(m)> en Y

NB-DEFI:  7   Y-formet rørforgreningsdel.
NB-MRKN:      Se grenrør.


**************************************************************  LNR: 00002
                                              PROSJ: aaa5         DEL: 04
NB-TERM:      delesirkeldiameter
NN-TERM:      delesirkeldiameter

DE-TERM:      Teilkreisdurchmesser <(m)>
EN-TERM:      pitch diameter
FR-TERM:      diam$2etre <(m)> primitif

NB-DEFI:      Den diameter en *deling refererer seg til.
```

Oppstarting av Norsk tekstarkiv.

Per Vestbøstad

Norsk tekstarkiv starta opp 1. mars i år med underskrivne som prosjektmedarbeidar i heil stilling. Prosjektet skal samordna og styrkja arbeidet med å samla og tilretteleggja moderne norsk tekstmateriale for datamaskinell utnytting med tanke på forsknings- og utviklingsarbeid. Rådet for humanistisk forskning, NAVF, har gjeve lovnad om 5 års finansiering av ein konsulentstilling.

Den faglege leiinga av prosjektet er lagt til eit fagleg råd, som har representantar frå fagmiljøa ved dei fire universiteta, NAVFs EDB-senter og Prosjekt for datamaskinell språkbehandling (PDS) i Bergen. Professor Egil Pettersen, Nordisk institutt, Universitetet i Bergen, er formann.

Pr. 1.7.80 kan tekstarkivet rapportera følgjande:
Bjørn Eide ved PDS har laga ei utgreiing om standard lagringsformat for tekstar. Fagleg råd fann at dei ville venta med å velja standardformat til ein hadde vunne meir røynsler med ulike lagringsformat.

Ved PDS er det utvikla eit programsystem for lesing av trykkeritekst frå magnetband. Tekstarkivet skal ta i bruk dette systemet, og underskrivne samarbeider for tida med Bjørn Eide for å gjera seg kjent med bruksmåten for dei ulike programma.

Tekstarkivet har samla opplysningar om norsk språkmateriale som finst i datamaskinleseleg form. Eit oversyn vil verta sendt til alle aktuelle institusjonar i haustsemesteret. Dei einskilde personar som vil ha tilsendt dette og framtidige oversyn, må melda frå til Norsk tekstarkiv, postboks 53, 5014 Bergen-Universitet.

For 1980 har fagleg råd vedteke å prioritera innsamling av romantekstar. Ein ventar førebels på klarsignal frå Norsk forleggerforening.

Det er laga ei utgreiing for Norsk språkråd om ulike alternativ

for å tilretteleggja tekstar til ei jamføring av rettskrivinga hjå tre forfattargenerasjonar. Dersom språkrådet gjev løyvingar, vil Tekstarkivet stå for teksttilrettelegging, ordlisteproduksjon og vidare databehandling av materialet.

Dag Worren (Norsk leksikografisk inst.) og Kolbjørn Heggstad (PDS) har gjort ei rundspørjing til nynorskavisene for å finne ut kven som har datamaskinleseleg avistekst. Tekstarkivet vil særskilt samla nynorske avistekstar, ettersom det alt fins ein god del datamaskinlagra bokmålstekstar. På lengre sikt vil vi også freista betra tilgangen på datamaskinlagra nynorsk sakprosa.

For å informera fagmiljøa vil Norsk tekstarkiv senda halvårlege oversyn over dei tekstar som er lagra. Vi maktar ikkje å tilretteleggja på standardformat alle tekstar vi får hand om, så vi reknar med å måtta lagra store delar ubearbeidd. Her er det viktig at forskarane melder ønska sine til fagleg råd, som skal prioritera innsamlings- og tilretteleggingsarbeidet.

Sixth International ALLC Symposium.

Knut Hofland

Det sjette internasjonal ALLC (Association for Literary and Linguistic Computing) symposium ble arrangert i Cambridge, England, 28. mars - 3. april 1980. Det er 10 år siden dette symposiet ble holdt for første gang, også det i Cambridge. Siden har dette vært arrangert i Edinburgh, Cardiff, Oxford og Birmingham og foredragene fra de tidligere konferansene, unntatt den i Oxford, er utgitt i bokform.

Det var 150 deltakere fra 20 land og nytt medlemsland var i år Kina. Det ble holdt 40 foredrag og arrangert omvisninger i det lokale EDB-miljø på Universitetet. Arrangøren hadde i år bevisst plukket ut nye foredragsholdere slik at det skulle være minst mulig gjengangere. Foredragene ble holdt innen emneområder som bibliografi, parsing, leksikografi, tekstkritiske utgaver, metrikk, vokabularstudier, maskinoversetting, datamaskinassistert undervisning, litterær statistikk, kvantitative metoder, statistisk analyse, programmeringsspråk og programpakker, behandling av naturlig språk, konkordanser og tekst-korpus.

Konferansen åpnet med et invitert foredrag av Jitze Couperus fra CDC. Han er medlem av den amerikanske CODASYL-komiteen. Foredraget omhandlet hvorledes lingvistiske og datamaskinelle disipliner kan forenes og da særlig innen databaseområdet, som var foredragsholderens spesialfelt. Dette ble utdypet i et åpent forum arrangert som en kveldssesjon om bruk av databasesystemer innenfor litterær og lingvistisk databehandling. Behovet for et kraftig, men samtidig enkelt, verktøy ble her understreket. Videre ble det hevdet at forskere inne språkfagene og de humanistiske fagene generelt må være aktivt med og stille krav til programmeringsspråk og systemer før det blir etablert en standard som f.eks. CODASYL standard for databasesystemer. Disse standardene er i neste omgang utgangspunkt for maskinleverandørenes programvare, og dette tilbudet er det som de enkelte brukere i hovedsak må leve med. Det samme gjelder innenfor den tekniske maskinutrustning. Et konkret eksempel på bruk av databasesystemer fikk vi i Linda Miseks foredrag "New and Simple Ways

to Gain Multiple Views of the Patterns in Text". Hun er ansatt ved et IBM forskningssenter og har tidligere laget en tradisjonell konkordans til "Paradise Lost". Foredraget viste hvorledes det var gjort bruk av et nytt interaktivt høynivå spørrespråk på den samme teksten. Teksten var nå strukturert som en database der det var tilføyd opplysninger om personers kjønn, tilhørere, tidsforhold, tema o.l. Uten programmering kunne en søke frem tekstutsnitt basert på kombinasjoner av de forskjellige opplysninger. Forskeren kunne fokusere sin oppmerksomhet på resultatene og unngå tidkrevende programmering. Teksten kunne stadig angripes fra nye synsvinkler.

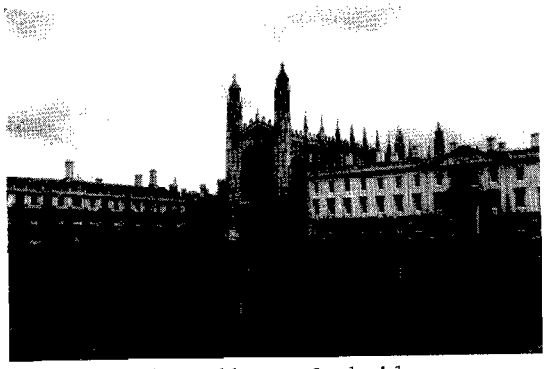
Det var 3 norske foredragsholdere ved konferansen. Geir Kjetsaa holdt et foredrag om sammenligning av 12 omstridte Dostojevskij artikler med 26 kjente artikler av Dostojevskij basert på 15 tekstlige parametre. Statistiske tester utelukker Dostojevskij som forfatter til 4 av de 12 artiklene. Den parameter som slo sterkest ut var type/token forholdet. Stig Johansson ga en oversikt over LOB-corpuset av engelske tekster som nylig er fullført og trakk frem noen frekvensresultater fra studiet av grammatikk og vokabular. Unde tegnede orienterte om det prosjektet som pågår for å lage en Ibsen konkordans og om de prinsipper og metoder som brukes for automatisert lemmatisering.

Det ble holdt to foredrag om mikrodatamaskiner brukt i språkopplæring Begge omhandlet drill med vokabular og uttrykk og grammatiske øvelser Det ble understreket at dette var et tillegg til klasseromsundervisning og et tilbud som tidligere ikke har vært dekket. Det ene prosjektet gjorde bruk av en avansert grafisk fargeskjerm, og det var særlig lagt vekt på presentasjonsformen ved bruk av farger, inndeling av skjermen i deler og bevegelse på skjermen. Mikromaskinene er spesielt godt egnet for slik bruk, og i fremtiden vil de sammen med lydkasseter bli brukt som språklaboratorier. Et av mikroanleggene som ble demonstrert var for øvrig norsk (Tandberg). Undervisningsprogrammene var skrevet i et standard programmeringsspråk slik at systemene var portable.

Foredragene ved konferansen spente over svært mange områder, ofte med bare ett foredrag innen hvert område. Ved utplukk av foredrag

bør en i større grad få frem foredrag som viser hvorledes EDB er brukt og særlig da i nye sammenhenger. Et tilbakevendende trekk ved disse symposiene er mangelen på skriftlig omtale av foredragene på forhånd. Først etter mange anmodninger til arrangørene ble det mot slutten av konferansen distribuert sammendrag av foredragene. De var for øvrig skrevet ut ved hjelp av datamaskin og kunne lett ha foreligget ved starten av konferansen. Ved å legge på deltageravgiften, som er moderat, bør det i fremtiden være mulig å trykke opp resymeene som foredragsholderne likevel sender inn. En del av foredragene fra denne konferansen vil komme i ALLC Journal nr. 1, 1980.

Det hadde ikke meldt seg noen arrangører i England for den neste konferansen, og formannen hadde derfor tatt imot et tilbud fra Pisa om arrangement der i 1982.



King's College, Cambridge

Sommerkurs i statistikk for språk- og litteraturforskere.

Roald Skarsten

9.-13. juni 1980 ble det på Panorama sommerhotell i Oslo arrangert et kurs i bruk av statistiske metoder for språk-og litteraturforskere Kurset kom i stand ved et samarbeid mellom EDB-tjenestene for humanistiske fag ved universitetene og NAVFs EDB-senter for humanistisk forskning. Hovedforeleser var professor Henning Spang-Hanssen fra Institut for anvendt og matematisk lingvistik ved Københavns Universitet. Han foreleste om følgende tema: deskriptiv statistikk, sannsynlighetsregning, statistiske modeller, hypotesetesting og beregning av korrelasjon. Kurset var et elementærkurs i statistikk for språk- og litteraturforskere og det var lagt opp i nær tilknytning til en grunnbok som deltakerne hadde orientert seg i på forhånd. Grunnboken var C. H. Muller, Innføring i metoder for LINGVISTISK STATISTIKK.

Kurset hadde 21 deltakere. Deltakerne kom fra Oslo, Bergen og Trondheim i noenlunde likt antall fra hver av byene. Dagsprogrammet var lagt opp slik at det etter hver forelesningstime ble holdt gruppe øvelser. Deltakerne ble delt inn i 3 grupper, og de måtte der løse statistiske regneoppgaver i tilknytning til temaet for forelesningen. Disse gruppeøvelsene ga et verdifullt bidrag til selve kurset og fungerte samtidig effektivt som en indikator på hvor mye stoff som kunne "fordøyes". Videre fungerte de effektivt som repetisjon av det foreleste stoff og som et pustehull for deltakerne fordi de i en engere krets lettere kunne snakke om de problemer de hadde. Representantene fra EDB-tjenestene for humanistiske fag ved universitetene var gruppeledere.

Nå kan man selvfølgelig ikke i løpet av en uke lære seg alt det som dette kurset inneholdt,og kurset må derfor sees som et ledd i den enkeltes opplæring, som vil skje ut fra grunnboken. Kurset ga imidlertid inspirasjon og perspektiv på stoffet og vil utvilsomt være betydningsfullt for det videre arbeid med boken.

Foreleseren, professor Spang-Hanssen, hadde en heldig hånd med den

pedagogiske tilrettelegging av stoffet. Han maktet å knytte sin fremstilling til grunnboken,samtidig som han satte boken og dens forfatter inn i en internasjonal sammenheng (forskjellige skoleretninger o.l.). Med sin entusiasme og lune danske humor var han et stort aktivum for kurset. Dette kan vi illustrere med å sitere hans bemerkning da han avsluttet forelesningene: "En oppsummering skal være kort, ellers mister man oversikten to ganger!"

I tillegg til det ordinære programmet hadde vi noen kveldssesjoner med inviterte gjesteforelesere. Professor G. Kjetsaa, Slavisk-baltisk institutt i Oslo presenterte emnet: "Statistiske metoder ved bestemmelse av forfatterskap". Han ga her et spennende innblikk i sitt arbeide med å bestemme om 12 artikler av Dostojevski kunne sies å være skrevet av ham eller ikke, og om de burde inngå i den nye utgave av Dostojevskis skrifter.

Professor K. Fintoft fra Lingvistisk institutt, Universitetet i Trondheim, foreleste om "Fonetikk og statistikk". Han kom særlig inn på de metodiske problemer og de problemer man møter ved måling av fonetiske data.

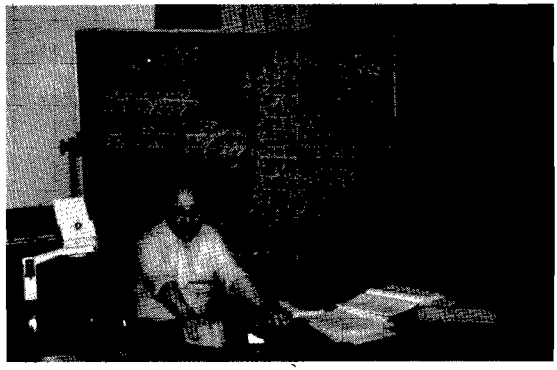
Dosent H. Thorén Amundsen fra Sosialøkonomisk institutt, Universitetet i Oslo, foreleste om "Teoretiske modeller i hypoteseprøving". Hun ga oss en perspektivrik innføring i emnet og særlig var det interessant med hennes drøftelse av forholdet "utvalg-populasjon", alternativt "utvalg-modell". Hun mente at det ikke alltid var så påkrevet med å tenke ut fra en populasjon, men at man med stor fordel kunne tenke "utvalg - modell". Her fikk hun også klar støtte av Spang-Hanssen som siterte seg selv: "Forholdet mellem modelfordeling og empirisk fordeling oppfattes ofte som forholdet mellem en given mængde (population) og en prøve udtaget deraf. Dette synspunkt er naturligt i mange statistiske anvendelser (----) men i al fald i mange sprogvidenskabelige problemstillinger er et sådant synspunkt en kunstig analogi, der let implikerer et sagligt overflødigt eksistenspostulat for en population av vedkommende art, hvad forholdet målerække-model ikke i samme grad frister til".

Kveldssesjonene ga perspektiv på det vi arbeider med og fra deltakerne kom det frem ønske om en oppfølging av dette tiltaket i

form av et forskerkurs om et par års tid, og en oppfordring om også å utgi den fortsettelsesbok som Muller har skrevet. På kurset diskuterte man også utførlig bruken av EDB i statistikk. Selve kurset var lagt opp med bruk av lommekalkulator, men i større prosjekter med maskinleselige data er det naturlig å tenke seg at det vil være behov for programpakker tilpasset de problemer som språk- og litteraturforskere har. Dette spørsmålet burde man også arbeide videre med, og arrangørene fikk av konferansen i oppdrag sammen med professor Henning Spang-Hanssen å undersøke mulighetene for en fortsettelse av det som var satt i gang. Bl.a. ble det nevnt behovet for en nordisk bibliografi som kunne komme med i en eventuell oversettelse av Muller fortsettelses-bok, "Principes et méthodes de statistique lexicale". Spang-Hanssen var særlig opptatt av at man burde skaffe seg corpora som ble statistisk bearbeidet og som kunne tjene som "norm" eller "modell" for den enkelte forskers arbeide med enkeltverk.

Stedet, Panorama sommerhotell, (Studentbyen på Kringsjå) var utmerket med sin frie beliggenhet og gode mulighet for badeliv og turliv. Programmet var hardt, for hardt hvis kurset skulle stå som enkeltstående tiltak, men fordi det sto i en videre sammenheng kunne vi tillate oss et så tettpakket program. For ordens skyld gjøres det her oppmerksom på at den som skriver dette, også var kursleder.

Til slutt kan nevnes at noen av deltakerne hadde sendt inn spørsmål om problemstillinger til Spang-Hanssen på forhånd. Da konferansen startet, hadde han satt seg inn i problemene slik at deltakerne i løpet av konferansen kunne diskutere emnene med ham.



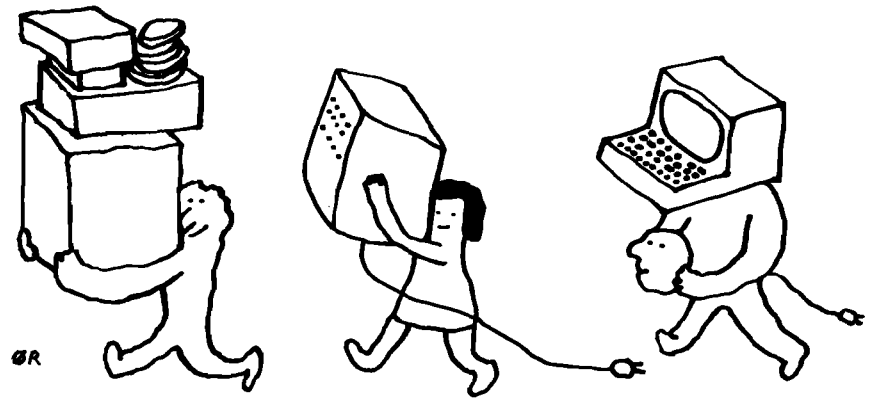
Professor Henning Spang-Hanssen var hovedforeleser på statistikkseminaret.

Meldinger.

NYE LOKALER

Som et ledd i samarbeidsavtalen med Universitetet i Bergen flyttet NAVFs EDB-senter i slutten av juni til nye lokaler i Harald Hårfagresgate 31, 3. og 4. etasje (i samme gate som Park Pension). Gjennom stor velvillighet fra Universitetet i Bergen og ved økonomisk medvirkning fra NAVF er lokalene her blitt pusset opp slik at de i dag gir en godt tjenlig ramme om senterets arbeid. Det må imidlertid også sies at vi allerede ved flyttingen har tatt i bruk hele arealet, og vi håper derfor på utvidelser i huset. Dette vil bl.a. gi oss en mulighet til bedre å ta hånd om de korttidsstipendiater, hospitanter og gjester for øvrig som jevnlig besøker senteret. Vår adresse er heretter: NAVFs EDB-senter for humanistisk forskning
Harald Hårfagresgt. 31, Boks 53
5014 Bergen-Universitet

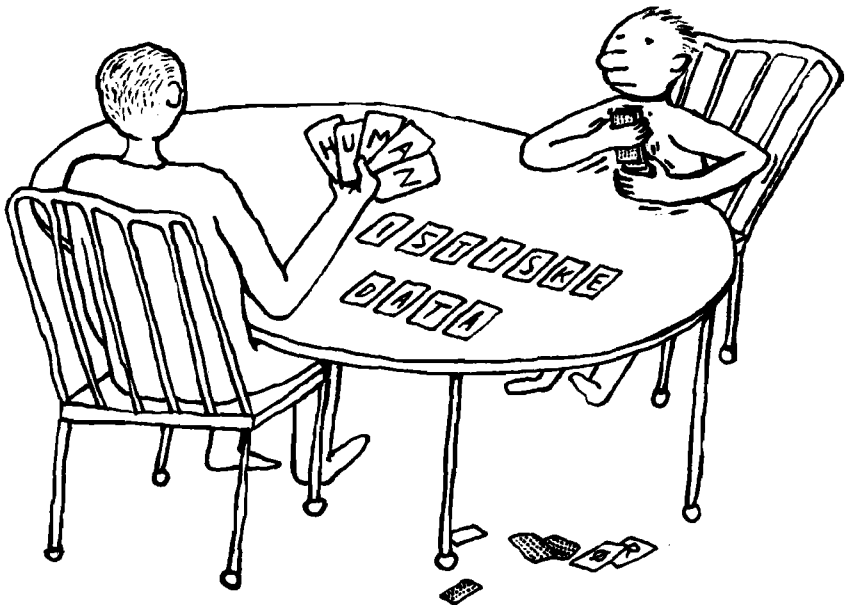
Vi er fremdeles knyttet til Universitetets sentralbord, tlf. 05/210040. Vi tar gjerne imot besøk i våre nye lokaler.



SENTERET FLYTTER

HUMANISTISKE DATA I OPPLØSNING

Vi beklager sterkt at Humanistiske Data nr. 1-2 1979 p.g.a. en produksjonsfeil gikk i oppløsning etter første gangs bruk. Uhellet ble først oppdaget da størstedelen av opplaget var gått ut av huset hvilket gjorde saken enda verre. Dersom det er noen som ønsker å erstatte sitt eksemplar med en bedre holdbar utgave, ber vi om å få beskjed. Så håper vi på bedre lykke denne gang------.



SENTERETS ÅRSMELDING 1979

NAVFs EDB-senter har i år valgt å sende ut årsmeldingen for det foregående år til et større antall kontakter, samarbeidspartnere og potensielt interesserte. Målet har vært å gjøre senterets mangesidige virksomhet kjent for de humanistiske fagmiljøene og utvide kontaktflaten med andre FOU-miljøer innen- og utenlands. De som ønsker årsmeldingen tilsendt, kan henvende seg til NAVFs EDB-senter.

UTREDNINGSARBEID OM DATATEKNOLOGIENS KONSEKVENSER

Norges almenvitenskapelige forskningsråd vedtok i januar 1980 å be de enkelte faglige råd i NAVF om å utrede datateknologiens konsekvenser for de ulike forskningsfelt. Utredningens formål er i første rekke å gi NAVF et bedre fundament i arbeidet med å klarlegge fremtidige forskningspolitiske mål og strategier.

I tillegg til å klarlegge datateknologiens innvirkning på forskningsarbeidet, ble også rådene bedt om å vurdere både de umiddelbare og de mer langsiktige samfunnsmessige konsekvenser av denne teknologi.

Utarbeidingen av grunnlagsdokumenter vil være sluttført i sommer, og i løpet av høsten vil det bli utarbeidet et sammenfattende dokument fra NAVF med bistand fra Norsk regnesentral.

NAVFs EDB-senter har ved eget utredningsarbeid og med støtte i konsulentuttalelser utarbeidet et omfattende dokument for Rådet for humanistisk forskning. Det er planen å bearbeide dette materialet med tanke på utgivelse av en separat rapport til spredning i de humanistiske fagmiljøer.

Senterets personale og prosjektmedarbeidere på trappen til Villavei 10.

KONSULENTTJENESTE FOR HUMANISTER VED UNIVERSITETET I TROMSØ

Rådet for humanistisk forskning, NAVF, har på sitt 1980-budsjett satt av midler til ½ konsulentstilling til delfinansiering av en konsulenttjeneste i EDB for humanister ved Universitetet i Tromsø. Rådet satte som vilkår at Universitetet i Tromsø etter en 2-års periode ville overta den konsulentstillingen som opprettes. Det viser seg nå at Universitetet i Tromsø vanskelig kan gi slike forhåndstilsagn, og det arbeides for øyeblikket med å finne alternative organisatoriske løsninger.

I første halvår 1980 har cand. philol. Gunnar Thorvaldsen, vikarierende amanuensis ved ISV, vært engasjert av NAVFs EDB-senter som konsulent i Tromsø i deltidsstilling. Han har særlig gitt konsulent bistand til fagmedarbeiderne ved Institutt for språk og litteratur.

Vi regner med å kunne få i stand en lignende ordning i høstsemestere

LITTERATURVEILEDNING

I løpet av den siste 10-års perioden er tilbudet av bøker og tidsskrift som omhandler databehandling i ett eller flere humanistiske fag, øket sterkt.

NAVFs EDB-senter har lenge vært oppmerksom på behovet for jevnlig informasjon om viktige bøker og tidskrift, men har til nå ikke vært i stand til å ta opp denne oppgaven på en planmessig måte.

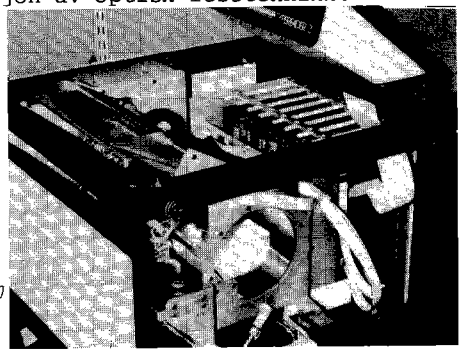
Vi ønsker imidlertid fra nå av å satse mer på dette informasjonsarbeidet og har således planer om å utarbeide en oversikt over sentra litteratur til neste nummer i Humanistiske Data.

I denne forbindelse vil vi også gjerne ha kontakt med noen som kan hjelpe oss i dette arbeidet, enten ved enkeltstående oppdrag eller med faste oppgaver knyttet til orientering om nye bøker og tidsskrif eller ved mer inngående faglig vurdering av enkelte utgivelser. Bidragsyterne kan regne med honorar for arbeidet.

OPTISK LESING (OCR)

Universitetet i Bergen og Rådet for humanistisk forskning, NAVF, har samarbeidet økonomisk om innkjøp av et optisk leseutstyr av enklere type som leser tall og bokstaver (små og store) skrevet på skrivemaskin med spesiell skrifttype (OCR-B kulehode). Utstyret (Hendrix' Typereader 2) ble installert rundt årsskiftet 1979/80 og er nå etter en del innkjøringsproblemer i regulær drift. Maskinen, som er installert ved EDB-avdelingen, Universitetet i Bergen, kan betjenes av brukeren selv eller av en operatør. I løpet av det siste halvåret er leseren blitt brukt til meget varierte oppgaver (arkivdata, trykte skjema, løpende tekst og tall) og registreringsmåten har svart til forventningene.

De som ønsker nærmere opplysninger om utstyret eller bruken av det, kan henvende seg til NAVFs EDB-senter. Det foreligger også en håndbok om registrering av materiale for optisk lesing og om betjeningen av Typereader 2. Senteret kan som en prøveordning og mot en godtgjørelse forestå innlesing av data på den optiske leseren for humanistiske fagmiljøer. Også ved Universitetet i Tromsø har det vært utført prøvearbeid med optisk lesing med tanke på å etablere dette som et standardtilbud ved EDB-sentret i Breivika.
For øvrig var det (for senteret i alle fall) et gjennombrudd på dataregistreringsfronten da det ved årsskiftet forelå muligheter til å få lest bøker og andre trykksaker på en avansert optisk leser i Stockholm. Etter de prøver senteret har foretatt, er kvaliteten på lesingen meget høy og leseren kan, i alle fall når det gjelder store datamengder, konkurrere prismessig med andre registreringsmåter. I et senere nummer vil vi gi en mer utførlig presentasjon av optisk leseteknikk.



En optisk dokumentleser er i drift ved Universitetet i Bergen fra 1980 av. De skrevne arkene leses i trommelen midt i maskinen.

REGISTRERING AV EDB-PROSJEKTER I HUMANISTISK FORSKNING

NAVFs EDB-senter oppfordret i januar 1980 de humanistiske fagmiljøer til å sende inn opplysninger om de EDB-prosjekter som på det tidspunkt var i gang innenfor og utenfor universitetene.
For å lette arbeidet fulgte det med bladet et skjema som kunne nytte i arbeidet.

Som resultat av miljøenes egeninnsats og senterets oppfølgings-/purringsarbeid er det i dag innkommet opplysninger om ca. 60 prosjekter.

Siden det i utgangspunktet ikke var lagt opp til en direkte henvendelse til enkeltpersoner i miljøene, er det vanskelig å si i hvilken grad vi har maktet "å sope rent" i de ulike fagmiljøene. Vi finner imidlertid at opplysningene som er fremsendt, gir et meget interessant bilde av den utvikling som har skjedd siden forrige gang vi laget et oversyn. (Humanistiske Data nr. 1 1974).

En fyllestgjørende presentasjon av prosjektene vil sprenge rammen for Humanistiske Data. Vi er derfor kommet til at vi vil presentere prosjektopplysningene i en egen rapport i senterets rapportserie. De som har levert bidrag, vil få rapporten tilsendt - andre kan bestille den hos NAVFs EDB-senter. Kunngjøring om rapporten vil bli gitt ved egne meldinger, og den vil bli omtalt i neste nummer av bladet.

PROSJEKTHÅNDBOK FOR HUMANISTISK FORSKNING

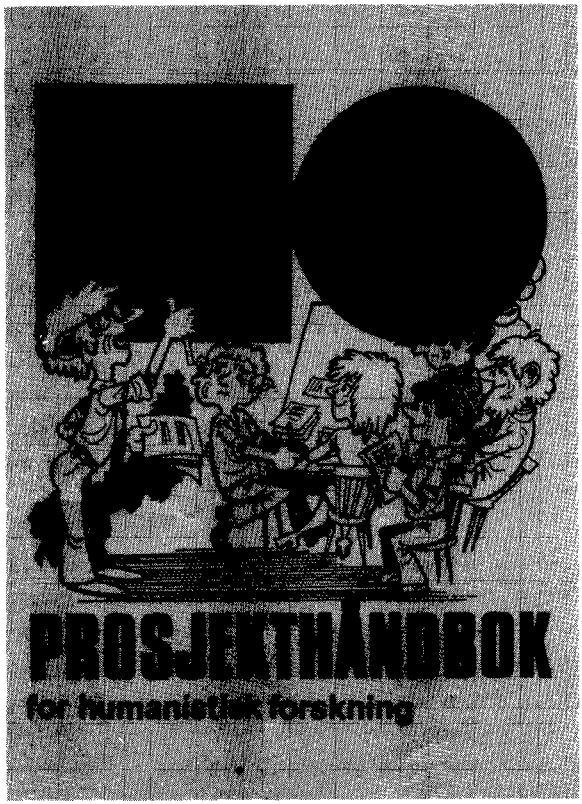
Bruk av prosjekter har i de senere år blitt mer vanlig også i humanistisk forskning. Forskere har imidlertid hatt vansker med å finne frem til hensiktsmessig litteratur for å sette seg inn i problemkomplekset rundt prosjektplanlegging og prosjektledelse.

Rådet for humanistisk forskning vedtok derfor høsten 1977 å nedsette en redaksjonskomité for utarbeidelsen av en håndbok i prosjektledelse for humanistisk forskning. Komitéen fikk følgende sammensetning:

Professor Jørn Sandnes, formann

Forskningsstipendiat Eskil Hanssen
Ass. direktør Anne-Lise Hilmen
Førstekonservator Arne B. Johansen

Komitéen fikk i oppdrag å lage en håndbok i prosjektledelse i løpet av 1 år. Arbeidet ble startet våren 1978, og håndboken forelå i juni 1979.



Første utgave skal revideres senere etter at kommentarer og synspunkter fra leserne er kommet inn.
Håndboken kan bestilles ved henvendelse til:
NAVF, Rådet for humanistisk forskning
Munthesgt. 29, Oslo 2, tlf. 02/565990

UTVIKLING AV ET GENERELT TEKSTSØKESYSTEM (SIFT-PROSJEKTET)

Siden januar 1980 har det pågått arbeid med å utvikle et generelt EDB-system for søking i fri tekst (SIFT-prosjektet).

Arbeidet ledes av en styringsgruppe med representanter for R-direktoratet, Lovdata, Institutt for privatrett, Avdeling for EDB-spørsmål v/Universitetet i Oslo, NAVFs EDB-senter for humanistisk forskning og Norsk Data. Formann i styringsgruppen er direktør Kåre Fløisand, R-direktoratet.

Bakgrunnen for arbeidet er de positive erfaringer med et nasjonal koordinert utviklingsarbeid knyttet til programsystemet NOVA*STATUS i årene 1974 - 1979. I dette prosjektet samarbeidet bl.a. R-direktoratet, Institutt for privatrett, Avdeling for EDB-spørsmål og NAVFs EDB-senter for humanistisk forskning om en videreutvikling av det engelske tekstsøkesystemet STATUS I.

I løpet av de siste årene er NOVA*STATUS tatt i bruk til en rekke oppgaver knyttet til humanistisk og juridisk databehandling, og systemet er brukt i flere statlige institusjoner bl.a. R-direktoratet Forbrukerrådet og Forbrukerombudsmannen. Systemet er i dag operativt på Honeywell-Bull, UNIVAC, DEC og NORD maskiner og er fritt tilgjengelig for statlige institusjoner. En CYBER-versjon er under implementering i Tromsø.

Utgangspunktet for arbeidet i SIFT-prosjektet, som flere av samarbeid partnerne i NOVA*STATUS står sammen om, er meget gunstig idet en kan bygge på de tidligere erfaringer fra systemarbeid med tekstsøking. På en rekke punkter vil imidlertid det nye systemet bety en forbedring av NOVA*STATUS. Følgende kan nevnes:
Oppdateringsmulighetene forbedres, og parallell søking og oppdatering tillates. Filstrukturen effektiviseres for å kunne håndtere meget store dokumentsamlinger. Kommandospråket forbedres. Søkespråket gjøres enda kraftigere. Parallell søking i flere databaser blir mulig. De høyest prioriterte generelle krav vil være:

- maskinuavhengighet: systemet skal gjøres fritt tilgjengelig, og bør følgelig ikke være avhengig av en spesiell type maskinutrustning

- modularitet: systemet må kunne tilpasses en rekke forskjellige behov. Implementering på et nett av mikromaskiner er også ønskelig.
- effektivitet: enkelte tekstsøkerutiner er svært tidskritiske.
- brukervennlighet: systemet vil i stor utstrekning bli benyttet av personer uten EDB-kunnskaper.

Prosjektarbeidet har i første halvår i hovedsak dreiet seg om systemspesifikasjon. Arbeidet med spesifikasjonen har tatt noe lengre tid enn på forhånd antatt da det ble funnet nødvendig med en meget grundig gjennomgang av de forskjellige funksjonene i tekstsøkesystemet. Det har også vært kontakt med en rekke miljøer hvor det arbeides med de samme problemene, f.eks. med MERLIN-prosjektet ved British Library og Siemens' utviklingsavdeling i München, som i flere år har arbeidet med utvikling av et meget avansert tekstsøkesystem (CONDOR).

Det foreligger pr. 20.6.1980 et eget dokument med detaljerte spesifikasjoner for det nye SIFT-systemet.

I løpet av utviklingsperioden vil det bli lagt vekt på å opparbeide en nær kontakt med brukermiljøer og interessenter både i Norge og Europa for øvrig, bl.a. i form av seminarer og mindre konferanser.

Arbeidet med grunnversjonen av SIFT-systemet vil kreve ca. 7 årsverk. Det er ventet at arbeidet vil gå frem til sommeren 1981.

NAVFs EDB-senter for humanistisk forskning er representert i prosjektgruppen ved cand.real. Øystein Reigem.

TEKSTSAMLINGER FRA INTERNATIONAL COMPUTER ARCHIVE OF MODERN ENGLISH (ICAME), BERGEN

På neste side er gjengitt en del av det bestillingsarket som brukes av ICAME i forbindelse med distribusjon av engelske tekstsamlinger. Det vises for øvrig til meldingsbladet ICAME NEWS, neste nummer sept./okt. 1980. Se også Humanistiske Data nr. 1-2 1979 s. 8.

Brown=	The Brown University Corpus of Present Day American English

LOB= The Lancaster - Oslo - Bergen Corpus, består av tekster på tilsammen 1 mill. ord fra britisk-engelsk, sammenstilt etter de samme kriterier som the Brown Corpus

Brown Text I Den orginale versjon av tekstene (bl.a. bare store bokstaver)

Brown Text II En versjon som er typografisk bearbeidet i NAVFs EDB-senter.

London-Lund Text= Talemålstekster på ca. 170.000 ord fra The Survey of English Usage, London, tilrettelagt for data-behandling av Engelska Institutionen, Lunds Universitet.

Magnetic tapes	9-track, 1600 FPI		7-track, 800 FPI		7-track, 556 FPI	
	Number of tapes	Price N.kr.	Number of tapes	Price	Number of tapes	Price N.kr.
Brown: Text I	1 1200 ft.	150	1 2400 ft.	275	1 2400 ft.	275
Brown: Text II	1 1200 ft.	175	1 2400 ft.	300	1 2400 ft. 1 1200 ft.	300
Brown: Texts I+II	1 1200 ft.	200	1 2400 ft. 1 1200 ft.	525	2 2400 ft.	550
Brown: KWIC concordance	4 2400 ft.	1100	11 2400 ft.	3375	15 2400 ft.	3900
LOB: Text	1 1200 ft.	150	1 2400 ft.	250	1 2400 ft.	250
LOB: KWIC concordance	5 2400 ft.	1250	12 2400 ft.	3650	16 2400 ft.	4100
London-Lund:Text	1 1200 ft.	150	1 1200 ft.	200	1 1200 ft.	200

Brown: KWIC concordance (microfiche) Price: 350 N.kr.

LOB: KWIC concordance (microfiche) Price: 350 N.kr.

Please, add for postage and handling:

for each 1200 ft. tape 25 Norwegian kroner

for each 2400 ft. tape 35 " "

for each microfiche set 10 " " (overseas air mail: 20)

SEMINAR OM "HISTORISKE DATABASER I NORDEN", UTSTEIN KLOSTER 12. - 15. MAI 1980

I tiden 12. - 15. mai i år ble seminaret "Historiske databaser i Norden" holdt på Utstein kloster. Seminaret ble arrangert av Universitetet i Bergen med amanuensis Jan Oldervoll, Historisk institutt som kursleder. Det deltok i alt 20 personer med tilknytning til universiteter og forskningsinstitusjoner i Danmark, Sverige og Norge. Deltakerne hadde sin faglige bakgrunn i historie og/eller databehandling.

Programmet besto av flere arbeidsøkter pr. dag med veksling mellom innledningsforedrag og diskusjon. I løpet av seminartiden ble det i alt tatt opp 6 hovedtema som dekker mange viktige sider ved bruk av EDB i historie.

Følgende innledningsforedrag ble holdt:

Registrering og oppretting av data.
Hovedinnleder: Gunnar Thorvaldsen, Universitetet i Tromsø.

Koding av data. Korleis overføra dataene til eit format ein kan kjøra statistikk på?
Hovedinnleder: Jan Oldervoll, Universitetet i Bergen.

Lagring av data, statistikkjøring.
Hovedinnleder: Ivar Fonnes, Universitetet i Oslo.

Lenking av data (Record linkage).
Hovedinnleder: Erik Søderlund, Demografiska Databasen, Umeå-Haparanda.

Søking. Hvordan hente frem indivudaldata?
Hovedinnleder: Eva Appel, Demografiska Databasen, Umeå-Haparanda.

Oppbygging av en database. Hvilke data; for hvem og til hvilken pris?
Hovedinnleder: Hans Chr. Johansen, Universitetet i Odense.

Konsentrasjonen om EDB og historie ga anledning til en utførlig presentasjon av temaene og en grundig diskusjon av dem. Samtidig ble det rikelig høve til å omtale personlige erfaringer fra tidligere

eller pågående EDB-arbeid i tilknytning til historie.

Ideen bak seminaret var å skape et alternativ til de store og tematisk omfattende EDB-konferansene som holdes internasjonalt, og seminaret må sies klart å ha vist sin berettigelse.

Rammen om seminaret var den beste, og deltakerne ble tatt meget godt hånd om av det lokale vertskap.

SECOND INTERNATIONAL CONFERENCE ON DATA BASES IN THE HUMANITIES AND SOCIAL SCIENCES, MADRID 16. - 19. JUNI 1980

Som en fortsettelse av den første konferansen i denne serie ved Dartmouth College, New Hampshire, USA i august 1979 ble det i tiden 16. - 19. juni holdt en tilsvarende konferanse ved Fakultet for informatikk, Det polytekniske universitet i Madrid.

Vi siterer fra den folder som ble utsendt før konferansen:

"The aims of this International Conference on Data Bases in the Humanities and Social Sciences are essentially to offer the possibility of personal contacts and the discussion of a widely interdisciplinary matter: the redefinition of objectives and methods in human sciences contributing to the exchange of ideas and opinions between humanities and social sciences on one hand and between both and computer science on the other.

The common subject of the Conference is data bases both from the operative and methological standpoints. This amounts to many different problems ranging from the establishment of data - not basically numerical to their use in Human Sciences.

Contributions on specific experiences or on technical aspects of data bases will be regarded in the same way as those devoted to methodological or theoretical aspects. It should also be pointed out that if computer science has a considerable bearing on human sciences studies, the latter is influencing computer science very significantly. Therefore, contributions on this kind of mutual

effects are required. Furthermore, papers on implications of data bases on the organizations and research economics and pedagogy in Human Sciences would be desirable. Finally, the conference would welcome any discussion on changes likely to come about in the role traditionally assigned to social sciences and to the humanities not only from the cultural standpoint but also in the political and economic fields.

The technical and methodological maturity of the Humanities and Social Sciences with regard to computer science was indeed proved at the Dartmouth College Conference. The 1980 Madrid Conference should deeply consider this matter and make clear the role of data bases through interdisciplinary discussions".

NAVFs EDB-senter var ikke representert på denne konferansen. Dersom noen av våre lesere deltok på konferansen, stiller vi gjerne spalteplass til disposisjon for en faglig omtale av begivenheten.

EDB-KURS FOR ANSATTE I ARKIVVERKET

NAVFs EDB-senter vil i samarbeid med Riksarkivaren arrangere et EDB-kurs for ansatte i arkivverket. Kurset vil bli holdt i senterets lokaler og varer fra 25. august til 5. september.

Arkivverket har i lengre tid vært oppmerksom på de utfordringer som datateknikken vil gi for arkivsektoren.

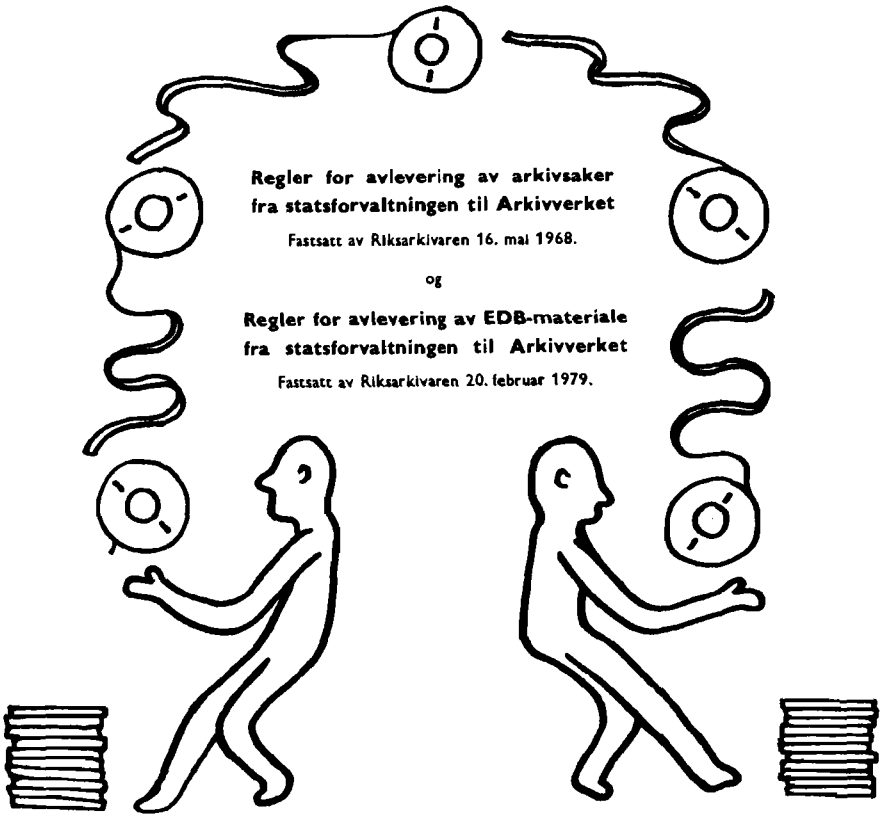
Det kan i denne sammenheng nevnes at det nylig (20.2.79) er fastsatt regler for avlevering av EDB-materiale fra statsforvaltningen til Arkivverket. Disse regler fastslår at EDB-materiale som oppstår som resultat av en offentlig institusjons virksomhet, er å betrakte som arkivmateriale og er underlagt de samme regler for arkivbegrensning, kassasjon og avlevering til Arkivverket som tradisjonelt materiale.

Riksarkivaren har også nedsatt en nasjonal EDB-komite som skal drøfte EDB-spørsmål og legge planer for å sette etaten i stand til å utnytte de muligheter som EDB kan gi i arkivarbeid.

Det kan videre nevnes at flere arkivarer i løpet av de siste årene har deltatt som korttidsstipendiater i senterets stipendprogram.

På kurset i Bergen vil deltakerene få innføring i generelle EDB-emner og trening i selv å bruke terminal på enklere EDB-oppgaver i tilknytning til et arkivmateriale som tilrettelegges for kurset. Gjennom forelesninger og demonstrasjoner vil også aktuelt prosjekt-arbeid på feltet i Norge bli gjennomgått, og tilgjengelig program-utrustning for arkivformål vil bli demonstrert. Kurset i Bergen vil bli etterfulgt av et kurs hvor særlig spørsmål i tilknytning til bruk av EDB i statsforvaltningen vil bli tatt opp.

I neste nummer vil det bli gitt en rapport om kurset i Bergen.



COLING 80

The 8th International Conference on Computational Linguistics

第8回計算言語学国際会議

Tokyo, Sept.30—Oct.4, 1980

Organizations

Sponsored by: ICCL (International Committee on Computational Linguistics)

In association with:

情報処理学会　電子通信学会

計量国語学会　視聴覚情報研究会

Date:

September 30 (Tue.) — October 4 (Sat.), 1980

Location:

Nippon Toshi Centre Hall
2-4-1 *Hirakawa-cho*
Chiyoda-ku, Tokyo
100 Japan
Phone: (03) 265-8211

Office of COLING 80 Organizing Committee
Room 362, Hotel New Japan
2-13-8 Nagata-cho, Chiyoda-ku
Tokyo 100 Japan
Phone: (03) 581-5511 ex 8362
Telex: COLING 80 No. 362
New Japan J-22499

1. THE 8TH COLING IN TOKYO

The 8th International Conference on Computational Linguistics (COLING 80) will be held on September 30 — October 4, 1980, at Nippon Toshi Centre, Akasaka, Tokyo, Japan.

This conference, *following seven successful previous* conferences, offers a chance for all researchers in this field to become acquainted with the most advanced results of computational linguistics research all over the world, and especially in Japan. It will provide a place for the discussion of a variety of computational linguistics topics.

The official language of the conference is English and *French.*

2. CONFERENCE REGISTRATION

For those of you who wish to register in advance (this is highly recommended), please complete and return the attached form to the *Office of Organizing Committee* as soon as possible. Your name will then be filed for further correspondence.

REGISTRATION FEE

The registration fee for the conference *is payable in* Japanese currency and is listed as follows.

early registration	¥22,000 (until July 31, 1980)
late registration	¥27,000 (from August 1, 1980)
student rate	¥ 8,000 (include letter from a professor which confirms your student status.)

The fee covers the cost of participation in the conference, a copy of the proceedings, and cocktail reception.

To Register

Complete Application Form A and return to:

Office of COLING 80 Organizing Committee
Room 362, Hotel New Japan
2-13-8 Nagata-cho, Chiyoda-ku
Tokyo 100, Japan

Remittance should be made by bank transfer at your earliest convenience to:

COLING 80 Office, Account No. 055-1284972
DAI-ICHI KANGYO BANK, Akasaka Branch
Minato-ku, Tokyo 107, Japan

ISSUANCE OF RECEIPT

Upon receiving your registration form and confirming your remittance of the appropriate fees, the secretariat *will* send you a receipt containing your registration number.

CANCELLATIONS

All cancellations shall be made in writing to the *secretariat of Organizing Committee.* Up to July 31, 1980, 80% of the registration fee will be refunded. From August 1, 1980, no part of the registration fee will be refunded, however, proceedings of the conference will be sent after the conference has concluded.

REGISTRATION

Registration will start in September 29, PM 3:00 at Toshi-Centre Hall, and *will continue until the end* of the conference. JTB will also open their desk at Toshi Centre Hall, where you can arrange tours in Japan.

See Map A of page 16 showing the location of conference site.

3. PROGRAM SUMMARY

COLING 80 AT A GLANCE **(subject to change)**

September 29 (Mon.) PM 3:00 Registration (at Nippon Toshi Centre Hall)

September 30 (Tue.)	Opening session Dictionary Data Base Linguistic Meaning Panel: Computational Linguistics and Linguistic Theory	Montague Systems Syntax	Cocktail Reception
October 1 (Wed.)	Thesaurus Generation Fuzzy Word Meaning Analysis Formal Representation Machine Translation Montague Grammar Panel: Terminology	Kana and Kanji Kana and Kanji Programming Languages Miscellaney Data Bases Linguistic Topics Parsers in Prolong Machine Translation	Pragmatics Dialogue Miscellany Semantic Topics Semantic Processing Semantic Processing
October 2 (Thur.)	Full-Day	Excursion of Kamakura	
October 3 (Fri.)	Analysis of English Keyboard Assignment of Key Codes Interactive Translation System Paradigms of MT Hierarchical Meaning Analysis Panel: Education	Semantics Syntax Speech Recognition Syntax Information Systems	Some other Topics Banquet
October 4 (Sat.)	First Language Acquisition Conjunctions and Modularity Natural Language Communication Dialogue Error Correction Panel: Machine Translation	Concordances Input and Output Machine (Aided) Translation Poetics Miscellany Semantic Topics	

De som er interessert i flere opplysninger om konferansen og evt. kopi av informasjonsbrosjyre kan kontakte NAVFs EDB-senter.

Senterets rapportserie.

RAPPORT NR. 1 EDB i gjenstandsfagene. Rapport fra en konferanse i Bergen, 18. og 19. april 1978. September 1978. Pris kr. 10,-

RAPPORT NR. 2 Et norsk datamaskinelt tekstkorpus. Rapport fra en konferanse i Bergen, 19. og 20. oktober 1978. Februar 1979 Pris kr. 17,50

RAPPORT NR. 3 Rapport fra den nasjonale konferanse om EDB i språk og litteraturforskning, 4. og 5. januar 1979. Mars 1979. Pris kr. 20,-

RAPPORT NR. 4 Oppbygging av EDB-katalog for folkemusea i Hordaland og kulturgeografisk registrering på Vestlandet. April 1978. 2. opptrykk oktober 1979. ISBN 82-7283-000-0. Pris kr. 10,-

RAPPORT NR. 5 Rapport fra NKKM's EDB-komite. August 1979. ISBN 82-7283-001-9. Pris kr. 10,-

RAPPORT NR. 6 Prøveprosjekt med EDB ved Norsk Folkemuseum. Oktober 1979. ISBN 82-7283-002-7. Pris kr. 10,-

RAPPORT NR. 7 Ivar Fonnes: Norsk landbruksordbok. Prosjektrapport om databehandling og tilrettelegging for trykking. September 1979. ISBN 82-7283-008-6. Pris kr. 20,-

RAPPORT NR. 8 SEFRAK. Rapport frå prøveprosjekt for databehandling av kulturminneregisteret. Oktober 1979. ISBN 82-7283-003-5. Pris kr. 15,-

RAPPORT NR. 9 Jostein H. Hauge og Sigbjørn Århus: Dataregistrering i humanistiske fag med vekt på optisk lesing. August 1978. 2. opptrykk oktober 1979. ISBN 82-7283-004-3. Pris kr. 10,-

RAPPORT NR. 10 Roald Skarsten: Innføring i SPSS for humanister. November 1977. 2. opptrykk november 1979.
ISBN 82-7283-005-1. Pris kr. 10,-

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 Sciences. November 1979. ISBN 82-7283-007-8.

RAPPORT NR. 12 EDB og manuskriptregistraturer. Oktober 1977. 2. opptrykk november 1979.
ISBN 82-7283-009-4. Pris kr. 15,-

RAPPORT NR. 13 Datatjenester for og datasamarbeid mellom kunst- og kulturhistoriske museer. Februar 1980.
ISBN 82-7283-010-8. Pris kr. 25,-

RAPPORT NR. 14 NOVA*STATUS, systemdokumentasjon. Brukerveiledning. 2. opptrykk februar 1980.
ISBN 82-7283-011-6. Pris kr. 15,-

RAPPORT NR. 15 Ivar Fonnes: Tekstsøking på tegnnivå. Januar 1980.
ISBN 82-7283-012-4 Pris. kr. 10,-

Summary.

AN APOLOGY

The Editor regrets very much the defective binding of the previous issue of HUMANISTISKE DATA caused by a production error.

On request the Editor will be pleased to supply the reader with a more durable copy.

SEFRAK. PRØVEPROSJEKT FOR DATABEHANDLING AV KULTURMINNEREGISTERET

The pilot study reported by Ove Magnus Bore, cultural advisor to the County Curator, Hordaland, is connected with the nation-wide registration of information on important cultural objects dating from the Reformation onwards.

Of special importance to the study, were the possibilities of utilizing both formated and free text descriptions in the computer analysis of the data. Additionally new available methods for the presentation of data (the use of microfiche and automatic card production) were demonstrated.

As a part of the study a data base was produced for on-line processing purposes. Spurred by the positive outcome of the pilot-study the Ministry of the Environment has decided to continue with the computer work, with the ultimate goal of creating a series of regionally-steered computer projects to handle this kind of cultural material. Moreover, the intention is to coordinate such work with parallel activities in archaeology.

It is hoped that information thus made available will be of great value both for purposes of future societal planning and for cultural research.

The design of the array of computing techniques used was the responsibility of the Norwegian Computing Centre for the Humanities.

BRUKEN AV EDB I TEATERVITENSKAPELIG FORSKNING

Within the field of the History of Theatre in Norway the study of scenography is an underdeveloped field according to Rune Johansen (Research Fellow at University of Bergen) in his article. Having expressed this view he goes on to show how traditional archival research facilities can be made more effective and flexible by converting the catalogue material to computer form.

Today the efficiency of conventional methods of retrieving basic research data depend largely on the experience and resources of the staff in charge of the archives and on their willingness to cooperate. In many specific studies there is a conspicous waste of time when trying to locate a photograph or a set of data pertaining to a specific play. In his article, Rune Johansen is a spokesman for a coordinated, collective effort in establishing a computer data base comprising basic information related to the History of Theatre in Norway.

As a case in point he briefly comments on how he intends to study scenographic developments at the Nationaltheatret, Oslo, in the period of 1908 - 35. As a by-product he will obtain a computerized exemplar of the archive developed for his own field of study.

ER TILRETTELEGGING AV PRIMÆRKILDER MERITERENDE ARBEID?

This article questions the nature of scholarly work in the preparation of primary source-material for processing by computers, particularly in historical research. The author, Eirik Lien, computer consultant at the College of Art and Science, University of Trondheir finds that data-preparation frequently is both onerous and time-consuming. While such work obviously has routine aspects much of it presupposes the insight and careful judgement of an experienced scholar in the field. However, this state of affairs is insufficiently recognized in Norway at least in the way the system of academic rewards functions.
In the article a case is also made for a more fitting evaluation of the role of the computer consultant who is often connected with such

activities. The present circumstances may lead to a situation where too few academics would be willing to devote themselves to such work which is, after all, fundamental to many research activities during their initial stage. This in turn might well bring about a dearth of sources available for computer-oriented studies which depend on effective means of using large amounts of well-prepared primary source-material.

EMIGRANTFORSKNING - HISTORIE PÅ INDIVIDNIVÅ

Gunnar Thorvaldsen, University of Tromsø, evaluates in his article the new dimensions opened up by modern computers within the field of microhistorical research tasks.

While scholars of the history of emigration normally used to focus on main tendencies, taking into account the large scale entities, (e.g. the situation of regions and of nations as a whole) as the basis for arriving at conclusions, computer facilities now enable the investigator to study more broadly the lives and destinies of individual emigrants and their families (record linkage).

A series of examples are given to demonstrate how tricky problems in detecting individuals and tracing the paths of their emigration history can be solved by using source-materials such as clerical records and census data.

There are plans to establish a Computing Centre for Historical Data at Tromsø. If these plans materialize on a permanent basis a huge amount of historical data will be processed at the Centre comprising the most important nation-wide censuses of the latter part of the 19th and the early 20th century.

It is hoped that this material will among other things create a fertile basis for studies in demography in Norway, and will function as a reservoir providing researches with an abundance of data to be utilized at the critical discretion of individual investigators.

TIENDPENGESKATTEN 1520/21 I EDB-VERSJON

Two research projects in the University of Trondheim are presented, one aiming at the study of land ownership in Trøndelag in the middle of the 16th century and the other connected with on-going research into the customs of name-giving in the same period.

For both purposes a comprehensive set of tax lists, the so-called "Tiendpengeskatten", has been prepared for data processing purposes in collaboration with Eirik Lien, the author of the report. One major objective has been to render a word to word computer-readable counterpart of the source. The tax lists include the name of the individual tax-payer, the amount of tax paid, and the means of payment (examples are given in the report). By means of careful coding a computer-accessible version was achieved which made feasible the tracing of individuals across a variety of tax-lists for the historical analysis. Specially designed programs systematically brought together all orthographical variants of the first and second names in a way which facilitated further linguistic and cultural analysis.

NORSK TERMBANK

Under the joint auspices of the Norwegian Council for Language Planning, the Board of Technical Terminology, The University of Bergen and The Norwegian Universities Press, an institution called The Norwegian Term Bank has lately been established in Bergen.

The object of the Norwegian Term Bank is to coordinate, promote and initiate projects within the field of technical terminology.

To this end the institution has developed flexible computer facilities as a framework for handling a variety of different projects in terminology now in progress in Norway.

The standard internal format of the data and the related data base system makes it easy to access, extract or correct all or part of the information provided for a given item in the data base. As of

June 1980 a total of 20.000 entries had been included in the data base, and a test version of the on-line accessible form will be implemented in the autumn.

Those interested in the work should write to the author:
Håvard Hjulstad
Norsk Termbank
c/o PDS, Nordisk institutt
Harald Hårfagresgt. 29
5014 Bergen-Universitetet

OPPSTARTING AV NORSK TEKSTARKIV

The Norwegian Text Archive has been established as a five-year project to provide a variety of research milius with modern Norwegian text data in computer form. The main user groups are assumed to be linguists and scholars within the field of applied linguistics generally who are responsible for developing teaching materials for special purposes, e.g. in the teaching of groups of immigrants.

Thus far the Text Archive has been developing a standard way of storing text material from newspapers and publishing houses and implementing methods for the convenient transformation of the type-set texts to a format more suitable for language study.

An application has been sent to The Norwegian Publishers' Association to obtain permission for using the material for bona fide research and developmental purposes.

Moreover steps have been taken to collect information on the considerable volume of texts already available at the universities, with a view to issuing an inventory.

The Norwegian Council for Language Planning has plans for a long range study of the effects of the three major spelling reforms in the 20th century on the writing style of contemporary authors. The Text Archive has been called upon to review and estimate the expenditure involved in these plans.

In carrying out its clearing-house functions the Text Archive will specially attend to the need for acquiring texts in the New Norwegian language form (ny-norsk), both newspaper material and novels.

SIXTH INTERNATIONAL ALLC SYMPOSIUM

The Sixth International Symposium of the Association for Literary and Linguistic Computing (ALLC) was held in Cambridge, 28 March - 3 April.

Knut Hofland, computer consultant at the Centre, reports that the Symposium was attended by some 150 participants from 20 countries. China was a newcomer this time. The organizing committee had taken great pains to recruit new speakers to supplement the group of well-established researchers in the field.

The topics included parsing, bibliography, lexicography, prosody, textual criticism, concordances, computers in education, data base technology, natural language processing, quantitative methods and text corpora.

Three Norwegian papers were read at the Symposium.
Interesting and stimulating as the Symposium was, the lack of pre-prints circulated in advance was, however, a flaw of the arrangement. With modern computing facilities at hand, it should be an easy matter to supply participants with such basic material for personal orientation and follow-up work.

SOMMERKURS I STATISTIKK FOR SPRÅK-OG LITTERATURFORSKERE

Roald Skarsten, computer-consultant at the Faculty of Arts, University of Bergen, gives a report of a national seminar held in Oslo, June 9 - 13 on the use of statistical methods in literary and linguistic research.

Attended by some 20 participants from the Universities of Oslo, Bergen and Trondheim, the seminar was arranged by the Norwegian

Computing Centre for the Humanities in cooperation with the consultant services of the Faculties of Arts at the universities.

Prof. Henning Spang-Hanssen of the Institute of Applied and Mathematical Linguistics, University of Copenhagen, was guest-lecturer. Discussions and group work on the application of statistical methods formed the major work of the seminar.

The seminar turned out to be a success, and plans will be made for future cooperative efforts in the field - the next time, perhaps, within a Nordic framework.

HUMANISTISKE DATA is published by the Norwegian Computing Centre for the Humanities. — The Editor is: *Jostein H. Hauge*, Director of the Centre. Issues are free. Contributions are welcome.

HUMANISTISKE DATA blir utgitt av NAVF.s EDB-senter for humanistisk forskning i Bergen. Senterets leder, *Jostein H. Hauge,* har det redaksjonelle ansvar for bladet. De som ønsker å få bladet tilsendt, kan bestille det ved henvendelse til senterets adresse: Villavei 10, Boks 53, 5014 Bergen-Universitetet.

Innlegg kan sendes til samme adresse.

Merk ny adresse: Harald Hårfagresgt. 31, boks 53, 5014 Bergen-Universitetet.