# humanistiske data

Meldingsblad for
NAVF's EDB-senter
for humanistisk forskning

*Norges almenvitenskapelige forskningsråd*

1977

INNHOLD:




HUMANISTISKE DATA blir utgitt av NAVF's EDB-senter for humanistisk forskning i Bergen.
Senterets leder, Jostein H. Hauge, har det redaksjonelle ansvar for meldingsbladet.
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.
Redaksjonen avsluttet 1.11.77


## *MEDARBEIDERE I DETTE NUMMER:*

*Nils Georg Brekke,* fylkeskonservator i Hordaland. Arbeider bl.a. med et prosjekt for EDB-katalogisering av kulturhistorisk materiale i distrikts- og bygdemuseer i Hordaland.

*Patricia Galloway,* systemprogrammerer ved Computer Unit. Westfield College, London.

*Ivar Fonnes,* førstekonsulent ved NAVF's EDB-senter for humanistisk forskning. Arbeidssted: Oslo.

*Knut Hofland,* Konsulent ved NAVF's EDB-senter for humanistisk forskning.

*Roald Skarsten,* førstekonsulent ved NAVF's EDB-senter for humanistisk forskning.


Sats og trykk:
Universitetets trykkeri,
Bergen

# REDAKTØRENS SPALTE

I forrige nummer ble det slått fast at virksomheten i Norge innenfor humanistisk databehandling øker fra år til år. Rapporter som kommer inn fra konferanser og seminar utenlands melder likeledes om aktiviteter over et bredt register av fagfelt og om upåklagelig entusiasme.

Siden vi ennå står midt oppe i den perioden da EDB ble introdusert i de humanistiske vitenskaper, er det for tidlig å skrive introduksjonsperiodens historie. På et senere tidspunkt vil det være lettere å beskrive de forhåpninger som en satte til bruk av databehandling i humanistisk forskning og gi et sant bilde av suksess og feilslag. Det vi nå kan gjøre er å håpe at EDB-bruken stadig blir bedre tilpasset de humanistiske fags egenart, faglige tradisjoner og metoder.

Økning i EDB-bruk i de humanistiske vitenskaper er slett ikke noe mål i seg selv. Datamaskinen må være et redskap som lar seg utnytte på humanistenes egne premisser. I senterets målsetting heter det at senteret skal arbeide for en *fornuftig* EDB-bruk i humaniora. De som gir råd i bruk av EDB har et klart ansvar for å forhindre EDB-bruk som ikke er arbeidssparende for forskeren, som ikke løser de konkrete problemer som er formulert eller som vil legge beslag på unødvendig store maskin- og driftsmidler. Her stanser imidlertid rådgiverens ansvar. Det er forskeren selv som må bedømme om datamaskinen kan brukes som et virkemiddel for fagets utvikling.

Derfor er det behov for kritisk å tenke gjennom og å videreformidle synspunkter på den plass datamaskinen bør ha i de humanistiske forskningsdisipliner. For å si det enkelt: Hvor passer datamaskinen, og hvor passer den ikke?

Gjestespalten denne gangen streifer dette temaet, og det ville være nyttig om flere ville bidra med innlegg i tilknytning til sine fagfelt.

Den berømte lingvist Noam Chomsky uttaler bl.a. i boken »Aspects of the Theory of Syntax»: »The social and behavioral sciences provide ample evidence that objectivity can be pursued with little concequent gain in insight and understanding». Hvis dette er riktig, bør vi også spørre: Kan datamaskinen føre til at en lignende tendens oppstår eller forsterkes i de humanistiske fag?

## Permanente EDB-organ for humaniora og samfunnsvitenskap under NAVF.

Etter forslag fra Rådet for humanistisk forskning og Rådet for samfunnsvitenskapelig forskning behandlet Styret for NAVF den 15.6. framlegg om permanente EDB-organer for humaniora og samfunnsvitenskap.

Styret vedtok å etablere NAVF's EDB-senter for humanistisk forskning og Norsk Samfunnsvitenskapelig Datatjeneste som permanente organer fra 1.1.1978. Ved disse institusjoner etableres følgende faste stillinger:

Ved NAVF's EDB-senter for humanistisk forskning:

1 stilling for administrativ leder,
2 førstekonsulent/konsulentstillinger,
1/2 driftsassistentstilling,
1/2 operatørstilling og
1 kontorfullmektigstilling

Ved Norsk Samfunnsvitenskapelig Datatjeneste:

1 stilling for administrativ leder,
1 konsulentstilling og
1 kontorfullmektigstilling

Styret delegerer den faglige og administrative myndighet til Rådet for humanistisk forskning og Rådet for samfunnsvitenskapelig forskning.

**Gjestespalten:**

# KULTURHISTORIE, GJENSTANDSFORSKNING OG EDB

*av Nils Georg Brekke*

Datateknikken har etablert seg i informasjonssystemet i samfunnet i eit slikt omfang at spørsmålet om personvern og kjeldevern byrjar å bli eit påtrengjande problem. Næringsliv og offentleg forvalting brukar store summar på datatenester, fordi dei representerer eit uvurderleg praktisk hjelpemiddel når det gjeld kritiske tidsrutinar og innsparing av personellressursar.

I forskningssamanheng har mange innan humanistiske fag vore reserverte mot å ta i bruk EDB. Rettnok har lingvistar, litteraturforskarar og historikarar alt lenge arbeidd systematisk med datahandsaming i analystisk samanheng, og store databasar for humanistisk primærmateriale er etablerte i Europa og USA, medan kulturhistorikarane her heime enno berre har sysla med planar om ein nasjonal museumskatalog på EDB eller eit landsomfattande register over gamle hus.

Er EDB eit aktuelt hjelpemiddel for gjenstandsforskarane? Vil datateknikken gjera det kulturhistoriske grunnlagsmaterialet meir tilgjengeleg, slik at vi kan frigjera forskningsressursar til andre oppgåver enn å leita i manuelle arkiv?

Det er reist mange og vektige motforestillingar når det gjeld verdien av EDB i humanistisk forskning. Det har kome til uttrykk otte for at datateknikken skal fremja »teknifiserte» problemstillingar, og at all tid skal drukna i endelaus koding, korrekturlesing og feilsøking. Ein ser for seg at det er tall-mennesket som nå rykker fram med ja/nei svar og statistikk, og at den gamaldagse, loslitne humanisten snart høyrer med til det historiske materialet.

Utan tvil har det vist seg å vera grunnlag for mange av desse motforestillingane, og ein bør respektera dei som førebels er noko avventande med å ta i bruk EDB. For oss som har det handicap å måtta klara oss gjennom livet utan særleg teknisk innsikt, kan det andsynes datateknikken liggja ein naturleg reservasjon mot det vi ikkje har skikkeleg greie på. EDB kan verka komplisert for ein lekmann på same måten som ein fullautomatisert telefonsentral (men vi dristar oss likevel til å bruka telefonen).

Ein del av reservasjonane kan hengja saman med litt overdrevne forestillingar om datateknikkens reelle funksjon i kommunikasjonsprosessen. Det har vore hevda at å ta i bruk EDB i humanistisk forskning er det same som å selja sjela si. Nå skal ein ikkje ukritisk gå inn på datateknikken berre fordi den »har framtida føre seg», men det kan vel vera turvande med ei viss »avmytologisering» av EDB-omgrepet.

La oss gå litt nærare inn på einskilde av motforestillingane:

1. EDB verkar kanaliserande på val av problemstillingar og vil gje preferanse for EDB-venleg materiale.
2. Koding av materialet i maskinleseleg form vil låsa fast datagrunnlaget til det vi i dag har oversyn over og interesse for, og kan medføra ei problematisk sementering av dagens kunnskapsnivå.
3. Vinninga går opp i spinninga.

For det første er det utan tvil rett at EDB kan verka inn ved valg av problemstillingar. I kulturhistorisk forskning er det også eit faktum at problemstillinga igjen kan verka inn på gjenstandsbeskrivelsen, der dei aspekt blir lagt serleg vekt på som er i fokus for forskaren.

Dette har to sider. På den eine sida stiller datateknikken eit enkelt og nyttig krav om å uttrykkja seg klart og eintydig. Dette reiser krav om eintydig terminologi og klare kategoriar, noko som vil provosera gjenstandsforskarane til eit sårt tiltrengt terminologisk oppryddingsarbeid som vi lenge har skuva framfor oss. I den mon EDB verkar fremjande på slike problemstillingar er det avgjort positivt.

Men behovet for standardisering av terminologi og rubrisering av opplysingar vil også kunna føra til at ein avskjer seg frå viktige opplysningar som ligg utanfor »skjemaet». I den utstrekning kravet til framtidig EDB-handsaming verkar som ei »tvangstrøye» i feltarbeidsfasen er dette nagativt, men her vil dei nye system for søking i fritekst og ustrukturerte tilleggsopplysningar vera eit avgjerande steg i retning av ein meir »åpen» bruk av EDB.

For det andre er det eit faktum at utvalet av kriterier i ein gjenstandsbeskrivelse i stor grad vil måtta vera subjektiv, avhengig av forskarens interessefelt, og seinare generasjonar vil kunne leggja vekt på andre aspekt ved materialet. Dette er serleg gyldig i eit fag der tolking av data er ei viktig oppgåve, men generelt vil dette vera eit problem som hefter ved all innsamling av grunnlagsmateriale. Einskilde vil derfor hevda at feltregistreringar berre har verdi der forskaren går ut med klare problemstillingar og avgrensar si datainnsamling til desse.

Problemet ligg her i at slike selektive registreringer av primærdata avgrensa til eitt tema vil føra til at eit potensielt forskningsmateriale kan gå tapt fordi informanten eller gjenstanden blir borte.

Eit materiale innsamla etter eit tilfeldig utval av kriterier og med eit avgrensa opplysningsspekter vil i alle høve gje grunnlag for å stilla visse spørsmål i framtida. Problemet med å låsa fast materialet til dagens kunnskapsnivå må seiast å vera mindre enn faren for at materialet går tapt.

I denne samanheng dreiar det seg først og fremst om å ta i bruk EDB til å få orden på »lageret». Her ligg verdien av eit datalager som gjev oversyn over museumsmateriale og feltmateriale, og ordner det topografisk, kronologisk eller systematisk til bruk i forsking, planlegging eller i pedagogisk samanheng.

Databasen som informasjonslager har to nivå: som reint lagringsmedium og som grunnlag for meir avansert analytisk handsaming av materialet. I kva grad EDB kan åpna nye vegar i analyse av eit materiale som ikkje går på den formale overflatestruktur, men på eventuelle underliggjande strukturar t.d. i eit ikonografisk eller eit stilhistorisk materiale, skal eg la liggja i denne samanheng.

For det tredje kan ein stilla spørsmålet om gjenstandsforskaren og musea har slike kritiske rutinar innanfor sitt arbeidsfelt eller slike rasjonaliseringsgevinstar å henta at dette kan forsvara kostnadene med å ta i bruk datateknikken i forvaltinga av det kulturhistoriske grunnlagsmaterialet.

Nytteverdien kan sjølvsagt ikkje målast på same måte som i næringslivet, men det er likevel klare grunnar som talar for:

1. Sikring av materialet.
2. Lettare informasjonssøking med samankopling til andre databaser vil gjera materialet meir tilgjengeleg og frigjera forskningsressursar.
3. Dette vil igjen åpna for betre utnytting av materialet til planlegging og pedagogiske formål.

Sikring av det historiske grunnlagsmaterialet i samfunnet gjennom regionale databasar er eit viktig punkt i kostnadssamanheng og treng ikkje nærare grunngjeving. Arkiv- og katalogsamarbeid mellom museer og forskningsinstitusjoner i ein regional eller i landssamanheng vil kunne gje ei heilt anna utnytting av materialet og spara inn ein vesentleg del av den tida som går med til å leita i manuelle arkiv. Forskningsprosessen bør i framtida kunne verta frigjort frå mykje rutinemessig arkivarbeid.

Etterkvart som stadig fleire typar grunnlagsmateriale blir registrert for datahandsaming vil den gjensidige nytteverdi auka. Det vil til dømes vera av stor interesse for kulturhistorikarane at historikarane registrerer folketeljingane på EDB, og det vil i framtida på ein enkel måte kunna gjennomførast kronologisk eller topografisk kryssøking mellom samankopla register over gamla hus, gamle fotografi og gjenstandskatalogar ved musea.

*overgang til side 13*

# EDB-SENTERETS LANGTIDSPLAN 1978-81.

Våren 1977 utarbeidet styret for NAVF's EDB-senter en ny langtidsplan som skal dekke perioden 1978-81. Langtidsplanen ble godkjent av Rådet for humanistisk forskning i mai.

Langtidsplanen bygger på senterets egen driftserfaring og på en rekke interne utredningsdokumenter og rapporter som er produsert i de siste år. Dessuten har Rådet for humanistisk forskning i den siste tiden gitt veiledende vedtak om senterets fremtidige funksjon.

**Et nasjonalt EDB-senter.**

I langtidsplanen blir det understreket at EDB-senteret i den kommende 4-års perioden vil bli et klarere markert nasjonalt senter ettersom universitetene vil opprette egne stillinger som EDB-konsulenter for de humanistiske fag. Senterets nasjonale karakter fremtrer bl.a. ved at senteret i første rekke søker å dekke nasjonale fellesbehov og løser oppgaver som er for store til at en enkelt institusjon kan makte dem. På den annen side må senteret også fortsatt samarbeide med lokale tiltak dersom tiltakene har en nasjonal interesse, eksempelvis i forbindelse med metodeutvikling eller programutvikling av generell karakter. I tiden fremover bør senteret også påta seg et større ansvar for humanistisk FOU-arbeid utenfor universitetene.

**Senterets samarbeidspartnere.**

I langtidsplanen blir det gitt en presentasjon og karakteristikk av senterets samarbeidspartnere. Disse er omtalt som:

- arkivinstitusjoner med vitenskapelig personale
- vitenskapelige museer
- regionale høgskoler
- de historisk-filosofiske fakulteter og fagavdelinger ved universitetene
- EDB-konsulentene for de humanistiske fag
- øvrige vitenskapelige eller kulturelle institusjoner (Kulturråd, Språkråd, fylkeskonservatorer, lokalmuseer etc.).

**Innsatsområder.**

Senterets virksomhet vil i den kommende 4-årsperioden foregå innenfor følgende områder:

- konsulentassistanse
- prosjektassistanse
- metodisk utrednings- og forsøksarbeid
- generell programutvikling
- produksjon og vedlikehold av og service på sentralt humanistisk kildemateriale (fellesdata)
- opplærings- og informasjonsvirksomhet (kurs, stipendprogram og hospitantopplegg, nasjonale og regionale kurs og seminar, dokumentasjonstjenester).

**Prioriterte målområder.**

Ifølge langtidsplanen vil følgende områder bli prioritert:

a) **Arbeid med fellesdata:**

Det er et klart behov for å etablere en regulær dokumentasjonstjeneste for EDB-orientert humanistisk forskning og å ta hånd om og formidle data som tilrettelegges i EDB-prosjekter i vårt land. Likeledes bør det arbeides for å legge forholdene bedre til rette for dataoverføring av sentralt humanistisk forskningsmateriale og utvikle de nødvendige servicetjenester.

**b) Informasjon.**

Det er behov for å styrke informasjonsarbeidet ikke minst for å kunne samle inn relevant informasjon nasjonalt og internasjonalt om prosjekter, data og program.

**c) Konsulenttjeneste i Tromsø.**

Senteret vil sammen med Universitetet i Tromsø arbeide for å etablere en regulær EDB-tjeneste for humanister etter mønster av ordningen ved de øvrige universiteter.

d) **Samarbeid med institusjoner utenfor universitetene.**

Etableringen av lokale EDB-tjenester vil frigjøre kapasitet ved senteret til å styrke kontakten med institusjoner utenfor universitetene.

**Utbygging av senterets stab. Driftskostnader.**

I langtidsplanen blir det foreslått en stillingsutbygging for å oppfylle det arbeidsprogram som er beskrevet. I langtidsplanen blir det også gjort greie for den betydelige

støtte som Universitetet i Bergen gir til driften av EDB-senteret.

Ved sin behandling av langtidsplanen uttalte Rådet for humanistisk forskning at spørsmålet om utvidelse av staben vil bli tatt opp i forbindelse med den årlige budsjettbehandlingen.

Rådet for humanistisk forskning har videre vedtatt en ny målsettingsparagraf for senteret. I den nye målsettingsparagrafen kommer det klarere fram at senteret

- er et nasjonalt organ med ansvar for humanistiske fagmiljøer både utenfor og innenfor universitetene
- skal legge vekt på å dekke fellesbehov
- skal videreutvikle EDB-miljøer

Ved sin omtale av arbeidsoppgavene understreket Rådet for humanistisk forskning bl.a. samarbeidet med universitetenes EDB-konsulenter og senterets medansvar for at kildemateriale som er tilrettelagt for datamaskinell behandling kommer sekundærbrukere til nytte.

**Rettelse til D. Wickmanns artikkel i forrige nummer (Humanistiske Data nr. 1/2 1976).**

I artikkelen: »Statistics in the Humanities. Some Epistemological Remarks» er det på side 17, sp. 1, siste avsnitt kommet inn en meningsforstyrrende feil. Vi gjengir derfor hele avsnittet i rettet stand:

»Now, some remarks on the limit of improbability beyond which we are no longer willing to accept the null-hypothesis. The limit depends, of course, on the risks we are ready to undergo. There are two risks, as we have seen, namely to commit the error of line (2) and that of line (3). These risks are measured in probabilities too, the probability $\alpha$ to commit the error of the first type and the probability $\beta$ to commit the error of the second type. The examinator expresses his own urge for scrutiny by means of $\alpha$ and $\beta$. But, worse luck, there is a nasty relation between $\alpha$ and $\beta$: the one is to the detriment of the other. The smaller the risk $\alpha$ to commit the error of the first type (this is to reject a valid null-hypothesis) the greater the risk $\beta$ to commit the error of the second type (this is to accept an invalid null-hypothesis). It is like cruising between Scylla and Charybdis. This is the dilemma of each decision-making. By means of statistics, nevertheless, it is possible to make decisions depending on the values of $\alpha$ and $\beta$ prefixed by the examinator at the beginning of an investigation. An examinator's readiness for risk or, as we put it before, an examinator's urge for scrutiny, is »condensed» in these two probabilities, all the rest follows by deductive steps. This, indeed, may be seen as a great progress towards inter-subjective testability.»

## The fifth International Symposium on Computing in Literary and Linguistic Research.

Det er kommet melding om at The Fifth International Symposium on Computing in Literary and Linguistic Research vil bli holdt ved The University of Aston in Birmingham, England fra 3. – 7. 4. 1978. Den utsendte folder gir følgende stikkord for temavalget:

- Authorship Studies
- Concordances
- Classical Studies
- Education
- Input/output
- Language-oriented Groups (English, French, Dutch. . .)
- Lexicography
- Literary Statistics
- Oriental Studies
- Software
- Stylistic Analysis
- Syntactic Analysis
- Text Editing

Kontakt:

The Secretary (CLLR),
Modern Languages Department,
The University of Aston in Birmingham,
Birmingham B4 7ET,
England.

som også gir veiledning i forbindelse med innsending av foredrag.

# Senterets arbeid våren 1977

Våren 1977 har vært viktig for senteret idet fundamentale spørsmål om senterets fremtid er blitt avklart.

De vedtakene som er truffet, (se egen melding) betyr bl.a. at senteret kan legge langsiktige planer for sin virksomhet. Ettter hvert som ordningen med universitetsansatte EDB-konsulenter for de humanistiske fag utbygges, vil senteret i større utstrekning enn før kunne konsentrere seg om nasjonale oppgaver. For at senteret skal kunne arbeide effektivt med de høyest prioriterte oppgaver i fagmiljøene, vil det imidlertid fortsatt være nødvendig med et nært samarbeid med universitetenes EDB-konsulenter.

Våren 1977 har bl.a. følgende oppgaver stått sentralt:

**I Programutvikling.**

Arbeidet med videreutvikling av programsystemet NOVA-STATUS har fortsatt samtidig som det er demonstrert og lagt til grunn for flere faglige prosjekter. I vinter ble det etablert en nasjonal interessegruppe omkring dette tekstsøkesystemet med representanter både fra universitetene og statsforvaltningen. Gruppen drøfter aktuelle oppgaver vedrørende videreutviklingen av systemet og prioriterer og fordeler dem. For tiden yter R-direktoratet økonomisk støtte til videreutviklingen. NAVF's EDB-senter har valgt å satse på oppgaver som vil gjøre behandlingen av resultatdata fra søking mer fleksibel og øke anvendbarheten av systemet i arkivsammenheng. I sommer har flere assistenter arbeidet med slike oppgaver.

Senteret implementerte NOVA-STATUS på universitetsanlegget i Trondheim i januar. Fra i vinter av har systemet også vært tilgjengelig på DEC-10, Universitetet i Oslo.

**Syntaktisk analyse.**

Programsystemet for syntaktisk analyse (se HD nr. 1—2, 1976, s. 9) ble i juni implementert på DEC-10 og siden videreutviklet via oppringt terminalsamband Bergen-Oslo.

**Lemmatisering.**

Det er utviklet programmer for å knytte bøyningsformer i tekster sammen med en grunnform. En første versjon som bygger på en grammatisk tabell for bøyning av ord i moderne bokmål, er implementert. I neste

fase vil en trekke inn ytterligere språklig informasjon ved bestemmelse av ordene, bl.a. statistisk informasjon om endingstyper, og utnytte resultater fra tidligere utførte og kontrollerte lemmabestemmelser.

Det er ført forhandlinger med Prosjekt for datamaskinell språkbehandling, Universitetet i Bergen om å få adgang til de språklige grunndata som finnes der.

**Søking i store tekstsamlinger.**

På grunn av mange forespørsler om fremsøking av bestemte ord og ordkombinasjoner fra The Brown Corpus er senteret i gang med å utvikle programmer for å korte ned søketiden ved å lagre hele tekstmaterialet i et bestemt konkordansformat. Metoden vil også bli brukt ved tilrettelegging av det moderne engelske tekstkorpuset CAMET som senteret fullfører i samarbeid med Britisk institutt, Universitetet i Oslo.

**Resultatdata i mikroformer.**

Ettersom humanistisk EDB-arbeid ofte medfører store samlinger av resultat- eller lagringsdata, er det aktuelt å ta i bruk mer kompakte utskriftsformer enn vanlig utskrift på papir.

I samarbeid med Bergen Datasenter har senteret i vår eksperimentert med overføring av tekst- og arkivdata til microfiche. På ett microfichekort er det plass til 208 A4 sider. Ved hjelp av indekser o.l. er det enkelt på et leseapparat å søke fram den relevante informasjon.

**II Prosjektassistanse.**

Senteret har som tidligere gitt prosjektstøtte til en rekke prosjekter i Oslo, Bergen, Trondheim og Tromsø. Det vil i neste nummer bli gitt en oversikt over de EDB-prosjekt som for tiden drives i vårt land, som en oppfølging av en oversikt gitt i Humanistiske Data for 3 år siden.

Noen av de største samarbeidsprosjekter nevnes:

Det omfattende redaksjonelle arbeid med *Norsk Landbruksordbok* ved Norsk leksikografisk institutt (se forrige nummer) er sluttført i løpet av våren. Parallelt med dette har det vært under oppbygging et trykkeklart magnetbånd for fotosetting. Dette arbeidet vil bli ferdig i høst. Vår EDB-konsulent i Oslo, Ivar Fonnes, har hatt ansvaret for EDB-arbeidet.

EDB-oppgavene i forbindelse med prøveprosjektet *»EDB og manuskriptregistraturer»* ble sluttført i januar. Det er utarbeidet en rapport om forsøksarbeidet hvor også planene om et EDB-basert sentralregister for privatarkivalia blir lansert. I forbindelse med prøveprosjektet er det utarbeidet en brevregistrant til privatbrev i Riksarkivet. Registeret vil foreligge i løpet av høsten på mikroformat.

For *De kulturhistoriske registreringer på Vestlandet* v/Historisk museum, er det produsert en serie kataloger for ulike lokalmuseer. Katalogene vil bli brukt til lokalt museumsarbeid og vil kunne danne grunnlag for en videre drøfting av et EDB-basert sentralregister for museumsmateriale og samarbeidsformer mellom et sentralregister og de lokale brukerne.

**Informasjons- og opplæringstiltak.**

I tillegg til de ordinære brukerkurs som regelmessig gis ved universitetene, har NAVF's EDB-senter etablert et 4 ukers stipendprogram for humanister som ønsker å sette seg inn i bruken av databehandling. Våren 1977 har 5 stipendiater oppholdt seg ved senteret og studert EDB i relasjon til språkforskning, litteraturforskning, klassisk filologi og anvendt språkvitenskap. Stipendiatene kom fra Oslo, Trondheim, Stavanger og Tromsø. I tillegg til å gi en individualisert undervisning utfører senteret dataregistrering for stipendiatene og samarbeider med dem om løsning av programmeringsoppgaver i tilknytning til deres eget forskningsarbeid (se melding om stipendiatopplegget i forrige nummer).

# Nasjonal konferanse om humanistisk databehandling Gol 4. og 5. november 1976.

(På grunn av en redaksjonell feil kom ikke denne meldingen med i nr. 1—2 1976).

NAVF's EDB-senter arrangerte 4. og 5. november 1976 den første nasjonale konferanse om status og fremtidige arbeidsoppgaver innenfor humanistisk databehandling. På konferansen, som ble holdt på Gol, deltok 47 representanter fra humanistiske fagmiljøer i Bergen, Oslo, Trondheim, Tromsø og Stavanger, samt representanter fra Rådet for humanistisk forskning og NAVF's EDB-senter. Det var også invitert representanter fra miljøer utenfor humaniora.Hovedarbeidet på konferansen ble utført på gruppenivå hvor det var satt opp temaer som: Datamaskinell språkanalyse. Tekstsøking. Språkstatistikk. Datamaskinell behandling av primærkilder i historisk forskning. EDB-opplegg for arkiv- og katalogdata.

I plenum ble det gitt oversyn over virksomheten i de humanistiske fag, en orientering om Norsk Samfunnsvitenskapelig Datatjeneste og om planene for en registreringssentral for historiske data i Troms (se egen melding i dette nummer).

De faglige plenumsforedrag var om »*Informasjonsvitenskapelige metoder som et* verktøy for humanister» v/professor Svein Nordbotten og »Metoder for trykking på grunnlag av data i maskinleselig form» v/ universitetsbibliotekar Hans Martin Fagerli.

I gruppene ble det holdt flere faglige innledninger som grunnlag for drøftingene. Her ble en rekke enkeltprosjekter behandlet (data, metoder og program), samtidig som behovet for fremtidige fellestiltak ble kartlagt.

I en avsluttende plenumssesjon ble gruppearbeidet referert og kommentert.

En del av konferansen var viet den plass NAVF's EDB-senter har i EDB-arbeidet i de humanistiske fag og hvilke tiltak senteret kan sette i verk for å dekke viktige behov i brukermiljøene.

Av synspunkter som kom fram, kan nevnes:

**Konferansevirksomhet.**

Det var allmenn enighet om behovet for ulike typer nasjonale sammenkomster, både tverrfaglige og konferanser med større faglig konsentrasjon. Av aktuelle temakonferanser ble nevnt konferanser i bruk av kvantitative metoder, datamaskinelle metoder ved studiet av syntaks og morfologi, EDB i gjenstandsforskning og bruk av datamaskin ved arkiv- og katalogopplegg.

**Informasjon.**

Ønske om mer informasjon om pågående arbeid i vårt land og i utlandet var fremtredende. Ved dette håpet en både

å hindre unødig dublerende prosjektarbeid og å skape relevante faglige kontakter.

Det kom klart fram på konferansen gjennom deltaking av representanter fra jus og samfunnsvitenskap at det også er behov for å bli orientert om det arbeid som pågår på humanistenes tilgrensende fagområder.

**Data.**

Behovet for større muligheter for forskningsmiljøene til å tilrettelegge data for maskinell behandling, ble ofte nevnt. Likeledes ble det understreket behovet for å ha en EDB-tjeneste som kunne samle, oppbevare og vedlikeholde viktig humanistisk forskningsmateriale.

Mulighetene for og verdien av å standardisere dataformater ble også berørt.

**Kurs- og opplæringsvirksomhet.**

Verdien av lokale og nasjonale kurs og seminar ble understreket. Variasjon i lengde, innhold og nivå burde tilstrebes. Opplæringsprogrammer som korttidsstipend for humanister ble positivt vurdert.

**Program og metode.**

Det ble også understreket viktigheten av å drive generell programutvikling innenfor de humanistiske fag og å utføre metodisk forsøksarbeid.

Allerede utarbeidede programsystemer som primært er tiltenkt andre felt, bør evalueres og eventuelt tilpasses bruken i humanistisk forskning.

Et viktig tema som det dessverre bare var anledning til å streife på konferansen, var spørsmålet om *den plass EDB bør få i humanistisk forskning.*

Videre var det interessant å få påpekt at flere EDB-brukere hadde registrert at EDB-metodikken hadde en direkte virkning på det faglige arbeid, slik f.eks. på arkivsektoren hvor sentrale, faglige spørsmål som enhetlig terminologi, dataformatering og datastrukturering aktualiseres. Det ble bl.a. hevdet at bruk av EDB kunne gi støtet til en høyst nødvendig systemanalyse av det faglige arbeid.

**Oppsummering.**

Konferansen viste at EDB-aktivitetene i dag har fått anseelig bredde, at entusiasmen er stor, og at også den nødvendige vurderende refleksjon over bruken av de nye redskaper er til stede.

**Registreringssentral for historiske data.**

Ved Universitetet i Tromsø har det i løpet av det siste året vært arbeidet med planer om en registreringssentral i Troms for historiske data.

Planene har sitt forbilde i dataregistreringssentralen »Demografisk databas» som er etablert i Umeå — Haparanda, hvor det tilrettelegges nominative kilder, dvs. kilder med individet som enhet.

På samme måte som i Sverige tenkes registreringssentralen drevet i økonomisk samarbeid med stat/fylke som nå vurderer prosjektet som sysselsettingstiltak. For 1978 er planen å starte et forsøksarbeid med tanke på full drift i 1979.

**Konsulenttjeneste for humanister i Tromsø.**

Rådet for humanistisk forskning har gjennom sin budsjett-tildeling til NAVF's EDB-senter avsatt midler tilsvarende 4 måneders konsulentlønn til EDB-assistanse ved Universitetet i Tromsø i 1977. Midlene, som blir disponert i samarbeid med Institutt for språk og litteratur og Institutt for samfunnsvitenskap, vil bli brukt som tilskudd til engasjement av datasekretær, konsulentopphold av senterets konsulenter og til timelønnet programmeringsassistanse.

# Pågående arbeid med tekstarkiv for engelsk språk og litteratur

*1. A Computer Archive of Modern English Texts (CAMET).*

Som nevnt i forrige nummer har senteret sammen med Britisk institutt, Universitetet i Oslo, innledet et samarbeid om CAMET. CAMET består av moderne engelsk tekstmateriale på 1 million ord og er bygget opp på samme måte som The Brown Corpus. Det ble i februar inngått en avtale med professor Geoffrey Leech, University of Lancaster om sluttføring av prosjektet.

Ifølge avtalen skal Britisk institutt ta på seg arbeidet med å løse copyright-problemene og foreta tekstkontroll. NAVF's EDB-senter står for EDB-driften.

Pr. 1. 8. 1977 er status slik: Hele korpuset fra Lancaster er lagret i datamaskinen ved Universitetet i Bergen. Det er foretatt en kodekonvertering for å gjøre tekstene lettere å bruke ved korrekturarbeidet. Arbeidet med tekstkontroll starter i høst. NAVF har bevilget midler til dette arbeidet.

Planen er å gjøre tekstkorpuset bruksklart i løpet av 1978. Alle henvendelser om tiltaket rettes til dosent Stig Johansson, Britisk institutt, Universitetet i Oslo.

*2. The Brown University Corpus.*

Etter de siste opplysninger fra professor Nelson Francis er arbeidet med grammatisk koding av hele korpuset nå avsluttet ved The Brown University. Ifølge professor Francis pågår nå »the laborious and boring but necessary work of checking the results». Når dette er gjort, vil det bli utarbeidet en lemmatisert frekvensliste samtidig som arbeidet med å utvikle teknikker for automatisk, grammatisk analyse starter.

Det kan nevnes at NAVF's EDB-senter i samarbeid med professor Francis har justert det typografiske oppsettet av tekstene i The Brown Corpus og tilrettelagt en versjon med store og små bokstaver. Med disse forandringer vil korpuset helt samsvare med CAMET – noe som er svært viktig for fremtidig, komparativ bruk av tekstsamlingene.

*3. Survey of Spoken English.*

Under ledelse av professor Jan Svartvik ved Lunds Universitet pågår det arbeid med å overføre til datamaskinleselig form et omfattende talespråkmateriale hentet fra Survey of English Usage, University College, London.

*4. International Computer Archive of Modern English (ICAME).*

Dette er en interesseorganisasjon som ble dannet i Oslo i februar 1977. Initiativtakere var:

Professor W. Nelson Francis, Brown University, USA.

Professor Geoffrey Leech, University of Lancaster, England.

*Dosent Stig Johansson, Universitetet i Oslo.*

Professor Arthur Sandved, Universitetet i Oslo.

Professor Jan Svartvik, Lunds Universitet.

NAVF's EDB-senter deltok ved de forberedende drøftinger.

Formålet med organisasjonen er:
»1. collecting and distributing information on English language material available for computer processing,
2. collecting and distributing information on linguistic research completed or in progress on the material,
3. compiling an archive of corpuses to be located at the University of Bergen, from where copies of the material could be obtained at cost.»

De tre tekstsamlingene som er nevnt foran, vil i første omgang danne basis for tiltaket.

I forbindelse med kunngjøringen om opprettelsen av organisasjonen til en rekke universiteter i Europa og USA er det kommet fram at flere universiteter har etablert, eller har planer om å etablere, tekstarkiv for litterær eller språklig analyse. Nedenfor nevnes noen slike tiltak:

5. *Oxford Archive of English Literature.*

Det kan nevnes at The Computer Laboratory, University of Oxford samordner arbeidet i England med å etablere et dataarkiv over engelsk litteratur. Det er pr. mai 1977 allerede samlet tekstkopier av en rekke verk i maskinleselig form. Nærmere opplysninger om arkivbestanden kan fås ved NAVF's EDB-senter.

6. *A Computer Archive of Language Materials (CALM).*

Ved Standford University, California, har det i de siste 6 år vært arbeidet med et prosjekt kalt »A Computer Archive of Language Materials» (CALM). Ifølge en orientering om tiltaket inneholder databanken for tiden:

»cross-linguistic typological files for phonetics and phonology: Lexicographic files representing one contemporary and several diachronic dictionaries, and a million-word corpus of present-day American English.» CALM vil gi databehandlingsservice til forskere som ønsker å stille spørsmål til materialet i arkivet og distribuerer resultatdata på microficheformat.

Nærmere opplysninger om CALM fås ved NAVF's EDB-senter.

7. *University of California, St. Diego.*

Også her arbeides det med planer om »An Archive of Computer readable texts in modern languages».

*En nærliggende konklusjon på* et slikt oversyn er at interessen synes å være stor for å etablere omfattende tekstarkiv og at tiden er inne til å koordinere bestrebelsene slik at de enkelte nasjonale bidrag kan utfylle hverandre.

---

*overgang fra side 5*

*gjestespalten:*

## *KULTURHISTORIE, GJENSTANDSFORSKNING OG EDB*

Kostnaden må også sjåast i samanheng med at eit informasjonslager i datamaskinen vil gje ei breiare utnytting av det kulturhistoriske grunnlagsmaterialet enn det som tidlegare har vore mogleg, jamvel om vi også vil kunna visa til ressursinnsparing når det gjeld søking i manuelle arkiv. I næringslivet vil det ofte vera storleiken på omsetjinga og rasjonalisering av rutinar som er avgjerande for vurdering av kostnad og nytteverdi. I forvalting av historisk grunnlagsmateriale vil kvalitative vurderingar vera like vesentlege som kvantitative: sikring av fullverdig dokumentasjon.

Det er vår oppgåve å sikra innhaldet i denne dokumentasjonen. Men *teknikken* overlet eg gjerne til teknikarane. Dei kjenner den betre enn oss.

## En EDB-tjeneste for de humanistiske fag ved Norges Lærerhøgskole.

I likhet med Universitetet i Oslo har Universitetet i Trondheim fra 1.7.77 etablert en EDB-tjeneste for de humanistiske fag (se også egen melding om Bergen). Som det første universitet i landet, overtar Universitetet i Trondheim fra samme tid hele det økonomiske ansvar for en EDB-konsulentstilling for de humanistiske fag. Tidligere har NAVF dekket halvparten av utgiftene til stillingen.

Det er nedsatt et styre for EDB-tjenesten sammensatt av representanter fra Avdeling for filologiske fag og Avdeling for samfunnsfag. EDB-konsulent Eirik Lien er sekretær for styret. Styret har fra 1.7.1977 til 30.6.1980 følgende sammensetning:

Professor Knut Fintoft (formann)

Amanuensis Jan Ragnar Hagland

Amanuensis Torstein Strømsøe

Det vil om kort tid bli ført forhandlinger mellom Lærerhøgskolen og NAVF's EDB-senter om en samarbeidsordning mellom institusjonene.

## Maskinlesbare tekstsamlinger for tysk språk og litteratur.

I fagtidsskriftet »Deutsche Sprache», 4/1976 er det artikkel om »Machinenlesbare Textcorpora des Deutschen und des Englischen» av Burkhard Schaeder.

I artikkelen blir det redegjort for spørsmålet om den vitenskapelige bruksverdi av tekstkorpora og hvilke kriterier man bør nytte ved utvalg av tekster til et korpus. Videre blir 10 tyske og 2 engelske tekstkorpora dokumentert.

I »Deutsche Sprache», 4/1976 blir det også gitt en bibliografi over litteratur i forbindelse med de ulike tekstkorpora og en adresseliste.

Det kan for øvrig opplyses om at forfatteren sammen med Henning Bergenholtz i 1978 vil utgi en artikkelsamling om »Text-Corpora-Materialen für eine empirische Sprach- und Literaturwissenschaft».

**Tabellarische Übersicht über Umfang und Aufbau maschinenlesbarer Textkorpora des Deutschen und des Englischen (Stand 1975)**

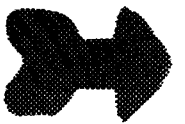
| | | (1) Sprache | (2) nationale Varietät | (3) mediale Varietät | (4) diastratische Varietät | (5) Genre | (6) Publikationsform | (7) Textstoff | (8) zeitliche Herkunft der Texte | (9) Anzahl der Texte | (10) Anzahl der Wortstellen pro Text | (11) Anzahl der Wortstellen des Korpus |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Aachener Textkorpus: Prosatexte | Deutsch | BRD | geschriebene Texte | Standardsprache | Romane, Erzählungen, Sachbücher | Buch | fiktional, nichtfiktional | ? | 42 | 32.000 – 295.000 | 4 Mio |
| 2 | Aachener Textkorpus: Studentenlyrik | Deutsch | BRD | geschriebene Texte | Standard- und Literatursprache | Gedichte | Buch | fiktional | 1822–1962 | 4.000 | 12–480 | 520.000 |
| 3 | Bonner Zeitungskorpus | Deutsch | BRD DDR | geschriebene Texte | Standardsprache | Zeitungstexte | Zeitung | nichtfiktional | (1949) 1954 1964 1969 (1974) | 6.832 | 3 Zeilen – 999 Zeilen | 1,7 Mio (3,16 Mio) |
| 4 | Brown Corpus | Englisch | USA | geschriebene Texte | Standardsprache | Romane, Erzählungen, Sachbücher, Reportagen | Buch, Zeitung, Zeitschrift | fiktional, nichtfiktional (15 Sachgebiete) | 1961 | 500 | 2.000 | 1 Mio |
| 5 | Bungarten-Korpus | Deutsch | BRD DDR Schweiz Österreich | geschriebene Texte | Standardsprache | Romane, Erzählungen, Sachbücher, Reportagen | Buch, Zeitschrift, Zeitung | fiktional, nichtfiktional (19 Sachgebiete) | 1954–1970 | 49 | 8.000 – 222.000 | 2,548 Mio |
| 6 | Freiburger Korpus | Deutsch | BRD | gesprochene Texte | Standardsprache | Dialoge Monologe | Vortrag, Gespräch | nichtfiktional (9 Sachgebiete) | 1967–1974 | 222 | 175 – 16.360 | 600.000 |
| 7 | Lancaster Corpus | Englisch | GB | geschriebene Texte | Standardsprache | Romane, Erzählungen, Sachbücher, Reportagen | Buch, Zeitschrift, Zeitung | fiktional, nichtfiktional (15 Sachgebiete) | 1961 | 500 | 2.000 | 1 Mio |
| 8 | LIMAS-Korpus | Deutsch | BRD | geschriebene Texte | Standardsprache | Sachbücher, Erzählungen, Berichte, Romane | Buch, Zeitschrift, Zeitung | fiktional, nichtfiktional (34 Sachgebiete) | 1970 | 500 | 2.000 | 1 Mio |
| 9 | LIMAS-Kfz-Korpus | Deutsch | BRD | geschriebene Texte | Fachsprache | Lehrbücher, Betriebsanweisungen, Berichte | Buch, Zeitschrift | nichtfiktional (Sachgebiet Kfzwesen) | 1956–1974 | 90 (200) | 600 – 77.500 | 750.000 (1 Mio) |
| 10 | Lunder Zeitungs-Korpus | Deutsch | BRD | geschriebene Texte | Standardsprache | Zeitungstexte | Zeitung | nichtfiktional (5 Sachgebiete) | 1966–1967 | 6.595 | ? | 3 Mio |
| 11 | Mannheimer Korpus | Deutsch | BRD | geschriebene Texte | Standard- und Literatursprache | Romane, Erzählungen, Berichte | Buch, Zeitschrift, Zeitung | fiktional, nichtfiktional (10 Sachgebiete) | 1946–1967 | 28 | 7.800 – 144.000 | 1,6 Mio (2,5 Mio) |
| 12 | Saarbrücker Korpus | Deutsch | BRD | geschriebene Texte | Standard- und Fachsprache | Fachbücher, Zeitungstexte | Buch, Zeitung | nichtfiktional | 1955–1957 1961–1963 | 11.000 Sätze aus 45 Buch- u. 1.340 Zeitungstexten | ? | 200.000 |

PATRICIA GALLOWAY:

# *An introduction to archaeological computing*

## SOME PROBLEMS AND METHODS

*This paper is intended to give a rapid survey, in layman's terms, of the major computer applications in archaeology. It has no pretentions to completeness, as it concentrates upon examples taken from the author's own experience. A short reading list directs the interested reader to further more technical works.*



*Information retrieval.*

Much has been said in recent years about the nearmiraculous prospects which could be offered to archaeology by the use of huge centralised computer data banks. That there will be advantages to be looked for in the future from such proposals is certain, but the need for coordinated study of standardised recording techniques and the expense of dealing with the enormous mass of existing material will make the national archaeological data bank a dream for some years yet. An organised approach to such a project will, however, rest upon the experience now being gained through the use of information retrieval systems on a smaller scale in archaeology.

For a medium-sized excavation, which will have more than, say, 5000 items to record and retrieve, computerised information retrieval is not only practicable, but also of immense economic importance for the acceleration of post-excavation analysis that it can make possible. Actual field recording using computer terminals has already been put into practice on some sites in America, where it has proved possible to give the excavator a daily overview of the work and thus to aid in the actual decision-making of the excavation.

An ideal system being discussed at present would make available to the recording staff an intelligent terminal consisting of a typewriter keyboard, television screen, and some sort of printing device, to be used as follows: the terminal is programmed to present a questionnaire or check-list, in the form of questions projected onto the screen; these questions are answered by typing the answers on the keyboard. All or part of the answered checklist can then be printed out for use as a label to be enclosed with the finds or attached to drawings or photographs.

Given a system of this kind, it is obvious that a number of decisions must be made in order to tailor the recording system thus implemented to the needs of the excavation, and that these decisions must be made in a more systematic way than is usually the case. The objectives of the recording system——the uses to which the records will be put——must be specified. Since such a system can dispense with numerical coding, some sort of attention must be given to the careful draughting of a questionnaire form which can adapt itself to the needs of each of the things which is to be recorded. It should be remembered that full formalisation of description is not necessary, since provision can be made for free comment by the recording staff following the completion of the checklist. But some clear structure must be imposed upon the recording system, as this structure will determine the programming of the inter-active machine-human system.

There are innumerable uses to which such a system can be put, once the information has been recorded. The material can be searched for the coordinates of certain classes of artifacts at specific stratigraphic levels, and distribution maps can be drawn automatically, this time in connection with a larger central computer. All conceivable sortings of the material can be performed, nearly as fast as they can be imagined, and catalogues representing these sortings can be generated. A program can be written which will 'translate' the information which has been recorded into the form required by museum records, and all the information can thus be transferred without the need for rerecording. Taking advantage of the 'free comment' feature suggested above, experts working on specialist reports on the material may add at any time to the record and have their additions included in the master files. The system can of course be protected from unauthorised access.

**II. General statistics.**

It is not our purpose here to discuss statistics *per se* or the enormous problems of sampling which seem to make the use of classical statistical inference rather doubtful for archaeological applications. In any case, only the largest research projects of this kind would call for the use of a computer. It is in the newer field of multivariate statistics that more promise has been found for archaeological work, and it is just in this field that the use of computers, due to the

complexities of the problems involved, is almost mandatory. The advantage to the archaeologist in the use of these methods is that by and large they leave the tasks of inference to him. In addition, archaeological problems being by nature multivariate, they enable the archaeologist to treat the complexities of his subject with sufficiently complex methods.

The basic concepts of multivariate analysis depend upon envisaging archaeological entities and the relations between them as points in a space which may have as many dimensions as there are attributes to be compared among the entities being studied. Within the multidimensional space,the similarities or dissimilarities which relate these entities are to be conceptualised as physical distances, and the shapes which are suggested by such configurations may be interpreted directly by the archaeologist in terms of the archaeological concepts relating to the entities in question. Clusters of points representing artifacts may group in a manner that suggests typology; the graves from a cemetery or the pottery retrieved from a pit may be represented by points which tend to string themselves out along a single direction, indicating perhaps a time-related development. Thus, in search of such indications, the points in the space are defined in terms of axes and a Cartesian grid system. The confusion of dimensions above the number of four may be reduced by means of dimension-reducing programs, which tend to consolidate the correlations between attributes of several dimensions: for example, decorative elements and shape in metalwork may show similar trends ofdevelopment in terms of similarly structured shapes of points-in-space in several dimensions, and these shapes are consolidated by superimposition and reduction to a mean.

A number of these multivariate methods have already been used in archaeology. A few of them are: principal components analysis, factor analysis, rotational fitting, discriminant analysis, nonmetric multidimensional scaling, and cluster analysis. Further discussion of these methods and their uses will be found in some of the references for further reading. The latter two methods will be discussed in some detail in order to give some notion of their actual use on two of the main 'archaeological problems': seriation and typology.

## III. Seriation.

Seriation, or the placing of similar archaeological entities in a chronological sequence, is a problem which has involved the use of computers since the 1950's. It is based on faith in the theory that stylistic changes are related in some way to chronological movement. If the theory has a basis in fact, it should be possible to arrange a group of artifacts or of graves so that a smooth change is seen through the series (barring conquest, etc.), and this change will be indicative of chronology. We will treat this problem simply, as though there were no question of e.g. cyclical styles, such problems can be coped with automatically, but would complicate matters unduly for a simple discussion.

Again, as we have mentioned with information retrieval, small problems can be solved without a computer; Flinders Petrie, who first systematised the theory and the problem, used many small slips of paper to do it as early as 1899. The problem as formulated by Petrie is as follows: if we have a group of artifacts to be seriated, then we can summarise the information about them in a matrix whose columns represent the attributes of the artifacts and whose rows represent the artifacts themselves. Each artifact can than be 'scored' for the attributes it has by entering a 1 in the row for the artifact under the columns corresponding to the attributes it possesses. The matrix thus obtained is known as an incidence matrix, and there are several ways which have been found to manipulate this matrix in order to obtain a matrix in 'Petrie form'—
—that is, a matrix in which the 1's are so concentrated that they form as dense a band as possible, running diagonally from top left to bottom right. This seems to be a fairly reliable method of arranging the artifacts, and the new order down the side of the rearranged matrix will represent a chronological series, correct in so far as it is based on the assumptions of the seriation theory mentioned above.

There are more complex methods of dealing with this same problem, theoretically they are meant to treat the evidence more faith-

fully because they exploit the similarities between the artifacts as well as the attributes possessed by the individual artifacts. These methods depend upon the conversion of the incidence matrix mentioned above into a similarity matrix by comparing the attributes of each pair of artifacts and arriving at a score which expresses the result of the comparison. These scores are called similarity coefficients, and they may take many forms depending upon the archaeologist's judgment of how he wants the comparisons to be made. For example, the 'simple matching' coefficient is obtained by comparing the attributes of two objects and arriving at a score consisting of the number of attributes possessed by both added to the number of attributes possessed by neither, divided by the total number of possible attributes. Similar coefficients using similar combinations are possible, subject to the archaeologist's choice.

Once a similarity matrix has been prepared, one has the choice of using it as it is or of using various matrix transformations to take advantage of the links between more than two artifacts. However one wishes to proceed, the next step is to utilise the technique of non-metric multidimensional scaling in order to obtain from the matrix of similarities a configuration of points representing the artifacts and the distances between them. This technique also makes possible the reduction of dimensions to three or two, with what will hopefully be an efficient and accurate combination of trends. The result will be a scatter diagram which will show, if the technique has been successful, a 'band effect' indicating a single major trend which should correspond with chronological movement. If a listing of the series is wanted, the points in the scatter diagram can be projected upon the first principal component of the configuration and read off in this way.

The use of these computer techniques for automatic seriation is not without problems; as mentioned before, it depends heavily upon the archaeologist's understanding and careful use of the techniques available. And most of these problems are of a kind which calls for the archaeologist's decision: appropriate similarity measures, deletion of attributes which are so common as to distort the seriation, choice of definitive typologies. But the seriation strategy can be used to some effect wherever a chronological series needs to be established with minimum aid from other evidence: the graves in a cemetery or the layers on an occupation site can also be seriated. However, as it is only a machine technique, it will of course be up to the archaeologist to tell which end of the series is which.

### IV. Typology (automatic classification).

The idea of finding some objective method for the classification of objects of all kinds has been much advanced in recent years by the development in zoology of numerical taxonomy. Some of these methods have been taken into use by archaeologists as well in an attempt to formulate objectively stated classification systems for artifacts while retaining the basic elements of archaeological method.

In the search for an automatic method, several theoretical points must first be settled. Does the archaeologist want classes which are strictly defined by reference to a specific and limited set of attributes ('monothetic' classes) or does he prefer the more natural polythetic classes, which include members which are not absolutely identical? If he chooses the latter, which seems the more appropriate to the archaeological case, he must be aware that such a classification will not yield unambiguous keys or lists of attributes for each class. Finally, he should of course be aware of all the rest of the hotly disputed issues surrounding the typology question in archaeology, since many of the decisions he takes for automatic classification will depend upon his answers.

The most commonly used multivariate technique in this field is cluster analysis. This method can be thought of as an examination of the clustering of points in a space according to their similarity to one another, though in actuality such a configuration need only be obtained for the convenience of the archaeologist using a dimension-reducing program. What the cluster analysis techniques do, starting with a similarity matrix calculated for the artifacts in question on the basis of their attributes, is to divide the

whole of the assemblage off into classes, to collect individual members into classes, or to partition into classes provisionally and then rearrange until some criterion is met. We will discuss three clustering methods briefly.

Some divisive strategies may be basically inappropriate because they yield monothetic classes, but they are fast and can be used for preliminary partitioning. The operative principle is to choose the attribute which best divides the entire assemblage into two groups, one which possesses it and one which does not. This is done progressively to each of the two groups thus formed, then to each of the four groups, and so on for some given number of divisions chosen by the user (five or six being usual). The attributes which have been used to make the divisions are thus those which define the resulting classes. The classes are also arranged hierarchically.

The agglomerative strategy tends to yield polythetic groups. It's principle is to proceed by considering all of the individual artifacts, pair by pair, joining together any pair that meets certain requirements of similarity. Continuing similarly, these requirements are relaxed by increments so that other members can join these original clusters, forming an inverted hierarchic tree with all clusters joined into a single stem at the 'top'. One has a choice of several methods for determining the 'admission requirement' of similarity. The single-link method allows a new member to join a class if it is sufficiently like any single other member. The complete-link method requires that it be sufficiently like all other members. Finally, the average-link method requires that the joining member be compared to an average of the existing cluster members. It will be obvious that there are numerous archaeological arguments to be advanced in favour of each of these methods, and all three have been used in archaeological applications.

A third method of clustering is the reallocation or K-means approach. One of the other methods is first used to create a clustering, and some hierarchic level of interest is chosen for further examination. At that level the cluster centre for each cluster is calculated, and any members which are sufficiently 'on the fringe' of their clusters are shifted if possible to clusters which they more closely resemble. Cluster centres are then recalculated and the process is repeated until the clusters have become sufficiently 'tight', a state which will be reached when no further reallocations can be made.

We do not argue for the use of any of these approaches, as the choice of the method appropriate to his problem should be the province of the researcher himself. Any of them will force the archaeologist to take a good hard look at the attributes in question and to consider very carefully just what it is that he does when he classifies. Again, a machine method is of most use with large problems of classification, and though it cannot promise perfect results from every point of view, it can be of immense help in making preliminary classifications of large groups of artifacts.

**V. Modelling.**

Every time an archaeologist makes use of one of the multivariate techniques we have discussed, he accepts its structure as a model of the theory——seriation, for example——which has supported his choice. Explicit model-building with computers is already well-advanced in locational geography, and similar uses have been made in archaeology with reference to geographical distributions. Models have been proposed as explanatory hypotheses to deal with the migration of populations and as heuristic tools for the analysis of kiln production. Models such as these seek to attain to an acceptable level of mathematical accuracy, and as long as they attempt to deal with no more than a few variables they can show quite effective results. But most archaeological problems have to be drastically simplified before they can be treated in this way, and few more ambitious projects have even been proposed, much less successfully completed. From the point of view of the use of computers, modelling for archaeology will probably be most effective in the foreseeable future as part of the process of hypothesis formation and testing.

This process is simple in outline. On the basis of already acquired data a problem is recognised and explanatory hypotheses are devised to cope with it. If these hypotheses

can be formalised, a computer model can be made which specifies a structure with variables at certain problem points where the hypotheses do not agree. Put onto the computer, the model can be used for a simulation, allowing the system to work with several sets of variables so that the results of the use of the different hypotheses are known. These results, in the form of model-generated artificial data, can then be compared with the original data and if desired with new data acquired as part of an organised research plan to test the results. The model or its variables can be adjusted if the authentic data seems to call for this; the process can be repeated as often as seems necessary.

We will discuss an experiment of this kind to suggest both how a model simulation of this kind is constructed for the computer and why the computer is necessary. The example is a model of a cemetery, constructed to compare with several excavated cemeteries under study. The first step is the specification of the main factors at work in the formation of the cemetery, which will be the parameters of the model. These will include the length of time that the cemetery was in use, the population from which the burials in the cemetery were drawn, the variety of types of artifact appearing in the graves, the popularity of the various artifact types over time, and the lifetimes of the various types. It will be seen from this list that we do not try to specify factors which the archaeological record cannot show directly, such as the distribution of social classes, but use what evidence there can be, as the variety of types in the graves. The values that these parameters will take are then specified. If, for example, we feel that the particular population in question was small at the beginning, grew to a peak, and then waned, we can approximate this hypothesis by using a Gaussian curve for the size of the population over time. Our aim is to specify the individual distributions, which are amenable to archaeological reasoning, and to use the computer to obtain the results of their interaction. We may further constrain the program which simulates our model by allowing it to generate only the number of graves and types which are known to us from the excavation of an actual cemetery. Such a program can then sample from the distributions and calculate the probability of each type's appearance in each grave, and an artificial cemetery which obeys the designated restraints can be produced by comparing this probability with a random value, allowing the artifact to appear in the grave if the comparison yields a certain range of values.

Many other examples of such modelling possibilities come to mind: the output of a potter's workshop based on demand, function of forms, number of workers, firing damage, availability of material, etc; the horizontal distribution of lithic assemblages based on function, discard rate, loss rate, a 'scuff index', etc.——the possibilities range as widely as the archaeologist's problems. And for this sort of sophisticated approach to his problems the computer can lend him all the aid he is capable of using.

The computer is a tool like any other in archaeology. It combines the delicacy and accuracy of the trowel with the power and speed of heavy earth-moving equipment. Like the rest of the archaeologist's tools, it demands informed intelligence to determine when, where, and why to use it. Very small problems are a waste of its power, but large and complex problems almost demand it. The computer itself is no more than a huge extension of the archaeologist's capability for making objective judgments and maintaining them with consistency. Its use will depend upon his desire to do so.

For further reading:


Doran and Hodson, *Mathematics and Computers in Archaeology*, Edinburgh, 1975.

Borillo (ed.), *Les Methodes mathematiques de l'archeologie*, Marseilles, 1972.

Hodson, Kendall, and Tautu (eds.), *Mathematics in the Archaeological and Historical Sciences*, Edinburgh, 1971.

*Newsletter of Computer Archaeology*, Dept. of Anthropology, University of Arizona.

*Computer Applications in Archaeology* (proceedings of the annual conference in Birmingham,*1973-*) Computer Centre, University of Birmingham.

ROALD SKARSTEN:

# THIRD INTERNATIONAL CONFERENCE ON COMPUTING IN THE HUMANITIES

2.–5. august 1977,
University of Waterloo, Canada

Den tredje internasjonale konferanse om databehandling i humaniora ble holdt 2.–5. august i år i Waterloo, Canada. Disse konferansene avholdes i Nord-Amerika annethvert år. I det mellomliggende år avholder Association for Literary and Linguistic Computing (ALLC) lignende konferanser i Europa. Neste konferanse blir holdt i april 1978 i Birmingham (jfr. egen melding). Disse konferansene har betydd mye for den kraftige ekspansjon som har funnet sted når det gjelder databehandling i humaniora. Konferanserapportene har vært en spesielt viktig litteraturkilde fordi feltet naturlig nok ikke har hatt så mye litteratur å bygge på til denne tid. På konferansen i Canada ble det dannet et canadisk »Learned Society for Computing in the Humanities.»

Årets konferanse var lagt til University of Waterloo, Canada, og var arrangert av dette universitet i samarbeid med University of Montreal. Universitetet i Waterloo er ungt, 20 år gammelt, men har allikevel markert seg i internasjonal sammenheng på bl.a. databehandlingsområdet. Den ikke ukjente WATFOR-kompilator kommer derfra (WATerloo-FORtran). Som universitet er det kjent for sitt »cooperative education system». Det bygger på den forutsetning at teori og praksis bør læres i sammenheng. Studentene studerer 4 måneder, arbeider 4 måneder og studerer 4 måneder igjen osv., inntil de er ferdige. Da har de til sammen vært ute i arbeidslivet i ett år. Ved at studentmassen deles i to, blir stillingene i arbeidslivet kontinuerlig besatt. Studentenes rapporter fra arbeidet inngår i den akademiske kvalifisering. Dette systemet har vist seg å være populært blant studentene.

Årets konferanse i Waterloo hadde samlet ca. 300 deltakere fra hele verden, og den adskilte seg fra sine forgjengere ved den bredde av aktiviteter som ble dokumentert. Tidligere har språk og litteratur og til dels historie vært dominerende. Denne dominansen var i år redusert til fordel for fag som musikk og dans og skapende aktiviteter innenfor musikk og grafisk kunst.

Innenfor de to sistnevnte områder var der konserter og utstillinger. Fra programmet kan nevnes titler som »A Personal View of Computer Graphics», »An Evening of Canadian Computer Music» og »A festival of computer – generated and computer-animated films» Til og med en opera »Daedolus», av Theo Goldberg ble presentert. Alt i alt ga dette konferansen et særpreg som gjør at den vil bli stående som en betydningsfull konferanse.

Prosjekter av umiddelbar nytteverdi tiltrekker seg alltid spesiell oppmerksomhet. På denne konferansen var det et prosjekt for datamaskinell oversettelse og automatisk produksjon av blindeskrift som fikk denne oppmerksomheten. Bokproduksjon med blindeskrift er en meget komplisert og kostbar affære. Et system med en database på mer enn 50 000 representative franske ord som kan brukes til automatisk oversettelse fra fransk tekst til blindekode og utlisting i blindeskrift, ble presentert. Det ble omtalt som det første i sitt slag i verden. Kostnadsbesparelsene var så store at man forutså en snarlig og vesentlig økning av fransk litteratur tilgjengelig i blindeskrift.

Publiseringsproblemet ble drøftet i en gruppe. P.g.a. store datamengder, f.eks. i datamaskinelle språkarkiv, er vanlig publisering i bokform økonomisk sett nærmest umulig. Microfiche blir nå mer og mer brukt

til publisering av store datamengder, og adskillige prosjekter rapporterte at de planla å bruke microfiche. Det ble hevdet at de psykologiske problemer som er forbundet med microfiche-lesing fra publikums side måtte vurderes som forbigående og at de økonomiske argumenter ville bli utslagsgivende.

Det ble sagt at microfiche ville bli forskningens »paperback».

En arbeidsgruppe i statistikk drøftet et forslag om bestemte minstekrav som burde kreves av publiserte artikler som inneholdt kvantitative utsagn om språklige ytringer. Bl.a. ble en del standard statistiske mål nevnt, og man ønsket at »rådata» skulle være tilgjengelig for publikum, slik at det var anledning til å kontrollere de publiserte resultater.

På området språk og litteratur ble det presentert prosjekter innenfor forskjellige språkområder: tysk, engelsk, gresk, fransk, spansk og hebraisk.

Et storstilt amerikansk prosjekt:

Standford Computer Archive of Language Materials (CALM) lager et datamaskinelt arkiv av språkdata fra temmelig mange språk, og ønsker å kunne yte service til alle forskere som ønsker komparativt lingvistisk materiale. (Se for øvrig omtale i forbindelse med melding om engelsk-språklige tekstarkiv i dette nummer).

Kvantitativ stilistikk (stylometrics) var også denne gang fyldig representert. Personlig syntes jeg å merke en tendens til større forsiktighet når det gjaldt den litterære tolkningen av statistisk sett signifikante resultater. Man nøyet seg ikke med å fastslå de signifikante utslag, men man brukte disse som utgangspunkt for en tilbakevending til teksten for nærmere studium og forsøk på å undersøke om alle de nødvendige betingelser for bruk av signifikanstesten var oppfylt. For å sitere en foreleser, Stephen V.F. Waite: »Ultimately, the judgment of the value of a result suggested by the computer must be decided by the scholar who returns to the text with a keener eye and a greater awareness».

Av nyheter festet jeg meg særlig ved en ny formel for vokabularmåling som ble presentert av J.A. Leavitt, og kalt for SPAN (SPan ANalysis). Hovedpunktet i den nye formel er at man måler grupperinger (clustering) for bestemte kategorier av ord (substantiv, verb osv.), slik at f.eks. tekster som er like store og med ord som har samme frekvens, kan adskilles fra hverandre ved å måle forskjellen i ordenes distribusjon (gruppering, intern avstand) i tekstene. Leavitt hadde funnet at adverbet var den beste tekstdiskriminator, og substantivet den dårligste. Det skal bli interessant å se om denne nye formel står sin prøve etterhvert som den blir testet på flere tekster.

Leavitt var selv ikke i tvil om verdien av den nye formel: »We claim to have developed a new lexicostatistical measure, SPAN, the unique feature of which is a clustering parameter». De som er interessert i å studere denne formel nærmere, bør skaffe seg konferansrapporten »Computing in the Humanities», editors Serge Lusignan and John S. North, The University of Waterloo Press, 1977).

Konferansens »Proceedings» forelå for øvrig til konferansen, og alle foredragene i stensilert form. Konferansen var i det hele velorganisert. Det illustreres også ved at deltakerne fikk utdelt en fyldig bibliografi, »A checklist of books and journals related to computing in the Humanities», og at det var en egen utstilling av bøker som var nevnt i denne bibliografien.

## IBSEN-konkordans

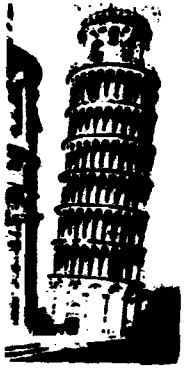
Som nevnt i forrige nummer har det i den siste tid vært i gang et større planleggingsarbeid i forbindelse med databehandling av Ibsens skuespill og dikt. Etter oppdrag fra en samarbeidsgruppe har George M. Gillow laget en rapport kalt Database Techniques in the Literary and Linguistic Research (se egen melding).

Dessuten har NAVF's EDB-senter utarbeidet en prosjektplan for en Ibsen-konkordans. Som ledd i prosjektarbeidet utfører NAVF's EDB-senter sammen med professor Harald Noreng en prøvestudie med lemmatisering »Når vi døde vågner». En søknad om støtte til tiltaket vil bli behandlet av Rådet for humanistisk forskning i oktober.

Også på engelsk side har arbeidet gått videre idet en ved The Literary and Linguistic Computing Centre i Cambridge har punchet ca. halvparten av tekstene etter Hundreårsutgaven.

NAVF's EDB-senter kan gi interesserte flere opplysninger.

Knut Hofland:



# DEN 4. INTERNASJONALE SOMMERSKOLE I PISA

Den 4. internasjonale sommerskole i datamaskinell og matematisk lingvistikk ble holdt i Pisa, Italia i tiden 8. – 26. august 1977 med 160 deltakere (derav 5 norske) fra ca. 40 land, også noen utenfor Europa. Deltakerne kunne velge mellom 12 kurs og 4 arbeidsgrupper med en varighet på 10-12 timer for hvert kurs. Det var 5 dobbeltforelesninger hver dag, og bare et par av kursene overlappet i tid slik at det var mulig å kombinere de kursene en ønsket. I tillegg til det oppsatte programmet ble det holdt et kurs i programmeringsspråket LISP, som er velegnet, og i USA nesten enerådende, for bruk innen datamaskinell lingvistikk. En del av deltakerne holdt også forelesninger om egne forskningsprosjekter.

Foreleserne var som følger:
Kurs:
J. BRESNAN: A realistic transformational grammar.
R. DILLIGAN: Introduction to computers *and programming fundamentals.*
M. HALLIDAY: A functional approach to grammar.
R. KAPLAN: Computational psycholinguistics: the syntactic component.
L. KARTTUNEN: Introduction to logic.
L. KARTTUNEN: Model-theoretic semantics for natural languages.
M. KAY: Computational linguistics and artificial intelligence.
E. KEENAN: Grammatical relations and universal grammar.
G. LAKOFF: Linguistic gestalts and cognitive grammar.
J. PETÖFIE: Structure and function of a text-theory.
Y. WILKS: Language, linguistics and processes.
L. STEELS: Introduction to computational linguistics.

Arbeidsgrupper (workshops):
E. Hajičová: Functionalism and topic focus in generative description.
M. Kay: Computational linguistics.
G. Ramcas: Introduction to linguistic dataprocessing.
I. A. Mel'čuc: Linguistic models of the meaning.

Kursene og arbeidsgruppene bestod både av emner som var rent lingvistiske og emner der EDB var en sentral faktor. Det ble holdt innføringsserier om datamaskin og programmering, EDB som hjelpemiddel i lingvistikk og om datamaskinell lingvistikk. Videre var det forelesninger over mer spesielle emner som syntaks, semantikk og kunstig intelligens (AI).

En svakhet ved sommerskolen var rekkefølgen på kursene. Innføringskurset i datamaskinell lingvistikk startet først andre uke, og da var det allerede holdt to kurs som krevde dette kurset som grunnlag. Dette var enda mer beklagelig siden innføringskurset var et av de mest vellykkete kursene ved sommerskolen. En annen ulempe var at ikke alle foreleserne hadde skriftlig materiale som dekket forelesningene.

Nærmere kommentarer til noen av kursene.

*Ron Kaplan:* Computational psycholinguistics: the syntactic component.

Kaplan har både jobbet sammen med Woods på hans ATN-parser og med Kay på hans chart parser, og hans modell er en variant av Kay's. Mens Kay's modell er i stadig utvikling, har Kaplan prøvd å holde modellen enklest mulig, men samtidig sette den i stand til å behandle de vanligste syntaktiske fenomen. Videre er han interessert i å trekke inn psykolingvistisk kunnskap slik at den syntaktiske modellen avspeiler den kognitive prosessen. Dette gjelder bl.a. spørsmålet om den datamaskinelle prosessen skal foregå fra bunnen av og opp i setningen og etter bredde først, eller på annen måte.

Forelesningene var utbytterike for dem som hadde noe kunnskap på forhånd, og Kaplan holdt flere forelesninger utover den oppsatte planen for å dekke spesielle ønsker.

*Y. Wilks:* Language, linguistics and processes.

Wilks gikk gjennom og kommenterte forskjellige system for språklig forståelse (natural language understanding).

Han delte systemene inn i to grupper. De første fra perioden 1972-1974 med bl.a. systemene til Winograd, Charniak, Schank og Wilks, og de som har fremkommet etter 74 og som bygger på begrepet ramme (frame). En trenger en viss mengde bakgrunnskunnskap for f.eks. å vite hva en preposisjon peker på, for å finne den rette syntaktiske struktur i en setning eller for å få tak i det semantiske innholdet. Problemet er hvorledes slik kunnskap skal organiseres, samt å knytte den kunnskapen som ligger i en tekst, til denne bakgrunnskunnskapen. Ved de nyere systemene lagres bakgrunnskunnskapene som rammer, (frames), d.v.s. som oppskrifter på vanlige hendelsesforløp og reaksjonsmønstre ved dagligdagse forhold som reising, kinobesøk, restaurantbesøk osv. Problemet er at en får et stort antall rammer, og på grunn av rammenes størrelse kan en bare ha tilgjengelig et par rammer på en gang. Det gjelder å vite hvilken ramme en skal velge og når en skal skifte ramme. I tillegg vil alle uvanlige hendelsesforløp falle utenfor rammeinnholdet.

Wilks, som er en sentral person på dette fagfeltet, kom med mange interessante personlige vurderinger. Men forelesningene ble skjemmet av et par pedagogiske svakheter som altfor hurtig og slurvete tale og klosset bruk av overheadprojektor.

*L. Steels:* Introduction to computational linguistics.

Dette var et av de lengste kursene med 20 dobbelttimer. Kurset startet med en oversikt over forskjellige former for lingvistisk representasjon, som listestrukturer, syntaktiske strukturer og semantiske strukturer. Videre gikk han inn på systemer for generering, parsing og produksjon av setninger. Til slutt ble det gjennomgått forskjellige former for representasjon av kunnskap. Til kurset var det utarbeidet en øvingsbok med henvisning til steder hvor teoriene ble behandlet. Forelesningene var en blanding av teori og gjennomgåelse av eksempler. Ved at man supplerte med en del selvstudium av teori, gav kurset en god innføring i emnet. Den eneste haken ved kurset var, som nevnt i innledningen, plasseringen på timeplanen. Dette kurset burde vært konsentrert til første del av første uke. En kunne da hatt større utbytte av flere av de andre kursene.

*Martin Kay:* Computational linguistics and artificial intelligence (Kurs).
Computational linguistics (Workshop).

Kay innledet sitt kurs med å beskrive sitt »reversible grammar system». Det nye i forhold til hans tidligere system er at det prinsipielt ikke er noen forskjell på produksjon og analyse. I tillegg inneholder systemet en grammatikk som dynamisk forandrer seg under analysen. Andre del av kurset handlet om semantiske nettverk og et språk for representasjon av kunnskap KRL (knowledge representation language). KRL er et deklarativt språk bygget på LISP. Ordrer for behandling av kunnskap vil også bli lagt inn

i språket, men må foreløpig skrives i LISP. KRL er et språk som en stiller visse forventninger til innen kunstig intelligens.

Til Kay's arbeidsgruppe var det avsatt 5 dager, og det meldte seg 25-30 deltakere. De som var interessert, kunne sende inn forslag om hva de ville arbeide med. Ved inndeling i grupper ble dette ikke tatt noe videre hensyn til både på grunn av det store antallet og på grunn av Kay's egne interesser. De forskjellige gruppene kom i høy grad til å dreie seg om syntaks, bl.a. om bygging av en parser og om grammatikk for et språk med fri ordstilling. På grunn av antallet grupper ble kontakten med Kay begrenset til et par timer pr. dag. Sammen med manglende forarbeid gjorde dette at arbeidsgruppen ikke helt svarte til forventningene.

Sommerskolen i Pisa er et utmerket tiltak. En får her grundigere kunnskap enn det som er mulig å skaffe seg på konferanser som varer i 5-6 dager. Foreleserne er de fremste på sitt fagfelt, og dette gjør forelesningene aktuelle. Ved sommerskolen var det et godt bibliotek som også inneholdt upubliserte artikler. En kunne der få en god oversikt over det som er skrevet på områdene. Det er meningen at forelesningene ved sommerskolen skal publiseres. Neste sommerskole skal etter planen holdes om 2 år. Det er å håpe at det kan være enda flere nordmenn til stede ved den sommerskolen.

# MICROFICHE-UTGAVE AV ENGELSKE TEKST-ARKIV

1. The Brown Corpus.

The Brown University Corpus of Present-Day American English (se omtale annetsteds i dette nummer) vil nå bli tilgjengelig også i mikro-form. Språkdata, Gøteborgs Universitet, som tidligere har arbeidet med planer om en microficheutgave, har inngått et samarbeid med NAVF's EDB-senter om saken.

Ifølge overenskomsten skal NAVF's EDB-senter stå for produksjonen i Bergen og samarbeide med Språkdata om konkordansformatet og ved distribusjonen. Det vil bli utarbeidet en fullstendig konkordans til korpuset (som er på 1 million ord) på microfiche.

Samtidig vil hele tekstmaterialet bli overført til microfiche slik at en også kan søke i grunnmaterialet når konteksten i konkordansen ikke er tilstrekkelig. Konkordansen og selve korpuset vil kreve ca. 90 kort (14,5 x 10,5 cm). Hver microfiche inneholder ca. 208 A-4 sider. Prisen er beregnet å ligge mellom kr. 300 - 430, avhengig av hvor mange kopier det blir aktuelt å produsere. Microficheversjonen vil bli klar i løpet av første kvartal 1978.

*NAVF's EDB-senter vil gjerne nå ha kontakt med institusjoner som ønsker å bestille en kopi.* De institusjoner som tidligere har meldt sin interesse til Språkdata, trenger ikke å bestille på nytt dersom de ønsker å opprettholde sin bestilling.

2. **CAMET (A Computer Archive of Modern English Texts.)**

Dette prosjektets status er beskrevet i artikkelen »Pågående arbeid med engelske tekstarkiv». Det er planen å produsere en microficheversjon av CAMET bestående av tekstene, en fullstendig konkordans og statistiske tabeller. NAVF's EDB-senter vil stå for produksjonen. Når korpuset er bruksklart, vil det bli gitt en ny melding.

3. **Survey of Spoken English.**

Det kan også meldes at professor Jan Svartvik, Lunds universitet, har planer om å utgi en microficheutgave av det engelske talespråkmaterialet fra Survey of English Usage. Det vil senere bli gitt nærmere melding om tiltaket.

# REFERAT FRA NORDISKE DATALINGVISTIKK-DAGER I GØTEBORG 10. – 11. OKTOBER 1977

De første nordiske datalingvistikk-dager ble arrangert av Språkdata, Gøteborgs universitet etter initiativ fra den nordiske samarbeidskomite for språklig databehandling.

Konferansen samlet 63 deltakere, hvorav 17 kom fra Norge.

Den norske deltakelsen var gledelig stor og vitner om en økende interesse for språklig databehandling i vårt land.

På konferansen ble det holdt i alt 12 foredrag og 5 av disse var av norske bidragsytere. Et konsentrat av alle presentasjonene ble utsendt før konferansen, noe som gav deltakerne anledning til å møte vel forberedt.

Foruten å gi innsyn i andres prosjektarbeid, hadde konferansen som mål å gi anledning til en drøfting av behov og utviklinglinjer innenfor datamaskinell lingvistikk og mer generelt på feltet språklig databehandling. Organisatorisk var det lagt opp til enkle, men møte-teknisk effektive løsninger. Den enkelte bidragsyter hadde ansvar for sin del av programmet, og dette gav konferansen et »selv-genererende» preg uten det arrangement-tekniske apparat som ofte kan føles som en belastning både for arrangør og deltaker. I det hele ble vertskapsoppgavene meget bra løst av Språkdata. Nedenfor blir det gitt en kort orientering om de foredrag som ble holdt. Det bør nevnes at det på konferansen ble vedtatt å publisere foredragene.

En rekke foredrag var viet automatisk språkanalyse. To av dem redegjorde for utviklingsarbeid bygd på professor Martin Kay's »chart-analysis». *Anne-Lena Sågvall* redegjorde for automatisk analyse av finsk morfologi, mens *Knut Hofland* presenterte arbeid i forbindelse med implementering av denne analysemetode for automatisk syntaktisk analyse av norsk.

*Mats Eeg-Olofsson* presenterte forskningsprosjektet »Algoritmisk text-analys» hvor målet er å utarbeide formaliserte metoder for grammatisk analyse av autentisk svensk tekst. I prosjektet er bl.a. emner som automatisk morfologisk analyse, homografi og lemmatisering sentrale.

*Benny Brodda* omtalte en metode for morfologisk analyse uten leksikon. I metoden, som bygger på substitusjonsgrammatikk, er det lagt vekt på å formulere enkle sett med substitusjonsregler.

Bruk av EDB ved etablering av leksikalske databaser og ved publisering av ordbøker var emne for 4 foredrag.

*Bo Ralph* redegjorde for forprosjekteringen av et større ordbokprosjekt »Lexikalisk databas». Prosjektet sikter mot å bygge opp en database som inneholder en stor mengde svenske ord med fyldig språklig informasjon. Alle ord blir knyttet sammen i et nettverk og vil bli definert ved hjelp av et på forhånd fastsatt grunnordvokabular.

*Harald Solevåg og Kolbjørn Heggstad* presenterte et pågående prosjekt om nyord-registrering i norsk og redegjorde for forhold som taler for at en i det fortsatte prosjektarbeid tar i bruk databaseteknikk.

*Hanne Ruus og Bente Maegaard* rapporterte fra prosjektet DANWORD. Formålet her er å undersøke et representativt utsnitt av moderne dansk med automatiske metoder. I presentasjonen ble det særlig gjort greie for prinsippene for utvalg av korpus og for det EDB-arbeid som er utført når det gjelder lemmatisering.

*Ivar Fonnes* gjorde greie for EDB-arbeidet ved tilrettelegging av Norsk Landbruksordbok for trykking. Ordboken er en vitenskapelig definisjonsordbok med et komplisert tegnsett og synonymer på en rekke språk. Ved hjelp av eksempelmateriale ble fordeler og problemer ved EDB-basert trykking drøftet.

Til slutt kan nevnes en del enkeltstående presentasjoner. *Viljo Kohonen og Jussi*

*Salmela* orienterte om et program, CHITAB, som er utviklet for å kunne håndtere detaljert språklig klassifikasjon til et tekstmateriale med vekt på krysstabulering og frekvensoppstilling.

*Hans Basbøll og Kjeld Kristensen* rapporterte fra sitt prosjekt om »Computergenerering af lydskrevet dansk ud fra en quasi-ortografisk notation ved hjælp af generative fonologiske regler». Prosjektet hadde i første omgang som mål å teste og forbedre en generativ fonologi for dansk. På det nåværende trinn ønsker en bl.a. å utvikle metoder for regelsyntese av dansk tale.

*Marina Mundt* presenterte resultater fra arbeid med språklig analyse av Håkonar saga med vekt på en diskusjon av funksjonsordenes verdi som analyse-enheter ved en språklig-stilistisk undersøkelse.

*Gulbrand Alhaug* drøftet emner i forbindelse med oppbygging av en morfemordbok for norsk. Særlig vekt ble lagt på de problemer som oppstår når man ønsker å oppnå en gjennomgående og konsistent morfeminndeling av eldre og nyere lånord.

Under konferansen var det også satt av plass til demonstrasjon hvor deltakerne fikk anledning til i større detalj å studere sentrale prosjekter ved Språkdata.

Den siste delen av programmet var, som nevnt, viet en drøfting av datalingvistikkens stilling i dag.

Av emner som særlig ble diskutert var den ulike forskningsinteresse innenfor datamaskinell språkbehandling mellom nordiske og ikke-nordiske forskere, databasesystemers anvendbarhet i språklig databehandling og diskusjon av forhold som avholder forskere og studenter fra å bruke datamaskinelle metoder.

Når det gjelder forskningsprofil innenfor datalingvistikk, ble det hevdet at nordiske forskere overveiende var empirisk interessert og at mangelen på metodiske og eksperimentelle prosjekter innenfor kognitiv psykologi, semantikk og kunstig intelligens var tydelig når en sammenlikner med forskningsaktiviteten f.eks. i USA. Om denne situasjonen ble det på den ene side hevdet at den var uheldig fordi den avskjærer nordiske forskere fra medvirkning ved forskningsfronten, men det ble også gitt uttrykk for at en empirisk holdning var en fornuftig strategi på feltets nåværende utviklingstrinn.

Det syntes å være enighet om at datalingvistikken er, eller rettere bør være, en viktig faktor ved utvikling av datamaskinelle informasjonssystemer og at ikke minst kommunikasjonsaspektet ved interaktiv databehandling, herunder talesyntese, burde være et viktig innsatsområde for datalingvister.

I drøftingene kom det også fram ulike synspunkter på databasesystemer. Slike systemers tekniske fortrinn ble fremhevet samtidig som flere så praktiske begrensninger på grunn av det forutsatte ressursforbruk og fordi den vanlige bruker ville bli avhengig av spesialisthjelp ved opprettelse av databasen og kanskje også ved den praktiske utnyttelse. Det ble også vurdert som problematisk at de ulike datamaskintyper har sine egne databasesystemer.

Flere deltakere fant det vanskelig selv å bruke datamaskinelle metoder og å anbefale studenter å bruke datamaskin på grunn av utilfredsstillende undervisningstilbud og maskinressurser. Andre mente med hell å ha introdusert datamaskinelle metoder i undervisning og forskning.

Det var imidlertid enighet om at spørsmålet om opplæring i språklig databehandling og valg av strategier for å øke anvendelsen av EDB-metoder var et meget viktig tema. Det ble derfor vedtatt å ta emnet opp igjen i større bredde på det neste møte, som vil bli holdt i København i 1979.

Ved avslutningen av konferansen ble en del praktiske spørsmål drøftet. Den nordiske samarbeidskomite for datamaskinell språkbehandling ble formelt valgt, meldingsbladet COMPILING ble ansett som et velegnet kontaktorgan, og det redaksjonelle ansvaret ble overlatt til arrangøren av de neste nordiske datalingvistikkdager, Bente Maegaard fra København.

# INNSTILLINGEN »EDB OG PRIMÆRKILDER».

En innstilling om »EDB og primærkilder» ble lagt fram i januar i år. Innstillingen er utarbeidet av et utvalg nedsatt av NAVF's styre. I innstillingen gjennomgår utvalget relevante sider ved den tidligere NAVF-utredningen om primærkilder og drøfter EDB-aspekter ved det remissmateriale som kom inn som svar på primærkildeinnstillingen. Utvalget utreder begrepet primærkilder og anskueliggjør hvordan man kan inndele primærkilder i en rekke kategorier. Videre gis det en analyse av de fordeler EDB kan gi i arbeid med ulike typer primærkildedata.

Utvalget slår fast at databehandling er ressurskrevende og at det må foregå en nøye planlegging og prioritering av primærkildetiltak.

Et eget kapittel er viet EDB-utstyr for primærkildebehov. Det gis også en oversikt over NAVF's engasjement på feltet EDB og primærkilder de siste årene.

Utvalget rår til at det etableres nasjonale EDB-tjenester for å sikre kompetanseoppbygging, konsultative funksjoner, informasjon og dataservice.

Det anbefales at det i tiden som kommer, blir satt i gang nasjonale fellestiltak hvor sentralt humanistisk kildemateriale kan bli overført til datamaskinell form. De forskningsdirigerende sideeffekter som databehandling kan gi, må motvirkes ved aktiv prosjektstimulering også i form av regionale tiltak. Arbeidet med faglig tilrettelegging av primærkilder i oppbevaringsinstitusjonene bør styrkes.

Konkret foreslås det opprettet en dokumentasjonstjeneste for feltet EDB og primærkilder og at et eget tverrfaglig primærkildeforum etableres.

Innstillingen har senere vært behandlet av NAVF's styre sammen med de øvrige utarbeidede saksdokumenter. Etter vedtaket i NAVF's styre oppfordres Rådet for humanistisk forskning spesielt til å føre arbeidet videre innenfor sitt område og å utarbeide et programforslag for organisering av konkrete tiltak.

**EDB- seksjon for de humanistiske fag ved Universitetet i Bergen.**

Det historisk-filosofiske fakultet ved Universitetet i Bergen vedtok høsten 1975 å opprette en EDB-seksjon ved Fakultetet. Det ble oppnevnt et interimstyre for seksjonen som skulle planlegge organisering og drift og uttale seg om EDB-saker.

I løpet av våren 1977 har Fakultetsrådet viderebehandlet spørsmål knyttet til EDB-seksjonen. Det er bl.a. vedtatt at EDB-seksjonen blir organisert frittstående i forhold til eksisterende EDB-miljøer, at EDB-seksjonen skal være rådgivende for Fakultetet i EDB-spørsmål og at det etableres et samarbeidsutvalg for Fakultetes EDB-tiltak, sammensatt av representanter fra NAVF's EDB-senter og de institutter som bruker EDB, samt representanter for EDB—seksjonen.

Når Universitetet i Bergen får en stilling som EDB-konsulent (prioritert for 1978) vil konsulentens tilknytning til seksjonen bli nærmere vurdert.

I det styret som ble oppnevnt i sept. 77 sitter universitetslektor Magnus Rindal (formann), universitetslektor Geir Berge og vit. ass. Jan Oldervoll.

# IVAR FONNES: *EDB-situasjonen for humanister ved Universitetet i Oslo*

I 1976 anskaffet Universitetet i Oslo et nytt data-anlegg av typen DEC-10. Anlegget har en del egenskaper som er av spesiell interesse for humanister. Følgende kan nevnes:

Anlegget er spesielt orientert mot interaktiv bruk og har gode hjelpemidler for tekstredigering o.l.

Fullt ASCII tegnsett (med store og små bokstaver) er standard i alle programprodukter og på linjeskriverne.

Masselageret er på 1000 mill. tegn med gode muligheter for betydelig økning i tiden fremover.

Tilknytning av terminaler til anlegget kan skje via linjer som er lagt opp til HF-bygget. Brukerne må selv anskaffe terminaler og sørge for å få linjene ført frem til det aktuelle rom. EDB-tjenesten ved HF (ved NAVF's EDB-konsulent) gir råd og hjelp ved anskaffelser og linjeopplegg, og forsøker å koordinere anskaffelsene slik at utstyret skal komme flest mulig brukere til gode. EDB-tjenesten anskaffer også utstyr over sitt eget budsjett og stiller dette til disposisjon for brukere ved fakultetet. Det største problem i øyeblikket er å finne egnet plassering for EDB-utstyr i HF-bygget, da alle institutter sliter med plassproblemer.

For tiden er det anskaffet 3 skriveterminaler i HF-bygget, og dessuten har Institutt for Musikkvitenskap en i Chateau Neuf. Det er bevilget penger til ytterligere 3 terminaler i inneværende år. Disse vil bli anskaffet i løpet av relativt kort tid. Ved utgangen av året vil HF-brukerne dermed ha adgang til 6 terminaler (den ene av de nåværende skal byttes inn). For 1978 er det foreslått 2 nye terminaler, noe det er rimelig grunn til å tro kan bli bevilget.

EDB-tjenesten har også fått bevilget til en Versatec printer-plotter, som dels kan brukes som vanlig linjeskriver, og dels (og samtidig) til å konstruere tegn som ikke finnes i det vanlige tegnsettet. Dette gir muligheter for å skrive ut f.eks. aksenter, kyrillisk og gresk alfabet osv. Et par tekniske problemer i forbindelse med tilknytningen skal avklares før vi går til bestilling.

Fra tidligere disponeres dessuten to hullkortpuncher, 3-4 hullbåndpuncher og en hullbåndleser.

På programsiden er ikke all overføring fra det gamle dataanlegget avsluttet. Men det meste av TEXT og VOTA samt deler av HISO er i operativ stand. Dessuten er det kommet til noen nye programmer for tekstanalyse, og DEC-10 har en rekke standardprogrammer som muliggjør lettvint håndtering av tekstdata.

## KONSULENTHJELP– PUNCHEASSISTANSE

EDB-spørsmål av interesse for humanistiske forskere kan tas opp med EDB-konsulenter i Bergen, Oslo og Trondheim.

Konsulentene, som har erfaring fra EDB-arbeid på ulike anvendelsesområder innenfor humanistisk forskning, vil også være behjelpelige med å formidle kontakt med andre fagfolk der det er ønskelig.

Særlig vil det være viktig å ta seg god tid til drøftinger med en EDB-konsulent ved planlegging av nye EDB-prosjekter.

NAVF's EDB-senter tilbyr også nye brukere puncheassistanse i forbindelse med prøveprosjekter innenfor de humanistiske fagområder.

*Bergen*

NAVF's EDB-senter for humanistisk forskning, Villavei 10, Boks 53
5014 Bergen-Universitetet.

*Oslo*

NAVF's EDB-konsulent Ivar Fonnes
c/o Historisk institutt,
Universitetet i Oslo
Postboks 1102 – Blindern Oslo 3

*Trondheim*

EDB-konsulent Eirik Lien
Norges Lærerhøgskole
Universitetet i Trondheim
7000 Trondheim

**Database Techniques in the Literary and Linguistic Research av Michael Gillow.**

Rapporten gir resultater av en prøvestudie utført i forbindelse med et planlagt Ibsen-prosjekt.

Etter å ha kommentert ulike mer konvensjonelle systemløsninger for tekstbehandling, konsentrerer forfatteren seg om å demonstrere hvordan et moderne databasesystem (DBMS-system) kan utnyttes i språklig og litterær databehandling.

Ved hjelp av databasesystemer kan man langt mer effektivt enn før lagre og søke fram data samtidig som systemet gir brukeren nye muligheter til enkelt å kunne endre databasen, f.eks. ved å legge supplerende data til den.

I egne kapitler blir det redegjort for utviklingen av programmer for interaktiv søking i tekstdata og programmer for strukturering av tekster for et databasesystem. Det gis også eksempler på hvordan et databasesystem kan tas i bruk ved tilrettelegging og behandling av Ibsens skuespill for ulike faglige analyser.

Rapporten kan fås kjøpt ved henvendelse til Prosjekt for datamaskinell språkbehandling, Nordisk institutt, Universitetet i Bergen.

**Data and Storage Structuring for Humanistic Data**

er tittelen på en studie skrevet av førstelektor Joan Veim, Institutt for Informasjonsvitenskap, Universitetet i Bergen. I repporten blir ulike slag humanistiske forskningsdata ordnet etter datatype, dataformat og bruken av data. Mot denne bakgrunn blir brukerne av humanistiske data, dvs. de humanistiske forskere, karakterisert med hensyn til EDB-anvendelser. I egne kapitler blir det gitt fremstilling om data og datalagringsstrukturer (sekvensielle, hierarkiske og nettverkstrukturer), og om ulike typer av lagringssystemer (filsystemer, »information retrieval systems» og databasesystemer).

I et avsluttende kapittel blir det redegjort for hvordan de ulike slag humanistiske forskningsdata krever forskjellige metoder og systemer for datalagring.

Rapporten kan fås kjøpt ved henvendelse til Institutt for Informasjonsvitenskap.

**EDB-prosjekt i norske bibliotek.**

Norsk utvalg for EDB i bibliotek har latt utarbeide en oversikt og beskrivelse av EDB-prosjekter i norske bibliotek. Oversikten kan fås ved henvendelse til Norsk Dokumentdata, Malerhaugveien 20, OSLO 6, tlf. 02/67 3480.

## COLING 78 CALL FOR PAPERS

The international Committee on Computational Linguistics is pleased to announce that the 7th International Conference on Computational Linguistics will be held in Bergen, Norway, from August 14 to August 18, 1978.
The Conference is being hosted by the Nordic Institute, Department of Computational Text Processing, University of Bergen, in collaboration with the Institute for Information Science and the University Computer Centre.
COLING 78 will consist of papers, panel-discussions, and demonstrations on the following themes:

- theories, methods and problems of computational linguistics, – its relations with the different branches of linguistics, artificial intelligence, mathematics, information science, etc.
- models for natural language processing: theory and implementation, – their different components (pragmatical, semantical, logical, syntactical, lexical, morphological, phonological, acoustical, etc.)
- applications of natural language processing: machine translation and machine-aided translation, man-machine communication, question-answering, database query languages, speech understanding and speech synthesis, text analysis and synthesis, automated terminology dictionaries, information retrieval and automated documentation, etc.
- automated processing of linguistic data collections: lexicology and lexicography; philological, literary, stylistic studies; textual and statistical processing (concordances, indices, etc.); historical linguistics; dialectology; processing of psycholinguistic and sociolinguistic data; archives and banks of texts and lexical information; etc.

The selections of papers and the themes for panel discussion will be made by a board of consultants coordinated by A. Zampolli.
The working languages will be French and English.

For each paper 40 minutes will be reserved, discussion included. Abstracts (1000 words minimum) or the complete text must be submitted by February 1, 1978 to:

Prof. A. Zampolli,
Chairman of Scientific Program Committee
COLING 78
Via S. Maria 36
56100 PISA ITALY
Tel. (050) 45245 – Telex: 50371 – CNUCE