humanistiske data

Utgitt av NAVF.s EDB-senter
for humanistisk forskning

The Norwegian Computing Centre
for the Humanities

NORGES ALMENVITENSKAPELIGE FORSKNINGSRÅD

Artikler
Konferanserapporter
Meldinger
Summary

NR.
1—2
1979

Rekv

INNHOLD

MEDARBEIDERE I DETTE NUMMER:

Ivar Fonnes, amanuensis, EDB-tjenesten v/ HF, Universitetet i Oslo

Stig Johansson, dosent, Britisk institutt, Universitetet i Oslo

Knut Kleve, professor, Klassisk institutt, Universitetet i Oslo

Erik Schia, prosjektleder, Gamlebyprosjektet, Riksantikvaren og Universitetets Oldsaksamling, Oslo

Roald Skarsten, EDB-konsulent, EDB-seksjonen v/ HF, Universitetet i Bergen

Gunnar Thorvaldsen, p.t. EDB-konsulent (NAVF) v/ Institutt for språk og litteratur, Universitetet i Tromsø

Redaksjonen avsluttet 1.12.1979

HUMANISTISKE DATA PÅ NY

Etter et opphold på 2 år sender vi nå ut et nytt nummer av Humanistiske Data. Vi håper at bladet vil bli vel mottatt og at det fremover kan bli et kontaktpunkt for dem som er interessert i databehandling i de humanistiske fag.

Humanistiske Data presenterer seg nå i en ny og omarbeidet skikkelse. Overgang til et nytt format og ny fremstillingsmåte er foretatt for å gjøre produksjonen enklere. Derved skulle forholdene være lagt til rette for en regelmessig utgivelse i årene som kommer. Antall nummer pr. år og sidetallet vil kunne variere alt etter stofftilgangen, men vi tar sikte på mellom 2 og 4 utgivelser årlig.

Når vi nå sender Humanistiske Data ut igjen, har vi sløyfet betegnelsen "meldingsblad". Dette er gjort for klarere å markere Humanistiske Data som talerør også for andre personer og miljøer enn NAVF's EDB-senter og dets personale.

Vi håper blant annet fremover å kunne bringe regelmessige oversikter over humanistisk EDB-arbeid ved universitetene gjennom bidrag ikke

minst fra EDB-tjenestene ved HF-fakultetene.

Dessuten ønsker vi at forskere og fagmedarbeidere som bruker EDB i sitt arbeid, orienterer om sin virksomhet, gir innblikk i faglige miljøer de har kontakt med i utlandet, eller orienterer om aktuell litteratur. Reaksjoner på stoff som blir presentert i bladet ønskes også velkommen.

Utgiveren vil på sin side orientere om arbeidet i senteret, rapportere fra konferanse- og kontaktvirksomhet, gi meldinger om nye hjelpemiddel i form av programmer og utstyr og rapportere fra prosjektarbeid.

I dette nummer forsøker vi å skaffe fram en oversikt over det EDB-arbeid som i dag drives i vårt land innenfor de humanistiske fag gjennom bruk av et spørreskjema. Vi ber om at så mange som mulig tar seg tid til å fylle ut skjemaet og sende det tilbake. Resultatet vil komme alle til gode gjennom et oversyn i bladet.

Humanistiske Data sendes ut gratis. De som ønsker bladet tilsendt og som ikke har fått det før, kan gi melding til NAVF's EDB-senter.

STATUTTER FOR NAVF'S EDB-SENTER FOR HUMANISTISK FORSKNING.

NAVF's EDB-senter for humanistisk forskning ble startet i 1972 som et 5-års utviklingsprogram. Fra 1978 av er senteret opprettet som et permanent EDB-organ under NAVF. I forbindelse med reorganiseringen ble det vedtatt nye statutter for senteret. En del av disse refereres nedenfor:

OPPRETTELSE

NAVF's EDB-senter for humanistisk forskning, (Sentret), er opprettet som organ under NAVF fra 1.1.78. Rådet for humanistisk forskning har den faglige og administrative myndighet over Sentret.

Sentret er underlagt NAVF's vedtekter og andre bestemmelser som gjelder for NAVF med mindre annet er bestemt.

FORMÅL

NAVF's EDB-senter for humanistisk forskning har som målsetting å legge forholdene til rette for en fornuftig bruk av EDB i humanistisk forskning og utviklingsarbeid. Sentret skal stimulere og hjelpe enkeltforskere og fagmiljøer til å bruke EDB. I samarbeid med fagmiljøene skal Sentret videreutvikle EDB-metoder med sikte på å utvikle de lokale miljøer slik at de blir stadig mer selvstendige i EDB-arbeidet. Sentret skal arbeide for en hensiktsmessig koordinering av EDB-tiltak. En hovedoppgave for Sentret er å bidra til at behovet for nasjonale EDB-tjenester i humanistisk forskning blir dekket.

Sentret skal bidra i utformingen av NAVF's EDB-politikk.

ARBEIDSOPPGAVER

Sentret skal særlig arbeide med følgende oppgaver:

1. drive informasjons- og opplæringsvirksomhet innen de enkelte fagområder og på forskjellige nivåer om bruk av datamaskiner i

humanistisk forskning.

2. gi konsulenttjenester i tilknytning til humanistiske forskningsprosjekter som gjør bruk av datamaskin,

3. gi programmerings-,maskin- og dataregistreringstjenester innenfor sin kapasitet,

4. skaffe til veie, tilrettelegge eller utvikle programutrustning for humanistisk forskning på nasjonal basis,

5. delta i prosjektsamarbeid,

6. bidra - i samarbeid med fagmiljøene - til en systematisk oppbygging av dataarkiv ved å utarbeide retningslinjer og metoder for slike arkiv. Sentret skal medvirke til at flest mulig får adgang til arkivene,

7. ta hånd om og vedlikeholde verdifullt humanistisk forskningsmateriale i EDB-prosjekter der dette er nødvendig for å sikre materialets senere bruk. Målet er å overføre denne funksjon til de ordinære oppbevaringsinstitusjoner så snart forholdene ligger til rette for det,

8. samarbeide med de humanistiske EDB-tjenester ved universitetene og med EDB-organer under NAVF, særlig Norsk samfunnsvitenskapelig datatjeneste,

9. holde seg løpende orientert om utviklingen i humanistisk databehandling såvel nasjonalt som internasjonalt og holde de humanistiske forskningsmiljøer og NAVF regelmessig informert,

10. holde kontakt med de humanistiske brukermiljøer, bl.a. gjennom arrangement av nasjonale konferanser og faglige brukermøter,

11. utføre andre oppgaver innen sitt ansvarsområde som NAVF pålegger Sentret.

LOKALISERING

Sentret skal være plassert i tilknytning til et universitet eller annet forskningsmiljø.

FINANSIERING

Sentret finansieres av NAVF gjennom bevilgninger fra Rådet for humanistisk forskning.

I tillegg kan Sentret finansiere deler av sin virksomhet gjennom betalte oppdrag fra andre organer under NAVF eller institusjoner utenfor NAVF. Planer om oppdragsvirksomhet skal forelegges Sentrets styre og deretter NAVF.

STYRE

Styret har for tiden følgende sammensetning:

Medlemmer:

Førstekonservator *Kjell Falck* (formann)
Dosent *Ådne Findreng*
Instituttstyrer *Sofie Rogstad*
Avdelingsleder *Carl. E. Ellingsen*
Førstebibliotekar *Nils L. Gram*
Adm. leder *Jostein H. Hauge* (sekretær)
Driftsass. *Per Vestbøstad* (representant for de ansatte)

Varamedlemmer:

Oppdragsleder *Kjell Fredriksen*
Riksbibliotekar *Gerhard Munthe*
Prof. *Gunnar Skirbekk*

IBSEN-KONKORDANS

Ved NAVF's EDB-senter har det fra 1.4.1978 vært i gang et prosjekt med det mål å utarbeide lemmatiserte konkordanser til Ibsens skuespill og dikt. Professor *Harald Noreng* er ansatt som faglig leder i full stilling, og det er engasjert personale til driftsoppgaver og vitenskapelig assistanse.

EDB-oppgavene i prosjektet utføres ved NAVF's EDB-senter. Det er etablert èn nasjonal styringsgruppe for prosjektet hvor professor *Bjarte Birkeland*, Nordisk institutt, Universitetet i Bergen er formann. Prosjektet,som finansieres av NAVF, skal avsluttes i løpet av 1. halvår 1981.

Det er bestemt at det skal utarbeides fullstendige konkordanser til hvert enkelt av Ibsens skuespill og til hans Digte. Som grunnlag anvendes Hundreårsutgaven av 1928. Tekstene er punchet dels i Cambridge, dels ved NAVF's EDB-senter, og et ikke uanselig antall trykkfeil blir rettet. I konkordansene blir det gjennomført homograf-separering og lemmatisering (ordene henføres til oppslagsform). Hvert enkelt ord Ibsen anvender, blir gjengitt hver gang det forekommer, innenfor en individuell og avpasset kontekst.

Det enkelte ord blir stilt i relasjon til den grunnform som kan stilles opp, med basis i den ortografi Ibsen valgte å følge etter rettskrivningsmøtet i Stockholm i 1969. Det gis opplysninger om ordklasse og bøyningsformer, og i mange tilfelle om ordets skrivemåte i moderne norsk.

Konkordansene angir også om det enkelte ord opptrer i replikk eller sceneanvisning, i prosa eller i versifisert språk, og for replikkenes vedkommende også hvilken av personene i dramaet som avleverer dem, og hvem som mottar dem. Når det enkelte ord står i rimposisjon, blir det opplyst hva slags rim (parrim, kryss-rim, omsluttende rim o.s.v.) ordet er en del av.

Endelig vil det i konkordansen under hvert ord og hver ordform bli tilføyd oppgaver over ordets eller ordformens totale og relative frekvens innenfor det enkelte verk.

På lengre sikt er det tanken å støpe verkkonkordansene sammen til en stor helhetlig konkordans over Henrik Ibsens produksjon.

Ved utarbeidelse av de lemmatiserte konkordansene tar en i bruk både manuelle og automatiserte rutiner. Det første skuespillet ble lemmatisert manuelt ut fra en KWIC-konkordans med 3 linjers kontekst. Det blir gitt et maskinelt forslag til kontekstavgrensing basert på skilletegn og rimmarkeringer. Forskeren har mulighet til å endre kontekstforslaget slik at hvert belegg får en nøye tilpasset kontekst. Ved lemmatiseringen av de påfølgende verk bygger en på den lemmatisering som tidligere er gjort. Det er utviklet datamaskinelle metoder som bl.a. medfører at det blir bygd opp en ordbok i datamaskinen over de godkjente grammatiske bestemmelser til ordene. Datamaskinen anvender så denne ordboka ved utarbeidelse av en lemmatisert konkordans til det neste verk. Dette forslaget blir gjennomgått av forskeren som foretar korreksjoner og tilføyelser. Disse blir deretter påført den maskinlagrede konkordansen via dataskjerm.

Ved det videre arbeid vil en automatisere lemmatiseringen ytterligere. Det vil da bli tatt i bruk et sett med bøyningsparadigmer for Ibsens språk, en tabell over ordklasser basert på endelser og en serie kontekstregler for å øke antallet rette lemmamarkeringer. Disse reglene blir satt opp bl.a. på grunnlag av de verk som er ferdig lemmatisert.

Pr. i dag har alle tekstene gjennomgått en nøye kontroll i flere korrektur-omganger, og det foreligger lemmatiserte konkordanser til en serie skuespill.

En slik konkordans vil kunne bli til stor nytte for både språk- og litteraturforskeren. Språkforskeren vil finne opplysninger om Henrik Ibsens ortografi og ordvalg, om hans grammatikk og syntaks, og vil kunne finne fram til både sammenhengen og utviklingen innefor Ibsens språk gjennom et halvt hundre år. Litteraturforskeren vil i konkordansene finne hjelpemidler til studiet av de enkelte ords plass innenfor mer og mindre faste uttrykk, sammenligninger, bilder og symboler. Konkordansene vil lette arbeidet med å foreta

språklige og litterære sammenligninger mellom de enkelte Ibsen-verk. I den grad der foreligger konkordanser over eldre og mer samtidige litteraturverk (f.eks. Bibelen, Søren Kierkegaards skrifter o.s.v.), vil de planlagte Ibsen-konkordanser kunne bli til god hjelp for forskere som vil arbeide med problemer i forbindelse med Henrik Ibsens forhold til fortid og samtid.

Den ferdige og fullstendige Ibsen-konkordans vil bl.a. kunne danne grunnlag for et Ibsen sitat-leksikon og en ny stor Ibsen-ordbok.

ICAME

INTERNATIONAL COMPUTER ARCHIVE OF MODERN ENGLISH

ICAME er en interesseorganisasjon med formål å samordne interessene hos forskere som gjør bruk av datamaskinlagret materiale fra moderne engelsk språk. Organisasjonen ble dannet i 1977 og har følgende oppgaver:

1. Samle og spre informasjon om engelsk språkmateriale til bruk i datamaskin.

2. Samle og spre opplysning om den lingvistiske forskning som planlegges eller er utført på slikt materiale.

3. Legge opp tekstsamlinger ved NAVF's EDB-senter for humanistisk forskning i Bergen til distribusjon for forskere i inn- og utland.

ICAME drives av en gruppe forskere bestående av:

Prof. *W. Nelson Francis*, Brown University, USA
Dosent *Stig Johansson*, Britisk institutt, Universitetet i Oslo

Prof. *Geoffrey Leech*, University of Lancaster, England
Prof. *Jan Svartvik*, Lunds Universitet, Sverige

Dosent Stig Johansson er faglig leder av ICAME. NAVF's EDB-senter er det operative EDB-organ.

ICAME NEWS er navnet på et meldingsblad som gir opplysninger om tilgjengelige tekster i de ulike forskningsmiljøene og informasjon om de tekstsamlinger som ICAME forvalter og distribuerer (siste nummer utkom i oktober 79). Bladet er gratis og kan bestilles hos redaktøren, dosent Stig Johansson.

I løpet av høsten er følgende tekstmateriale tilgjengelig fra ICAME/ NAVF's EDB-senter:

1. *The Brown Corpus*.

   Samlingen omfatter 1 mill. ord hentet fra amerikansk-engelsk bruksprosa og skjønnlitteratur. Det foreligger mikrokort- og magnetbåndversjon med store og små bokstaver av konkordans og grunntekster.

2. *The Lancaster-Oslo/Bergen Corpus (LOB)*.

   Dette er et britisk-engelsk motstykke til Brown Corpus basert på 1 mill. ord fra tekster utgitt i 1961. Prosjektet er avsluttet i Norge gjennom et samarbeid mellom dosent Stig Johansson og NAVF's EDB-senter.

3. *The London-Lund Corpus*.

   Dette maskinlagrede talespråksmaterialet er resultat av arbeid utført ved Engelska Institutionen, Lunds Universitet under ledelse av prof. Jan Svartvik. Tekstgrunnlaget er hentet fra The Survey of English Usage, University College, London. Samlingen omfatter 170.000 ord fra spontan tale.

Utførligere opplysninger om disse tekstsamlingene og den bearbeidelse

som er foretatt av dem, finnes i ICAME nr. 3, oktober 79.

Det vises også til separat referat i dette nummer av Humanistiske Data fra "A Symposium on Grammatical Tagging of English Text Corpora" holdt i Bergen, 29. og 30. mars 1979.

NORSK TEKSTARKIV

Det har i lang tid vært et ønske å få bedret tilgangen på opplysninger om moderne norsk språk og tekstsamlinger av moderne norsk som grunnlag for språklig og stilistisk analyse og til bruk i undervisningssammenheng. Dette er behov som har vært reist i like stor grad utenfor som innenfor universitetene.
Bruk av datamaskin ved utgivelse av bøker, aviser og tidsskrifter gir i dag helt nye muligheter til å fange opp aktuelt materiale på et datamaskinelt lagringsmedium.

For å få drøftet denne saken inviterte NAVF's EDB-senter for humanistisk forskning i oktober 78 et representativt utvalg av interesserte til en 2 dagers konferanse i Bergen. Som det går fram av konferanserapporten som ble utarbeidet, ble det her en omfattende og nyansert debatt om saken med utgangspunkt i presentasjoner av pågående arbeid og de ulike behov for språkdata. Resultatet ble at det fremlagte forslag om å legge til rette et standard norsk tekstkorpus ble modifisert. I overensstemmelse med tilrådingen på konferansen, ble det nedsatt en planleggingsgruppe, som har utarbeidet planer for et norsk datamaskinelt tekstarkiv.

Følgende personer og institusjoner har vært med i planleggingen:

Adm. leder *Jostein H. Hauge*, NAVF's EDB-senter
Førsteamanuensis *Kolbjørn Heggstad*, Nordisk institutt, PDS, UiB
Førstekonsulent *Aagot Landfald*, Norsk språkråd

EDB-konsulent *Eirik Lien*, EDB-tjenesten for humanistiske fag, Trondheim

Prof. *Egil Pettersen*, Nordisk institutt, UiB (formann)

Aman. *Jarle Rønhovd*, Nordisk institutt, UiTrheim

Aman. *Dagfinn Worren*, Norsk leksikografisk institutt

Norsk tekstarkiv har som mål å koordinere og øke innsatsen i arbeidet med å samle inn og tilrettelegge tekstmateriale fra moderne norsk til bruk i forsknings- og utviklingsarbeid. Tiltaket vil fra starten av bli nasjonalt orientert.

En forutsetning for at Norsk tekstarkiv vil resultere i en viten-ressurs om norsk språk, er at arbeidet med tekstinnsamling legges opp etter en nasjonalt koordinert plan og at materialet tilrette-legges på en standard måte. De datamengder som legges opp, må kunne utnyttes datamaskinelt i alle interesserte miljøer med et minimum av ekstra tilretteleggingsarbeid.

Basis for tiltaket vil være et formalisert samarbeid mellom PDS, Nordisk institutt, Universitetet i Bergen og NAVF's EDB-senter i Bergen, som igjen har EDB-tjenestene ved HF-fakultetene ved univer-sitetene som faste samarbeidspartnere.

Rådet for humanistisk forskning, NAVF, har for 1980 blant annet bevilget midler til en driftsmedarbeider som kan knyttes til Norsk tekstarkiv i full stilling i 5 år. Noe av virksomheten tenkes forøvrig finansiert gjennom betalte oppdrag.

Gjennom en egen bevilgning til Norsk tekstarkiv fra Universitetet i Bergen i 1979 vil det bli foretatt en utgreiing om standard-format for lagring av norske tekstdata.

Forholdene skulle således ligge godt til rette for ordinær drift av tekstarkivet fra vinteren av. En nærmere presentasjon vil bli gitt i et senere nummer av bladet

NORSK LANDBRUKSORDBOK - DEFINISJONSORDBOK OG DATABANK

Ivar Fonnes

I februar i år ble Norsk Landbruksordbok utgitt på Det Norske Samlaget. Boken er resultatet av mer enn 20 års arbeid med samling og systematisering av landbruksterminologi under ledelse av redaktøren, dosent Magne Rommetveit. Når den nå foreligger i trykt form utgjør landbruksordboken en av de største og mest omfattende definisjonsordbøker innen landbruksterminologi. Ved siden av termer på nynorsk og bokmål finner man angitt synonymer på inntil 6 andre språk - samisk, svensk, dansk, engelsk, tysk og islandsk (og dessuten finske synonymer i registeret).

Verket foreligger i to bind. Bind 1 (ca. 580 sider) inneholder selve definisjonsordboken. Den omfatter ca. 18 000 definerte og ca. 5000 udefinerte termer. Videre finner man ca. 5000 vitenskapelige navn (plantenavn m.v.) samt synonymer på inntil 6 språk, om lag 100 000 i alt.

Bind 2 (ca. 400 sider) er et registerbind over utenlandske synonymer og vitenskapelige navn. For hvert språk er det utarbeidet et alfabetisk register med referanse til de norske oppslagsord i bind 1. Et tilsvarende register for finsk er også med, selv om ikke finske synonymer er angitt i selve ordboken. Til sammen inneholder registerbindet ca. 120 000 oppslag.

Samtidig med at ordboken nå er utgitt i bokform foreligger det en tro kopi av materialet i en databank. Databanken er et produkt av arbeidet med å tilrettelegge ordboken for trykkking. Hele materialet ble skrevet inn i datamaskinlesbar form, og på dette grunnlag ble det produsert magnetbånd ("drivetape") med data klare

for kjøring i fotosetter. Vi har ikke tidligere anvendt EDB-basert tilrettelegging og fotosetting av slikt ordboksmateriale, og prosjektet har derfor i en viss utstrekning vært preget av forsøksarbeid. Av spesiell interesse har det vært å finne fram til hensiktsmessige metoder for håndtering av et stort og meget hyppig vekslende tegnrepertoar. Materialet er også av betydelig størrelse og utgjør den største databank med terminologisk materiale i Norge i dag.

EDB-prosjektet for etablering av en databank og trykking av Norsk landbruksordbok ble finansiert av NAVF og initiert av NAVF's EDB-senter for humanistisk forskning. EDB-arbeidet har vært planlagt og ledet ved EDB-tjenesten ved HF, Universitetet i Oslo og utført i samarbeid med Norsk leksikografisk institutt. Databehandlingen har foregått ved Universitetet i Oslo.

EDB i produksjonsarbeidet.

Det meste av arbeidet med utarbeidelsen av Norsk landbruksordbok var utført før spørsmålet om bruk av EDB ble tatt opp. Først da manuskriptet skulle klargjøres for trykking, ble det vurdert om EDB kunne være et hensiktsmessig hjelpemiddel i tilretteleggingen. Etter en del utredningsarbeid valgte man å satse på fotosetting via EDB, først og fremst av økonomiske årsaker. Men det ble også lagt vekt på at materialet ville ha større bruksverdi dersom det forelå i maskinlesbar form.

Når databehandlingen kunne bidra til å redusere kostnadene, skyldtes dette at datamaskinen kunne overta en del arbeidskrevende operasjoner i det avsluttende redigeringsarbeid og i tilretteleggingen for trykking. For det første kunne innskrivingen av data (punchingen) også fungere som renskriving av manuskriptet. Dermed behøvde man skrive materialet bare én gang. Når data var innlest i maskinen og korrekturarbeidet utført, var det en helt automatisert prosess fram til ferdig sats. Ved tradisjonell trykking måtte man først ha renskrevet manuskriptet, og deretter ville trykkeriet ha måttet skrive det hele om igjen for å framstille sats. Man ville også som følge av dette ha fått to omganger med korrekturarbeid.

For det andre kunne datamaskinen anvendes til å ordne oppslagene i alfabetisk rekkefølge. Før trykkeprosjektet var startet var materialet ordnet etter fagområde og alfabetisert innen hvert fag, mens det i den trykte utgaven skulle være ordnet i ett alfabet. En manuell sortering av ca. 25 000 oppslag ville være en møysommelig og tidkrevende oppgave. Datamaskinen kunne utføre arbeidet i løpet av et par timer.

Den tredje store oppgaven var å produsere registrene til ordboken, dvs. bind 2 i den trykte utgaven. Dette skulle gjøres ved å trekke ut alle utenlandske synonymer som er angitt i ordartiklene, samle dem i en liste for hvert språk med de norske oppslag som referanser, alfabetisere og bearbeide dem til registre samt klargjøre for trykking. Det aller meste av dette arbeidet ble utført av datamaskinen. Bare i tilfeller med synonymer på mer enn ett ord måtte det redaksjonell kontroll til for å bestemme hvilket ord som skulle stå først og utgjøre oppslag i registeret.

Det er vanskelig å anslå hvor mye man har spart ved bruk av EDB på disse oppgavene. Sikkert synes det imidlertid at synonymregistrene ikke ville ha kunnet produseres innen rimelig tid og realistiske kostnader uten bruk av EDB. Når registerbindet nå foreligger samtidig med hovedmaterialet er dette således en direkte gevinst av den produksjonsmåte som ble valgt. Datamaskinen ble også brukt til en del kontrollarbeid i forbindelse med den avsluttende redigering. Slikt automatisert kontrollarbeid kunne det vært ønskelig å utføre i langt større utstrekning, men økonomien tillot ikke det.

Nå er det imidlertid grunn til å understreke at bruk av datamaskin hverken er gratis eller uavhengig av arbeidsinnsats fra brukeren. Det koster penger å anvende datamaskinen, og det koster en god del arbeid å organisere og behandle såvidt store datamengder. Men framfor alt krever det arbeid å utarbeide systemopplegg og programmer som kan fortelle datamaskinen hva den skal foreta seg med data. I dette prosjektet har vi i liten utstrekning kunnet basere oss på ferdige tekstanalyseprogrammer p.g.a. materialets spesielle karakter og de oppgaver som skulle utføres. Vi har utarbeidet eget systemopplegg og egne programmer for prosjektet.

Denne investering vil imidlertid også komme til nytte i annet ordboksarbeid i og med at opplegg og programmer med visse justeringer kan benyttes (og blir benyttet) i lignende prosjekter. Likeledes har prosjektet gitt verdifull erfaring og øket kompetanse i bruk av EDB i leksikografisk arbeid og tilrettelegging for fotosetting.

Utnyttelse av databanken.

Databanken er som nevnt en tro kopi av ordboksmaterialet slik det foreligger i trykt utgave. Den maskinlesbare versjon gir imidlertid flere muligheter for utnyttelse. Den er f.eks. velegnet som grunnlag for en videre bearbeidelse av materialet. Det er planlagt å føre videre arbeidet med Norsk landbruksordbok med sikte på utvidelser, justeringer, utbygging av synonymapparatet m.v. Slike forandringer/tillegg kan lett føres inn i databanken, og denne vil da til enhver tid være ajour i forhold til det faglige arbeid som er gjort. Dessuten vil man lett kunne trekke ut deler av materialet for publisering.

I tilknytning til det videre arbeid med materialet er det også planlagt en avtale med EF-kommisjonens oversettelsesavdeling (Luxembourg) om datautveksling. I EF-kommisjonens termbank er landbruksterminologi en viktig del, og der er betydelig interesse for å få adgang til materialet i landbruksordboken. Fra vår side er det av interesse å få påført franske synonymer og få adgang til landbrukstermer som ikke finnes i ordboken i dag. Begge deler vil bidra til å øke materialets verdi.

Materialet i databanken vil også være av verdi for annet terminologisk arbeid innen fagområder som er med i Norsk landbruksordbok. På grunnlag av fagmerkinger i ordartiklene kan datamaskinen f.eks. skrive ut alle termer innen bestemte fag, eller eventuelt produsere et nytt datasett med bare disse termene. Likeledes er det grunn til å anta at materialet i databanken vil kunne inngå i og utgjøre en viktig del av en mer generell norsk termbank.

I språkvitenskapelig forskning som benytter materialet vil databanken gi betydelige tilleggsmuligheter i forhold til den trykte utgave. Informasjonssøking etter fenomener som ikke kan finnes ut fra oppslagsordenes alfabetiske rekkefølge, kan gjøres meget effektivt ved bruk av databehandling. Likeledes er materialet direkte tilgjengelig for å anvende EDB i kvantitative analyser.

Disposisjonsretten over databanken er tillagt NAVF's EDB-senter for humanistisk forskning. Det arbeides nå med retningslinjer for hvordan materialet kan utnyttes av andre.

EDB SOM HJELPEMIDDEL VED DE ARKEOLOGISKE UTGRAVINGER I GAMLEBYEN, OSLO

Av Erik Schia.

Innledning.

I 1970 begynte middelalderarkeologiske byutgravinger i Gamlebyen, som forprosjekt for motorveiutbygginger i bydelen. I alt er det undersøkt ca. 1400 m^2 med kulturlag i tykkelse fra 1-3 m, inneholdende gater, brønner, ca. 350 bygninger, brannlag, flislag etter gjenoppbygging (i alt ca. 1700 forskjellige jordlagsnummer) og gjenstander fra hverdagslivet med ca. 30 000 registrerte funnnr.

Under utgravingsarbeidet rådet en forsiktig skepsis til bruken av EDB. Manuelle systemer med krysskataloger og bl.a. funnkort i 3 eksemplarer for sortering på forskjellige måter, ble derfor utviklet. Det manuelle systemet fungerer, men det er tidkrevende for en del oppgaver.

I det manuelle systemet ble det tidlig bestemt at funnkortene skulle renskrives med skrivemaskin. Da maskinskrivingen og funnkortene fra "Søndre felt" begynte i 1977, valgte vi imidlertid å legge opp arbeidet slik at funnkortene også skulle kunne leses optisk. Det ville på den måten være mulig å legge inn opplysninger om gjenstandene i en database på et seinere tidspunkt om ønskelig. I motsatt fall ville vi for lang tid bli fastlåst til det manuelle systemet. Dermed var "EDB-snøballen" i realiteten begynt å rulle, og vi regner nå med at alle funnkortene fra "Søndre felt" (ca. 16800) vil være ferdig maskinskrevet for optisk lesing og lagret på magnetbånd i to databaser tidlig i 1980.

Fra en første skepsis til EDB, har vi nå tatt spranget fullt ut, takket være mulighetene for optisk lesing og tekstsøkesystemet NOVA*STATUS. Vi er optimistiske og vil bruke gjenstandsmaterialet

fra *"Søndre felt"* som et prøveprosjekt, for å innhente erfaring.

Denne EDB-prøvingen har i dag et kortsiktig og et langsiktig perspektiv for oss. Det kortsiktige perspektivet gjelder bruken av EDB i et forskingsprosjekt om de utgravde data fra Gamlebyen i Oslo. Det langsiktige perspektivet gjelder en tilrettelegging av EDB-bruk ved framtidige utgravinger. Vi samarbeider nært med NAVF's EDB-senter, som utfører alle EDB-oppgavene i prøveprosjektet.

Forskingsprosjektet og det kortsiktige EDB perspektivet (1979 - 1983).

Høsten 1978 bevilget NAVF midler til en helhetsanalyse av de innsamlete arkeologiske data i Gamlebyen fra 1970-1976. Hovedproblemstillingene i dette prosjektet er:

1) Datering av de ulike fasene ved hjelp av keramikk, sko og kammer.
2) Urbaniseringsprosessen i Oslo belyst ved de undersøkte områder.
3) Bygårdenes funksjon, bygninger, eiendomsgrenser etc.
4) Ervervsliv som husdyrhold, fiske, håndverk, husflid/hjemmesysler, handel.
5) Levestandard, materiell og åndelig kultur som hygiene, sosiale forhold etc.

På grunn av den store funnmengde, ca. 30 000 gjenstander, var det nødvendig å dele opp materialet i mindre enheter og fordele det på i alt 19 medarbeidere. Disse vil ta opp delspørsmål i prosjektet i sammenheng med hovedproblemstillingene, og det blir prosjektledernes (Petter B. Molaug og Erik Schia) oppgave å sammenfatte det hele i et avsluttende syntese-bind. Det er meningen prosjektet skal avsluttes i 1983. I dette arbeidet tror vi EDB vil være til stor hjelp når det gjelder ulike typer spørsmål og materialgrupper. Vi vil og få anledning til å sammenlikne EDB-bruk med tradisjonell manuell metode idet et utgravingsfelt, *"Mindets tomt"*, ikke blir overført til EDB, fordi funnkort herfra allerede var maskinskrevet med feil kulehode da spørsmålet ble

aktuelt. Fra *"Mindets tomt"* er det i alt registrert ca. 10 000 funn-nr.

De data fra *"Søndre felt"* som vil bli lagret i databasen, gjelder funnforhold og gjenstandsbeskrivelse. Gjenstandsbeskrivelsen er i hovedsak svært generell og fyller sjølsagt ikke de krav som stilles av de enkelte forskere i prosjektet. Slike data kan først etableres etter den vitenskapelige bearbeiding av hver funngruppe og kan eventuelt leses inn seinere som supplement/ rettinger i databasen. To materialgrupper, keramikk og lær, danner imidlertid unntak, idet den vitenskapelige analysen her ligger foran maskinskrivingen av funnkortene. Disse funnkortene blir dermed rettet opp og inneholder data som er av interesse for forskingsprosjektets problemstillinger.

For keramikkens del (behandles av *Molaug*) vil EDB være et nyttig hjelpemiddel for ulike summeringsoppgaver og spørsmål gjeldende kombinasjon av to eller flere elementer. F.eks. funksjonstyper innenfor de ulike keramikkgrupper (se bind I i serien om de arkeologiske utgravinger i *Gamlebyen*), og eventuelt om det kan sees konsentrasjoner av bestemte keramikk-typer til bestemte bygårder/ bygninger. Til stor hjelp for dateringsspørsmålet er en maskinell utlisting av de ulike keramikkgrupper ordnet etter kronologiske faser.

Tilsvarende spørsmål vil og bli aktuelle for behandlingen av skomaterialet (*Schia*) og kanskje spesielt i sammenheng med læravfallet som er inndelt i 3 kategorier og teller i alt ca. 100 000 enheter. En EDB utlisting med plassering av de ulike kategorier til faser i utgravingsfeltets koordinatsystem, for deretter eventuelt å tegne ut spredningskart, vil være svært arbeidsbesparende.

I sammenheng med prosjektets målsetting om helhetsanalyse av materialet, vil det være en fordel å kunne stille enkle spørsmål om funnfordeling av gjenstandsgrupper behandlet av ulike forfattere. Det er mulig at samarbeidet mellom de forskjellige forfatterne i

prosjektet vil kunne forenkles og at nye problemstillinger som reises under arbeidet med gjenstandsgruppene, kanskje lettere kan testes mot en annen gjenstandsgruppe når EDB tas i bruk. Oversikten over de ulike typer funnkombinasjoner og spredningskart vil ventelig framskaffes lettere med EDB enn ved manuelt arbeide.

I forskningsprosjektet har vi således forventninger om arbeidsbesparelse av rene rutineoppgaver ved EDB-bruk, i tillegg til at nye oppgaver av kvantitativ karakter vil kunne utføres. Rimeligvis vil derfor bruk av EDB også tilføre prosjektet kvalitative verdier.

Framtidige utgravinger i Gamlebyen og det langsiktige EDB-perspektiv.

På grunn av planer om motorvei der middelalder Oslo en gang lå, er det ventet store arkeologiske utgravinger i 1980 åra. I Oslo er det i dag forholdsvis små områder med intakte kulturlag igjen, og disse faller i tillegg hovedsakelig sammen med plasseringen av de prosjekterte motorveiene. Gjennomføres veiplanene kan i verste fall ca. 10 000 m^3 måtte graves ut og vår generasjon kan komme til å fjerne kanskje mesteparten av det som er igjen.

Utgravinger av en slik størrelsesorden vil ventelig frambringe store mengder data som vil bli vanskelig håndterbare. En overføring av den innsamlete informasjonen til en database og EDB-behandling av materialet, vil derfor tvinge seg fram, slik at det første etterarbeidet og den seinere vitenskapelige analysen av materialet kan forenkles.

På hvilken måte dette skal skje har vi ikke tatt stilling til ennå, og det er klart at erfaringene fra EDB-behandlingen av gjenstandsmaterialet fra *"Søndre felt"* her vil bli viktig.

Det vil og bli et spørsmål om hvordan gjenstandene skal katalogiseres, hvilken nomenklatur som skal brukes og i hvilken grad det vil være mulig å få med relevante spesialopplysninger. Forhåpentligvis vil forskningsprosjektet av de allerede utgravde data bidra med nye synspunkter på hva som er viktig for den EDB-orienterte katalogisering av gjenstandene i Oslo i 1980 åra.

REGISTRERINGSSENTRAL FOR HISTORISKE DATA

UNIVERSITETET I TROMSØ

Gunnar Thorvaldsen

1. KORT HISTORIKK.

Ved Institutt for samfunnsvitenskap, Universitet i Tromsø, har man siden høsten 1976 planlagt å opprette et historisk dataarkiv. Mønster for tiltaket er Demografiska Databasen i Norbotten i Sverige, som databehandler kirkebøker fra 1800-tallet.

Også i Norge har vi store kildeserier som først gjennom EDB blir reelt tilgjengelig for forskerne. Blant disse står folketellinger og kirkebøker fra forrige hundreår sentralt i flere forskermiljøer. Nå egner arbeidet med avskriving seg godt for desentralisering. Derfor kombinerte historikermiljøet de nasjonale behov med sysselsettingvanskene i utkantstrøk. Man innledet samarbeide med Utbyggingsavdelingen i Troms fylkeskommune.

Sentralens foreløpige styringsgruppe er utpekt av rådet ved ISV med dosent John Herstad som formann.

2. FAGLIGE MÅLSETTINGER.

Forskere innen en rekke fagområder søker nå i større grad enn før å trekke historiske funn inn i sitt forskningsfelt. De samfunn som var er interessante studieobjekter i seg selv. Og uten kjennskap til gårsdagens samfunn, kan forståelsen av dagens samfunn være vanskelig. Professor Kenneth Lockridge, University of Michigan, har beskrevet denne "nye historie" slik:

"The strongest trend in recent social science has been to seek longer series of relevant social data, series extending as far as possible into the historical past, in order to provide a firmer ground for theoretical insights and generalizations concerning the nature of social behavior. This movement began with the demographers and now includes economists, political scientists, and even social phsychologists. At stake is our understanding not only of such spesific processes as fertility control and economic development, but our whole conception of social change and of social modernization as these have been embodied in the history of western humankind."

"There is much that is fruitful in this "new history". If it is to be carried to completion it will be heavily dependent on materials such as those being provided by the Demografisk Databas in Sweden."

Det samme gjelder Norge: Vi vet lite om samfunnsmessige sammenhenger og endringer i historisk tid, om hvordan økonomiske og demografiske forhold, fenomener som yrkesrekruttering og organisering, flytting og klassedannelser gjensidig påvirket hverandre.

Mikrohistorie, dvs. historie på individnivå, har individet som kombinasjonsenhet for egenskaper forskningen er interessert i ved forskjellige typer analyser for å beskrive grupper, klasser og hele samfunn. Det kan hevdes at en slik tilnærmingsmåte har åpenbare metodiske fortrinn, idet den gir grunnlag for sikkerhet i generaliseringer, reduserer faren for nivåfeilslutninger og kan gi historen en spesiell dynamisk dimensjon ved at enkeltmennesker kan følges fra stadium til stadium i deres livsløp.

Slik mikrohistorie har vært applisert i Norge (bl.a. Kristiania-undersøkelsen), retter den oppmerksomheten mot massefenomener, eller som det har vært sagt, "mot de mange mer eller mindre anonyme aktører... det anonyme flertallets bidrag til det historiske forløp."

Teknisk har EDB-behandling av massedata muliggjort en langt mer inngående og sikrere beskrivelse av det brede folks sosialhistoriske utvikling enn tidligere gjennom nærstudium av de tusener av enkeltindivider som er registrert i folketellinger, kirkebøker, skattelister, stemmerettsmanntall, matrikler osv. Når forskningsprosjekter innen dette felt fortsatt er relativt få, henger det bl.a. sammen med at behandling og registrering av slike data er tid- og arbeidskrevende for forskerne.

På denne bakgrunn kan vi summere de grunnene som taler for å prioritere registrering av nominative historiske kilder.

1. Denne type data er tverrfaglige. I tillegg til samfunnsforskere og historikere har navnegranskere vist interesse for materialet. Dette går fram av innstillingene fra NAVF's primærkildekomiteer, i svarene på en brukerhenvendelse som registreringssentralen har foretatt, indirekte også av NAVF's prosjektkatalog "Humanistisk forskning". Interessen for individdata og forskning omkring dem er økende i mange land. Derfor skulle mulighetene for internasjonalt samarbeid være gode, særlig med Demografiska Databasen i Sverige.

2. Arkivtekniske grunner: Som følge av interessen for slektsgranskning er de nominative kildene utsatt for sterk slitasje. Ettersom forskningen i Norge desentraliseres, vokser behovet for spredning av kilder med tilknytning til lokalmiljøene.

3. Selv om man må regne med at trykte kilder i overskuelig fremtid kan behandles direkte med optisk dokumentleser, vil dette neppe være mulig med håndskrevet materiale. Det aktuelle nominative materialet må altså stadig gjøres maskinleselig ved avskrift.

4. I debatten om personvern har forskernes adgang til opplysninger om enkeltmennesker kommet i søkelyset. Hvis reglene om datavern innskjerpes for nyere materiale, kan man forvente at noen forskeres interesse vil svinge over mot historiske individdata.

De data man distribuerer til forskerne vil være av to hovedtyper. For det ene datalister hvor individene er sortert etter ulike kriterier, for det andre kodede kildeutgaver som er grunnlag for statistisk analyse.

Dette tilsvarer de to viktigste områdene for forskningsmessig

anvendelse av nominative listedata. For det første kan man følge enkeltindividers livsløp som utgangspunkt for kollektive biografier. For det andre at det foretas statistiske tverrsnittstudier av større befolkningsgrupper med utgangspunkt i enkeltkilder.

3. STATUS FOR PROSJEKTET.

Det har vært prøvedrift på prosjektet fra 1/8-78 med støtte fra NAVF, Distriktenes utbyggingsfond, Arbeidsformidlinga og Universitetet i Tromsø. Prøveprosjektet har registrert 12.205 individenheter fra folketellingene 1865 og 1875, samt kirkebøkene i mellomliggende tidsrom. Alt er registrert for optisk lesing og det meste to ganger. Dermed kunne overensstemmelse mellom de to versjonene være hovedkriterium for riktig avskrift. Sammenligningen foregikk maskinelt.

Kombinasjonen av optisk lesning og dobbel registrering har vist seg fullt anvendbar på nominativt materiale. Begrensede forsøk med registrering på mikromaskin viser imidlertid at denne metoden er raskere, men krever betydelig større investeringer i maskinvare.

Når det gjelder videre drift, foreligger detaljerte planer om et 3-årig pilotprosjekt i Utredning nr. 2 om Registeringssentral for historiske data. Heri inngår plan for samarbeide mellom prosjektet/Universitetet og Utbyggingsavdelingen i Tromsø fylkeskommune om finansiering av registeringsarbeidet. Planene har vært ført videre av utvalget til behandling av problemene for de ansatte i forbindelse med automatiseringen i Televerket (Myklevollutvalget).

4. ARBEIDSPLAN FOR PERIODEN 1978 TIL -81.

Arbeidet har hittil hovedsakelig bestått i systemering, programmering og dokumentasjon av registreringsrutiner, feilrettingsprosedyrer og utlistingsprogrammer. Man har også kommet godt igang med sorteringsrutiner.

I 1979 arbeider vi videre med registreringssystemer for folketellinger. Vi samarbeider om et system for maskinell koding av data som blir programmert ved NAVF's EDB-senter. Det skal også utvikles overgangssystemer for bruk av statistikkpakker.

I 1980 vil hovedvekta ligge på å videreutvikle registeringsprosedyrer for kirkebøker. Det er planen å implementere programmer for standardising av personnavn.

1981 skal vies registrering av 1910-folketellinga. På programmeringssida vil vi starte arbeidet med systemer for maskinell lenking av kildene. Utgangspunkt er fødselsdatoene i 1910-tellingen og dåpslistene.

5. VALG AV KILDEMATERIALE.

Registreringssentralen vil etter planene i tida 1979-81 kunne registrere 1.2 til 1.5 mill. individenheter fra norske 1800-talls-

kilder. For å gi forskerne optimal hjelp med et flertall forskningsoppgaver, er det ønskelig å behandle mer enn en nominativ kildegruppe. På den annen side må utvalget begrenses fordi RHD har en relativt liten ledelsesenhet. Forskningsmiljøene har hatt anledning til å uttale som om følgende løsning:

1. år: ca. 400.000 enheter fra folketellingene 1865-1900.
2. år: ca. 400.000 enheter fra kirkebøkene 1800-1900.
3. år: ca. 500.000 individer fra folketellinga 1910, forutsatt at denne frigis.

Valg av kilder er nærmere begrunnet i Utredning nr. 2 om RHD, og sentralen har ikke mottatt viktige innvendinger. Den endelige avgjørelse vil bli truffet av det styringsorgan som skal velges så snart det er bestemt at dataarkivet kan etableres utover februar 1979. Hvordan kan man sikre at forskernes prioriteringer blir bestemmende når endelig beslutning fattes? For det ene vil deres representanter ha flertall i RHD's styringsorgan. For det andre må forskerne bli holdt løpende informert og få anledning til å uttale seg. Endelig er mulighetene åpne for å registrere kilder fra ad hoc områder etter spesielle ønsker fra igangværende eller planlagte prosjekter. Det innebærer naturligvis at man må kutte i kjerneområdene. Av disse grunner er det viktig at miljøene så snart som mulig informerer oss om hvilke regioner de vil ønske behandlet både innenfor kjerneområdene og ellers.

Det relativt differensierte utvalget av kilder forutsetter at man begrenser seg til noen geografiske områder. I en brukerhenvendelse sendt ut i juni 1978 ble 4 regioner foreslått. Med utgangspunkt i brukermiljøenes reaksjoner har vi revidert forslaget.

1. Nord-Norge. Valget av Midt- og Nord-Troms ligger fast, med ialt 24.364 innbyggere i 1865. Mht. folketellinga 1910 må samarbeidet med Demografiska Databasen om migrasjonsstudier på Nordkalotten fremdeles veie tungt. Derfor opprettholdes forslaget om å dekke hele Troms og Finnmark (119.800 personer).

2. Midt-Norge. Forskere ved Universitetet i Trondheim som orienterer seg mot historiska individdata, ønsker Øvre og Nedre Stjørdalen pluss Selbu m/Tydal. Her bodde i 1865 tilsammen 18.528 innbyggere. Begge områder er forholdsvis klart avgrensede i forhold til nærliggende distrikter. Mens Stjørdalen er variert med bl.a. industriutbygging på 1800-tallet, ble Selbu tvert imot mer jordbruksdominert. Komparative studier med utgangspunkt i EDB-materialet kan her bygge videre på tidligere undersøkelser med mer tradisjonelle metoder.

For 1910-tellingas vedkommende er det aktuelt å dekke Sør-Trøndelag med unntak av Fosen fogderi (108.124 innbyggere). Samtidig med at mange av migrantene fra kjerneområdet kan fanges opp, dekkes Norges tredje største by m/omegn. Samfunnsvitere der planlegger et "Trondheimsprosjekt".

3. Vest-Norge.

På Vestlandet kan de foreliggende demografiske studier av Etne og

Os være utgangspunkt for lignende undersøkelser av et større sammenhengende område. Man vil da oppnå resultater som er gyldige for migranter som forlot sin hjemkommune, men ble i regionen. Hvis valget faller på Sunnhordland er det også mulig å fange opp dem som dro til Bergen i folketellinga av 1875, eventuelt i emigrasjonsprotokollene. I tillegg til at området "lekker" i sør, er problemet at Sunnhordland er svært folkerikt, (33.695 innbyggere i 1865). Derfor er det naturlig at dette området "rammes" først ettersom ad hoc oppgaver må løses.

Siden 1910-tellingas fødselsdatoer vil gi størst gevinst i store kommuner, er det naturlig å inkludere Norges nest største by. Sunnhordland og Bergen hadde tilsammen 110.201 innbyggere på dette tidspunkt. Man ser hvilke muligheter som åpner seg for komparative studier av de 3 største bysamfunn i landet.

4. Øst-Norge.

Valget av Sør-Gudbrandsdalen (Lillehammer m/Fåberg hadde ca. 8.000 innbyggere i 1865) er gjort med sikte på prosjektene som studerer hamskiftet i Fåberg, arbeiderbevegelsen på Lillehammer samt husmannsvesenet i Gudbrandsdalen. Østlandet var det mest folkerike område. Derfor kan mye tale for å tilgodese dette med enda en region. Som et eksempel kan nevnes at Edv. Bulls studier av industrialiseringa av Østfoldbyene etterlater mange interessante forskningsoppgaver som bare kan besvares med studier på individnivå.

Planene om å registrere ca. 100.000 enheter fra 1910-tellinga for deler av Kristiania og Akershus, må sies å ligge fast som en naturlig forlengelse av Ullensaker- og Kristianiaprosjektene.

Kildevalget er fremdeles gjenkallelig. Forskernes begrunnede ønsker må veie tungt når styringsgruppa treffer det endelige valg. Et historisk dataarkiv kan bare legitimere seg gjennom de resultater forskerne publiserer på grunnlag av registrert materiale.

REPORT FROM A SYMPOSIUM ON GRAMMATICAL TAGGING OF ENGLISH TEXT CORPORA

Stig Johansson

An international symposium on "Grammatical Tagging of English Text Corpora in Machine-Readable Form" was held at Bergen on March 29-30, 1979. The symposium, which was financially sponsored by the Norwegian Research Council for Science and Humanities and the Universities of Oslo and Bergen, was arranged as part of the work within ICAME. It was attended by 37 participants from 10 countries.

The background to the symposium was the realization that corpora of (unanalyzed) natural-language texts are insufficient for many types of linguistic investigation, coupled with the discovery that linguists in different parts of the world had embarked on projects of grammatical tagging, seemingly unaware of each other's work and in some cases applying different systems of analysis to exactly the same material. During the Bergen symposium representatives from different projects had an opportunity to describe their work and profit from each other's experiences.

It is impossible to adequately summarize the papers and discussions. Wherever feasible, references will be made to publications giving detailed information on the particular projects.

Randolph Quirk (University College London) gave an introductory lecture on "The Place of Corpus Study in English Language Research". He emphasized the special features of the new corpora compared with the sources of material used by traditional grammarians such as Jespersen and Poutsma. In particular, the new corpora have been systematically compiled to represent a broad range of text types. They are further intended to be subjected to "total accountability" rather than to analysis of selected features. Quirk, who in his talk also touched on the relationship between corpus and elicitation, has recently dealt with these matters in a joint article with Jan Svartvik, "A Corpus of Modern English", in H. Bergenholtz and B. Schaeder, eds., *Empirische Textwissenschaft: Aufbau und Auswertung von*

Os være utgangspunkt for lignende undersøkelser av et større sammenhengende område. Man vil da oppnå resultater som er gyldige for migranter som forlot sin hjemkommune, men ble i regionen. Hvis valget faller på Sunnhordland er det også mulig å fange opp dem som dro til Bergen i folketellinga av 1875, eventuelt i emigrasjonsprotokollene. I tillegg til at området "lekker" i sør, er problemet at Sunnhordland er svært folkerikt, (33.695 innbyggere i 1865). Derfor er det naturlig at dette området "rammes" først ettersom ad hoc oppgaver må løses.

Siden 1910-tellingas fødselsdatoer vil gi størst gevinst i store kommuner, er det naturlig å inkludere Norges nest største by. Sunnhordland og Bergen hadde tilsammen 110.201 innbyggere på dette tidspunkt. Man ser hvilke muligheter som åpner seg for komparative studier av de 3 største bysamfunn i landet.

4. Øst-Norge.

Valget av Sør-Gudbrandsdalen (Lillehammer m/Fåberg hadde ca. 8.000 innbyggere i 1865) er gjort med sikte på prosjektene som studerer hamskiftet i Fåberg, arbeiderbevegelsen på Lillehammer samt husmannsvesenet i Gudbrandsdalen. Østlandet var det mest folkerike område. Derfor kan mye tale for å tilgodese dette med enda en region. Som et eksempel kan nevnes at Edv. Bulls studier av industrialiseringa av Østfoldbyene etterlater mange interessante forskningsoppgaver som bare kan besvares med studier på individnivå.

Planene om å registrere ca. 100.000 enheter fra 1910-tellinga for deler av Kristiania og Akershus, må sies å ligge fast som en naturlig forlengelse av Ullensaker- og Kristianiaprosjektene.

Kildevalget er fremdeles gjenkallelig. Forskernes begrunnede ønsker må veie tungt når styringsgruppa treffer det endelige valg. Et historisk dataarkiv kan bare legitimere seg gjennom de resultater forskerne publiserer på grunnlag av registrert materiale.

REPORT FROM A SYMPOSIUM ON GRAMMATICAL TAGGING OF ENGLISH TEXT CORPORA

Stig Johansson

An international symposium on "Grammatical Tagging of English Text Corpora in Machine-Readable Form" was held at Bergen on March 29-30, 1979. The symposium, which was financially sponsored by the Norwegian Research Council for Science and Humanities and the Universities of Oslo and Bergen, was arranged as part of the work within ICAME. It was attended by 37 participants from 10 countries.

The background to the symposium was the realization that corpora of (unanalyzed) natural-language texts are insufficient for many types of linguistic investigation, coupled with the discovery that linguists in different parts of the world had embarked on projects of grammatical tagging, seemingly unaware of each other's work and in some cases applying different systems of analysis to exactly the same material. During the Bergen symposium representatives from different projects had an opportunity to describe their work and profit from each other's experiences.

It is impossible to adequately summarize the papers and discussions. Wherever feasible, references will be made to publications giving detailed information on the particular projects.

Randolph Quirk (University College London) gave an introductory lecture on "The Place of Corpus Study in English Language Research". He emphasized the special features of the new corpora compared with the sources of material used by traditional grammarians such as Jespersen and Poutsma. In particular, the new corpora have been systematically compiled to represent a broad range of text types. They are further intended to be subjected to "total accountability" rather than to analysis of selected features. Quirk, who in his talk also touched on the relationship between corpus and elicitation, has recently dealt with these matters in a joint article with Jan Svartvik, "A Corpus of Modern English", in H. Bergenholtz and B. Schaeder, eds., *Empirische Textwissenschaft: Aufbau und Auswertung von*

Text-Corpora. Königstein/Ts.: Scriptor Verlag, 1979.
(This is the final title of the book which was announced in *ICAME NEWS* 1.)

If the preceding talk dealt with general linguistic matters, the particular uses of the computer in linguistics were taken up in brief contributions by Alvar Ellegård (University of Gothenburg) and Geoffrey Leech (University of Lancaster). Ellegård emphasized the importance of the computer in handling large bodies of data and relieving the linguist of much routine work, whereas Leech focused his remarks on the special advantages and possibilities offered by computer corpora and the need for cooperation in computer corpus work.

W. Nelson Francis (Brown University) presented the system which has been used in the recently completed tagged version of the Brown Corpus (cf. p.1 above). The system, which is essentially that outlined in B.G. Greene and G.M. Rubin, *Automatic Grammatical Tagging of English* (Department of Linguistics, Brown University, 1971), involves the assignment of one of 80 tags to each word in the material, through a combination of automatic procedures (dictionary look-up, suffix list look-up, context frame rules) and manual pre- and post-editing. The Brown Corpus tagging project is described in a paper by W. Nelson Francis, "A Tagged Corpus: Problems and Prospects" (forthcoming) and in the manual mentioned on p. 2 above.

Henry Kučera (Brown University) reported on results from studies of the tagged Brown Corpus in his talk on "The Frequency of Grammatical Classes in the Brown Corpus". Statistics were given for the frequency of individual tags (singular common noun, plural common noun, singular proper noun, etc.) as well as for major classes such as nouns, pronouns, verbs, etc. The latter were also ranked and compared with the frequencies in a Czech corpus. Word-class distribution across genres was further studied in a way which revealed the varying degree of "contextuality" of the major tag classes.

Alvar Ellegård (University of Gothenburg) described his analysis of portions of the Brown Corpus. This very detailed system, which, in contrast to that used at Brown University, does not involve any automatic procedures, has already been presented in this newsletter (*ICAME NEWS* 2, pp. 3-7).

Jan Aarts (University of Nijmegen) gave a report on "Grammatical Tagging in the Dutch Computer Corpus Pilot Project". The system is being implemented on a corpus of modern English texts assembled in Holland. It involves the manual assignment of a four-digit code to each word in the text and includes word-class labels comparable with those used in the Brown University project (the first two digits) as well as boundary markers (the last two digits). Categorial and functional constituents are derived from the four-digit code by a series of algorithms. The system has been adapted from J. van Bakel, *Automatische Syntactische Analyse van Nederlandse Teksten* (Computer Centre, Katholieke Universiteit, Nijmegen, 1970). Information on the Dutch project has been given in a paper by Jan Aarts on "Syntactic Coding of a Computer Corpus", presented at the 5th International Congress of Applied Linguistics, Montreal, August 20-26, 1978. The system of analysis is described in detail in a *Manual for Coders*, which is available on request from: Jan Aarts, Department of English, University of Nijmegen, Holland.

Rudolf Filipović (University of Zagreb), who was unfortunately prevented at the last moment from attending the symposium, submitted a paper on "The Grammatical Tagging of the 'Zagreb Version' of the Brown Corpus". In the Zagreb project about half of the Brown Corpus has been selected and translated into Serbo-Croatian, with the object of providing a source of data for contrastive analysis. The text is tagged manually according to a system in part reminiscent of Ellegård's and in part similar to that of the Dutch project. Words are assigned a four-digit code corresponding to part of speech (the first two digits), function of words or phrases in clauses (the third digit), and function of clauses in the sentence (the fourth digit), though the last two digits are only used with the first word of a syntactic constituent

Information on the Zagreb project has been given in publications by Filipović from the Serbo-Croatian-English Contrastive Project.

Jan Svartvik (University of Lund) gave an outline of the plans for the grammatical analysis of the London-Lund Corpus of spoken British English. The plans include semi-automatic word-class tagging similar to the Brown model as well as higher-level syntactic analysis. In his talk Svartvik touched on the particular problems of tagging spoken material, e.g. those posed by having the tone unit rather than the sentence as the basic element. The projected system is described in Jan Svartvik, "Tagging Spoken English" (forthcoming). See further the information on the London-Lund Corpus, pp. 6-8 above.

Mamata Nakra (Maisonneuve College) gave a talk on "Grammatical Tagging of Journalistic Prose" based on her work on newspaper material from the Brown Corpus, presented in her thesis on the topic. Nakra's system has not yet been implemented computationally.

Viljo Kohonen (University of Turku) described the CHITAB program, which he has developed in cooperation with Jussi Salmela. The program operates on a coded version of the text (manually assigned) without direct access to coding and text at the same time. It has been used in Kohonen's recently completed thesis, *On the Development of English Word Order in Religious Prose around 1000 and 1200 A.D.* Publications of the Research Institute of the Åbo Akademi Foundation, No. 38. Åbo 1978, and is described on pp. 223-227 of his work.

Claus Faerch (University of Copenhagen) reported on the grammatical analysis of a corpus of learners' language collected in Denmark and consisting of English as spoken and written by Danes. The tagging is restricted to the assignment of word-class labels by semi-automatic techniques along the Brown model. Faerch touched on the particular problems caused by the learner-language material. Is it possible to characterize learner-language as a

system? If not, can you write rules for the assignment of tags? The Copenhagen project is described in Faerch's report on "Computational Analysis of the PIF Corpus of Learner Language", *PIF Working Papers*, No. 1, Department of English, University of Copenhagen, 1979.

Dirk Geens (University of Leuven) was the only one of the participants who gave a report on automatic syntactic analysis, based on his recently completed doctoral thesis on the topic. Geens' analysis, which has been implemented on the Leuven Drama Corpus (described in *ICAME NEWS* 2, pp. 7-9), included a "syntactic recognition procedure that was mainly used to produce a rough characterization of the apparent syntactic structure of each sentence" and a "full syntactic analysis procedure", both of which are too complex to be summarized here. Geens also dealt with some problems of semantic analysis.

In a lucid and relevant introduction to the final discussion period, Sture Allén (University of Gothenburg) presented a taxonomy of tagging systems based on the type of material (running text, sorted text, linked network), purpose of analysis (lexical, grammatical, communicative), and tagging technique (off-line encoding, on-line encoding, interactive procedure, automatic procedure). Through Allén's paper the work at the symposium was placed within a more general framework of computational-linguistic analysis.

The contributed papers presented a variety of tagging projects, from the point of view of the material (cf. Francis, Svartvik, Faerch) as well as aim (cf. Francis, Filipović, Geens) and technique (cf. Francis, Ellegård, Geens). An illustration of different ways of tackling the same material was given by four treatments of an extract from the Brown Corpus: the Brown model, Ellegård's and Filipović's systems, and the Dutch system (Jan Aarts was kind enough to provide comparative material, though

the text was not part of the corpus analyzed in the Dutch project). Unfortunately, space does not permit the reproduction of this material here.

The variation in the ways of handling the same material raises the question why a particular system is chosen. It was clear from the discussion that different projects had different aims, as already hinted at in the preceding paragraph, which partially explains the varying approaches (Naturally, these are also due to such more trivial matters as available funds, personnel, equipment, etc.) The Brown analysts were concerned with developing a source of data as neutral as possible with respect to future applications and were therefore content with a fairly uncontroversial, traditional, "surface" analysis, whereas Ellegård had the more specific aim of revealing the syntactic features of four categories of English texts and therefore considered it necessary to perform a fuller analysis. The varying aims in other cases (Faerch, Filipović, Geens) should be immediately recognizable to the reader.

The linguistic differences between the systems should, however, not be exaggerated. Most of them use traditional categories (parts of speech, parts of the sentence, clause types), though the delicacy of analysis and the labels may differ. The question of standardization of categories and labels was raised in the discussion. It was agreed by almost everybody that standardization is impossible, or even undesirable, in view of the varying aims, ambitions, and resources of projects. Instead, it was pointed out that it is desirable to have systems which are convertible to some extent and always to provide explicit descriptions of the systems used.

In conclusion (to take up just one further matter from the discussion), it was agreed that the exchange of information between researchers must be improved. To partially solve this problem, it was proposed that a follow-up symposium should be held in two years, if possible.

(Referatet er hentet fra ICAME NEWS no. 3, Oct. 79)

COMPUTER-SENTER FOR GRESK FILOLOGI (THESAURUS LINGUAE GRAECAE). UNIVERSITY OF CALIFORNIA, IRVINE.

Knut Kleve

For en tid tilbake oppholdt jeg meg tre uker ved TLG for å studere bruk av EDB-metoder i gresk filologi.

Senteret ble opprettet i 1972 ved en privat donasjon på 1 mill. dollar. Idag er TLG en avdeling av Klassisk institutt ved University of California, og dets 18 ansatte lønnes over universitetsbudsjettet. Leder er professor dr.Th.F.Brunner.

Formålet for senteret er å sette den overleverte greske litteratur fra oldtiden på maskinleselig form, hvilket vil si en tekstmasse på 90 mill. ord fra en periode som strekker seg fra Homer til år 700 e.Kr. Punchingen er satt ut på anbud og utføres av et firma i Korea. Tilretteleggingog siste korrektur foregår ved TLG. Man er nå på det nærmeste ferdig med litteraturen fra Homer til 200 e.Kr. (et materiale på 20 mill.ord). Man kan dermed få tekster på tape til kostpris av en hvilket som helst forfatter fra dette tidsrom. Avdelingen produserer dessuten korrektur-programmer og programmer for konkordanser, indekser og leksika. Konkordanser etc. utføres på bestilling. Korrekturprogrammene og lemmatiseringen er avanserte. Man kan slå ned på enhver uvanlig bokstavkombinasjon, få skilt språk i de forskjellige perioder og bøye alle substantiv, verb etc., selv de mest uregelmessige.

Avdelingen har også tatt de første steg for å få satt det store papyrusmaterialet som er kommet for dagen i Egypt de siste hundre år, på maskinleselig form. Når det gjelder papyrene fra Herculaneum, er et samarbeid innledet med Klassisk institutt i Oslo.

Under mitt besøk ble jeg satt inn i avdelingens generelle virksomhet. Av særlig interesse var det å få arbeide med store tekstmasser og se hvilke problemer som må løses for å lette søkingen i slike.

TLG er primært innrettet på å produsere, og produktene er til bruk for andre og anvendes ikke til forskning ved TLG selv. Dette er

både en styrke og en svakhet. Mulighetene for visse typer av programmer kan nemlig ikke oppdages uten aktiv forskning. Derfor kunne jeg fra norsk side bidra med et program TLG manglet, et som EDB-konsulent Ivar Fonnes har laget for meg og som muliggjør søking etter strenger med et ubegrenset antall variable. Programmet er av stor betydning i papyrologien der det gjelder å fylle ut lakuner (hull) i tekstene.

KONFERANSE OM DATATENESTER FOR OG DATASAMARBEID MELLOM DEI KUNST- OG KULTURHISTORISKE MUSEA, USTAOSET HØGFJELLSHOTELL, 23. OG 24. OKTOBER 1979.

NAVF's EDB-senter for humanistisk forskning arrangerte 23. og 24. oktober saman med Norske Kunst- og Kulturhistoriske Museer (NKKM) ein konferanse om datatenester for dei kunst- og kulturhistoriske musea. Bakgrunnen for konferansen var at ei rekkje musé no er i ferd med å orientera seg mot EDB i katalogiseringsrutinar og ved gjennomføring av forskningsprosjekt. NKKM har m.a. hatt ein EDB-komitè i arbeid, og denne komitéen har levert framlegg til eit felles katalogiseringskort som kan nyttast av dei kunst- og kulturhistoriske musea. NAVF's EDB-senter har dei siste åra gjennomført ei rad prøveprosjekt med kunst- og kulturhistorisk datatilfang. Tida var derfor inne til å få gjennomført ein prinsippdiskusjon med sikte på å koma fram til eit konkret program for det vidare arbeidet med å etablera fornuftige EDB-løysingar for musea.

Deltakarane på kurset kom frå musé og tilgrensande institusjonar som er i gang med prøveprosjekt, eller som har reelle planar om slike prosjekt, og frå andre humanistiske fagområde med aktive EDB-miljø. Dessutan deltok ei rad EDB-medarbeidarar frå alle dei 4 universiteta.

Som resultat av drøftingane vart det vedteke å arbeida for at det blir oppretta ei datateneste for musea knytt til NAVF's EDB-senter i Bergen. Dette vart sett på som eit naudsynt tiltak på kort sikt inntil det einskilde museum, eller fleire musé saman, kunne utvikla meir institusjonsbaserte EDB-løysingar. Med utgangspunkt i ei prinsippskisse for ei datateneste som vart lagt fram på konferansen vart det drøfta korleis ein kunne få finansiert ei slik teneste og kva verknader eit slikt service-organ kunne få for det faglege arbeid i musea. Det vart mellom anna peika på at bruk av EDB gjer det meir aktuelt enn før å etablera tiltak som kan auka kvaliteten i det faglege nomenklaturarbeidet. Fleire ønskte seg eit fagleg utviklingsprosjekt knytt til fagterminologi der ein stipendiat kunne fordjupa seg i dette emnet.

Det blei m.a. foreslått at NKKM's EDB-komité skulle arbeida vidare med dei tema som var tekne opp på konferansen. Særleg burde komitéen konsentrera seg om å laga detaljerte planar for ei datateneste og arbeida vidare med nomenklatur-saka.

Det er nå under arbeid ein større konferanserapport som kan tingast frå NAVF's EDB-senter.

For dei som er spesielt interesserte i emnet vil vi gje ei førebels orientering utarbeidd av førstekonsulent *Einar Ådland*, Opplysningsavdelingen, Universitetet i Bergen. Ådland, som var til stades på konferansen, laga opphavleg orienteringa for Nytt fra Universitetet i Bergen nr. 2,1979.

"Ein databank som skal vera til hjelp for norske museum er no i emning. Eit slikt datasenter vil gjera det mogleg for både store, sentralt plasserte museum og små bygdemuseum å henta fram opplysningar om gamle ting, bygningar og skip på ein mykje snøggare og meir effektiv måte enn det har vore råd til no.

På ein konferanse som Norske Kunst- og Kulturhistoriske Museer skipa til saman med NAVF's EDB-senter for humanistisk forskning på Ustaoset 23. og 24.okt. vart eit landsomfemnande data-samarbeid mellom musea her i landet grundig drøfta og dei om lag 40 deltakarane - museumsfolk og dataspesialistar - drog opp hovudlinene for eit slikt samarbeid.

Fylkeskonservatoren i Hordaland, *Nils Georg Brekke*, sa i si innleiing at eit datasamarbeid mellom musea vil kunna få ein svært positiv desentraliserande verknad på museumsarbeidet og kulturlivet reint ålment i vårt land. - Det vil ikkje lenger vera noko skilje mellom eit stort og eit lite museum når det gjeld tilgjenge til datatenester og kjeldetilfang og utveksling av opplysningar. På det viset vil både store og små museum kunne gjera museums-tinga og opplysningane om desse lettare tilgjengelege for eit stort publikum - og det vil og verta råd å driva *forskning* på musea i større mon nå enn før.

STOR TILGANG PÅ GAMLE TING.

Direktør *Halvard Bjørkvik* på Norsk Folkemuseum - formann i Norske Kunst- og Kulturhistoriske Museer (NKKM) - sa at det kvart år kjem veldige mengder av gamle ting til musea. Såleis kom det i tiårsbolken 1968/69-1978/79 heile 258.000 ting inn til dei norske musea som NKKM har opplysningar om. Til samanlikning kan nemnast at Norsk Folkemuseum i alt har 180-200.000 ting i sine samlingar.

Bjørkvik meinte at dette store tilfanget burde kunna nyttast betre ut, og sa at datateknikken kunne gje gode løysingar i så måte. Han vona det skulle la seg gjera å skaffa pengar til eit datasamarbeid for musea - ikkje minst fordi EDB-teknikken ville kunne fremja forskinga på musea - ei oppgåve som det i seinare år er vorte lagt meir vekt på. Bjørkvik trudde elles at overgangen til ny datateknikk burde kunna gjennomførast utan alt for store utgifter.

Leiar for Noregs allmenvitenskaplege forskningsråds EDB-senter for humanistisk forskning ved Universitetet i Bergen, *Jostein H. Hauge* gjorde greie for dei samarbeidsprosjekt som senteret dei siste åra har hatt med dei ulike arkeologiske og kunst- og kulturhistoriske musea. Han meinte at tida no var inne til å planleggja ei nasjonal datateneste for musea og la fram ei prinsippskisse for ei slik teneste. Hauge la vekt på at eit slikt organ måtte ha klårt definerte oppgåver og sa at NAVF's EDB-senter vil by seg fram som eit datafagleg støttemiljø for ei slik teneste. For at datateknikken skal kunna fremja arbeidet, blir det viktig at dei einskilde musea set i verk opplæringstiltak for alle dei som skal nytta det nye data-tilbodet.

OPTIMISME.

Konferansen vart opna av førstekonservator *Kjell Falck* ved Historisk museum, Universitetet i Bergen, som sa at det no truleg er betre grunnlag for optimisme med tanke på datasamarbeid mellom musea enn nokon gong før. Men skal ein lukkast i arbeidet trengst det grundig førearbeid, nasjonal samordning og nært samarbeid mellom kollegaer, la han til.

På konferansen gjorde registrator *Randi Johannessen* greie for "Prøveprosjekt om bruk av EDB ved Norsk Folkemuseum". Både ho og konservator *Jan Henrik Munksgaard* ved Historisk museum i Bergen - som tala om regionalt datasamarbeid mellom folkemusea i Hordaland og dei kulturhistoriske registreringane på Vestlandet - tok til orde for meir presise nemningar på dei gamle tinga.

Førsteamanuensis *Atle Thowsen* fortalde om planar for EDB-drift ved Bergens Sjøfartsmuseum. Han såg stort sett positivt på bruk av datateknikk i musea, men var og redd for at EDB kunne skapa avstand mellom forskaren og dei tinga han forska i.

Konsulent *Sigbjørn Århus*, NAVF's EDB-senter, gjorde greie for "Metoder for datamaskinell informasjonsøking" og heldt fram at det tilsynelatande ikkje er grenser for kor mykje informasjon vi kan leggja inn i datamaskinen.

Konsulent *Ragnar Dag Blekeli* frå Rasjonaliseringsdirektoratet kom i sitt foredrag "Mikroprosessoren som "museumsgjenstand"" m.a. inn på korleis datamaskinane er blitt stadig billegare og meir effektive og korleis det no er teknisk mogleg å desentralisera bruken av desse maskinane - nytta dei lokalt i mykje større mon enn før.

Fylkeskonservator *Nils Georg Brekke* fortalde om arbeidet i NKKM's EDB-komite, der han er formann.

Oppdragsleiar *Kjell Fredriksen* ved EDB-avdelinga på Universitetet i Bergen sa at det er ei ulempe at det til no er arbeidd ut for få program for dei små datamaskinane, men dette vil kunne retta på seg etterkvart. Han meinte at samspelet mellom datamaskinen og dei som søkjer kunnskap gjennom denne i "interaktivt program" - det vil seia ei stendig veksling mellom spørsmål/kontrollspørsmål og svar - gjer det mogleg å finna fram til veldige datamengder på brøkdelen av den tida det ville ha teke på manuelt vis.

Driftsassistent *Per Vestbøstad*, NAVF's EDB-senter, fortalde om utstyr til og organisasjon av EDB-drift, EDB-sjef ved Universitetet i Bergen, *Carl Ellingsen* spurde: "Er databehandling dyrt?", museumsdirektør *Halvard Bjørkvik* gjekk inn på musea sin økonomi og bruken av EDB."

ICCH/4

THE FOURTH INTERNATIONAL CONFERENCE ON COMPUTERS AND THE HUMANITIES 19. - 22. AUGUST.

DATA BASES IN THE HUMANITIES AND SOCIAL SCIENCES 23. - 24. AUGUST 1979, HANOVER, NEW HAMPSHIRE.

De to konferansene ble holdt på Dartmouth College og samlet ca. 250 deltakere hver.

Arrangementsteknisk var konferansene dårlige. Det ble gitt lite informasjon på forhånd og det forelå heller ikke noen form for sammendrag av foredragene da konferansene startet. Ettersom det bare på den første (ICCH/4) ble holdt ca. 70 foredrag fordelt på 20 sesjoner og 11 rundebordskonferanser sier det seg selv at det var vanskelig å beholde oversikten og å treffe begrunnete valg blant de mange muligheter som forelå.

Kvaliteten på foredragene var svært variert og grupperingen av dem virket til tider forvirrende. Konferanser av denne type spenner over et svært bredt spekter; det eneste som er felles er at forskerne har gjort bruk av datamaskin. Ofte var forøvrig også EDB-siden helt perifer i foredraget.

Det sosiale program var meget spinkelt, et unntak var en konsert med "New music with computers", som ga tilhørerene førstehånds kjennskap til hva som kan tilbys av elektronisk generert og styrt musikk i dag. Mange av de emnene som var tatt opp, var velkjente fra EDB-arbeid i vårt land. Flere foredragsholdere tok opp spørsmålet om bedre generelle programmer og programsystemer for tekstbehandling og ordboksproduksjon og demonstrerte nye og mer fleksible program for eksempelvis konkordansproduksjon.

En rekke presentasjoner og et par rundebordskonferanser var knyttet til arbeid med historiske primærkilder og arkivmateriale. Det er i gang flere interessante prosjekter knyttet til billed- og malerisamlinger og arkivalia mer generelt. Problemene er de samme som hos oss: enorme datamengder, tidkrevende datatilretteleggingsarbeid og mangel på standardisert faglig nomenklatur, noe som er nødvendig å etablere for å kunne dra nytte av automatiserte arkivsøkesystemer.

Det prøvearbeid som er utført i vårt land, synes å kunne hevde seg godt sammenlignet med tilsvarende arbeid i andre land. Det ble på konferansen vedtatt å nedsette en (amerikansk) interessegruppe for å fremme arbeid på dette feltet. Senteret kan formidle adresser til kontaktpersoner.

På den andre konferansen, databasekonferansen, var flere foredrag konsentrert om datatyper fra folketelling, kirkebøker, protokoller og offentlig statistikk. Andre presentasjoner tok opp emner som leksikografi, produksjon av ordbøker, maskinoversettelse, programpakker for statistikk, bibliografisk arbeid og dokumentasjonstiltak for gjenstandsfagene.

Begrepet database ble her brukt i populær forstand: en samling data i en datamaskin. Bruken av databaseteknikk (i egentlig forstand) ble bare berørt i et par foredrag.

Av de mange interessante presentasjonene kan nevnes orientering om en ny statistisk programpakke P-STAT 78 som er mer fleksibel enn SPSS.

Imponerende var de data som ble lagt fram om de nye datalagringsmedia som er i utvikling. F.eks. kan en på en laserplate av størrelse som en LP-plate lagre data tilsvarende 3 mill. skrevne A4 sider. Platekostnadene vil bli ca. $10.

Med bakgrunn i disse to konferansene kan en hevde at det EDB-arbeid som drives i Norge på mange områder er fullt på høyde med

tilsvarende arbeid i utlandet. Vi bør derfor med større frimodighet enn hittil stå fram med norske foredrag på slike konferanser. Denne gang var det bare ett norsk foredrag: Prof. Knut Kleve ved Universitetet i Oslo orienterte om bruk av EDB i forbindelse med studium av papyrusmateriale.

Amerikanske forskere synes å ligge langt fremme på områder som kunstig intelligens, semantikk og bruk av ny teknologi generelt.

Her i Norge er vi imidlertid spesielt ved universitetene i en gunstig stilling når det gjelder datamaskinressurser. Et inntrykk blant mange deltakere fra USA var at de hadde bygget opp dataarkiver som de hadde dårlig råd til å bruke.

I private samtaler kom det også til uttrykk en allmenn bekymring for den stilling den humanistiske forskning for tiden har i USA. Det synes å være langt vanskeligere enn før å få støtte til humanistiske forskningsprosjekter og å holde oppe kvaliteten på den forskning og undervisning som drives ved universitetene. Enkelte så på situasjonen med fortvilelse og desperasjon, og det ble også antydet at man forberedte lobby-virksomhet blant politikere for å gjøre de bevilgende myndigheter oppmerksom på den fare som truer de humanistiske tradisjoner i USA i dag.

Det foreligger en mer utførlig konferanserapport utarbeidet av Jostein H. Hauge og Knut Hofland som kan fås ved henvendelse til NAVF's EDB-senter.

17TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, SAN DIEGO, 11. - 12. AUGUST 1979

Konferansen ble holdt i naturskjønne omgivelser på University of California, San Diego og samlet ca. 170 deltakere, nesten utelukkende fra amerikanske universiteter og forskningsinstitusjoner.

Feltet datamaskinell lingvistikk har vært i en rivende utvikling siden interesseorganisasjonen ble grunnlagt. Fra å være konsentrert om mekanisert oversettelse, er interesseområdet blitt utvidet til å gi rom for arbeid med talegjenkjenning, automatisk morfologisk og syntaktisk analyse, semantiske og kognitive modeller, modeller for lagring av kunnskap, spørsmål/svar-systemer, datamaskinassistert læring og mange flere.

Det 17. møtet i rekken var meget godt organisert, og ikke minst var det kjærkomment at det før konferansen startet, forelå et hefte med sammendrag av alle foredragene.

Konferansen var delt i 4 sesjoner:

Language structure and parsing
Knowledge organization and application
Dialogue
Applications

I alt ble det presentert 24 foredrag under disse temaer.

Language structure and parsing.

I sin introduksjon til foredragene til denne sesjonen gjennomgikk *Martin Kay* utviklingen av metoder for automatisk språkanalyse, særlig syntaktisk analyse, de siste 25 årene. Den første perioden

av datalingvistikken var preget av arbeid med maskinoversettelse. Typisk for de første systemene her var at de i hovedsak besto av et leksikon bygget opp i maskinkode. Det tok imidlertid ikke lang tid før kontakten med virkeligheten, dvs. analyse av naturlig språk, viste at dette ikke var tilfredsstillende metodikk.

Den språklige virkelighet fremsto som langt mer kompleks enn antatt og bruk av maskinkoder for å representere andre sider ved språket enn ordbestanden (syntaks, semantikk) gjorde systemene til uhåndterlige mastodonter som til slutt brøt sammen under sin egen vekt. Dette førte til en oppdeling mellom data og program og satt fart i arbeidet med å utvikle algoritmer for syntaktisk analyse. Viktig tildriv til dette arbeidet kom også fra generell syntaktisk teori, ikke minst fra arbeidet med transformasjonelle generative modeller.

Dette gav grunnlag for systemer som gjennom bruk av grammatiske regler i strengt formalisert form kunne gi enklere, mer gjennomskuelige og kraftige redskaper enn før ved språklig analyse.

"There has been a shift of emphasis away from highly structured systems of complex rules as the principle repository of information about the syntax of a language towards a view in which the responsibility is distributed among the lexicon, semantic parts of the linguistic description and a cognitive or strategic component. Concomitantly, interest has shifted from algorithms for syntactic analysis and generation, in which the control structure and the exact sequence of events are paramount, to systems in which a heavier burden is carried by the data structure and in which the order of events is a matter of strategy."

Denne løsere organiseringen av programmer for syntaktisk analyse kommer delvis fra et generelt ønske om å bryte ned grensene mellom de vanligvis adskilte morfologiske, syntaktiske og semantiske delsystemer. Ikke minst har arbeidet med automatisk analyse av talt språk og arbeidet innen kunstig intelligens bidradd til denne nyformulering. På mange måter har den nye tendensen gitt *leksikondelen* en mer prominent stilling enn tidligere og redusert omfanget

av transformasjonsregler i språkanalysesystemene, idet en del av regelverket knyttes direkte til leksikondelen.

I foredragene i denne sesjonen ble bl.a. ulike analysesystemer sammenlignet (*Marc Eisenstadt*) og det ble referert fra arbeid med analysesystemer som selv lærer betydningen av ord ut fra bakgrunnskunnskap om teksten og bruk av kontekst (*Jaime G. Carbonell*).

Interessant var også foredraget *"Ungrammaticality and extra-grammaticality in natural language understanding systems" (Stan C. Kwasny)* hvor det ble drøftet metoder for å bygge et automatisk system for tolkning og forståelse av naturlig språk, slik at også inndata i grammatisk ukorrekt eller ufullstendig form ville bli riktig tolket.

Knowledge organization and application.

De fleste foredragene i denne sesjonen omhandlet kunnskapslagring i datamaskin og redegjorde for ulike typer av kunnskap som må tas hånd om for at datamaskinen skal kunne "forstå" naturlig språk.

Også utenfor USA har datalingvistene med interesse fulgt med i utviklingen av de automatiske såkalte "story understanding systems" kjent under navn som SAM, PAM og FRUMP. Systemene sikter mot å forstå tekster innenfor veldefinerte språksituasjoner (restaurantbesøk, bilulykker) ved bruk av strukturert bakgrunnskunnskap og viten om hvilke sosiale samhandlingsmønstre som er dominerende (kalt *"scripts", "plans", "goals")*. Mens SAM og PAM *"model the way people might read the story if they were expecting a detailed test on it"*, er FRUMP et robust system som skumleser en avistekst som en vanlig leser gjør det. Det nye systemet IPP (Integrated Partial Parser), som *Michael Lebowitz* gjennomgikk, har som mål å etterlikne den måten folk vanligvis leser en avisreportasje på, d.v.s. gjennom stoffutvalg basert på interesser etc.. IPP hadde spesialisert seg på reportasjer om vold og terroristhandlinger. Ifølge foreleseren innebærer denne målsettingen at en ved utviklingen av systemet bl.a. også tar hensyn til den kunnskap som er innvunnet innen kognitiv psykologi.

Discourse.

Under dette oppslaget ble det redegjort for metodiske forsøk innen et nytt avansert felt. Målet for arbeidet på feltet er ved hjelp av språkbruken å finne fram til kunnskap om de intensjoner språkbrukeren har. Det blir da viktig å kunne isolere de trekk ved språkbruken som signaliserer og gir hint om intensjoner og mening. Dette har en tid vært et studieemne innen allmenn lingvistikk *(discourse analysis, text grammar)*, men er nå også interessant innen datalingvistikk. Ved interaktiv databehandling er forholdene lagt til rette for en dialog mellom menneske og maskin. I relasjon til datamaskinen blir derfor ikke språket bare en kode som behandles paradigmatisk ved analyse etc., men språkbruk fremstår også som handling - en måte å opptre på *(behaviour)*.

Kathleen R. McKeown redegjorde i foredraget *"Paraphrasing Using Given and New Information to a Question-Answer System"* for systemet CO-OP som er et system for å parafrasere spørsmål som er stilt til en database i naturlig språk. Målet med parafrasen er å sikre at systemet oppfatter spørreren riktig.

Spørsmålet om hvordan vi reflekterer ved problem-løsning var temaet for foredraget *"Structure and Process of Talking About Doing"* av *James A. Levin og Edwin D. Hutchins*. Ved hjelp av såkalte *"problem solving protocols"* mente de å kunne studere de kognitive prosesser som aktiveres ved problem-løsning.

Andre foredrag tok opp mer spesielle emner bygd på språkhandlingsteori *(Speech Act Theory)* og spill-teori *(Dialogue Game Theory)*.

Applications.

Karakteristisk for denne konferansen var at de aller fleste foredragene knyttet seg til utprøving av metoder for språkanalyse, teksttolking og kunnskapslagring. Bare et fåtall (6) redegjorde for konkrete anvendelser av metoder. Dette er kanskje typisk for den situasjon datalingvistikken, spesielt i USA, er i for tiden, men som *David G. Hays* uttrykte det: *"Science needs applications, since contact with reality tends to remind us scientists that there are*

more things out there than are dreamed of in our theories". Et par av foredragene skal nevnes her:

I foredraget EUFID: *"A Friendly and Flexible Front-end for Data Management Systems"* ble vi presentert for et spørresystem som gav brukeren anledning til å spørre i en database i naturlig språk og få svar selvom spørsmålet ikke var grammatisk velformulert og inneholdt ortografiske feil.

Et interessant bruksområde for spørresystemer i naturlig språk er i medisinsk behandling hvor det frem til i dag er utviklet flere systemer som gjør bruk av naturlig språk.

I foredraget *"Natural Language Input To A Computer-based Glaucoma Consultation System"* redegjorde *Victor B. Cielsielski* for et system som brukes ved behandling av opplysninger om pasienter som har vært undersøkt med henblikk på grønn stær. Målet er å lage et system som er praktisk anvendelig og effektivt ved analyse av pasienters kliniske status, slik at legenes skriftlige diagnose i naturlig språk analyseres, og beskrivelsen reformuleres til de krav som settes av det EDB-baserte konsultasjons-systemet som skal bruke pasientopplysningene.

Sammenfattende kan en si at konferansen gav et nyansert bilde av datalingvistikkens stilling i dag i USA. Fremtredende trekk her er eksperimenter med modeller for lagring av semantiske opplysninger og kunnskap om verden mer generelt, og den sterke tilknytning til kognitiv teori og metode.

Virksomheten hos oss innen datalingvistikk synes derimot å være mer praktisk orientert og innrettet mot løsning av mer konvensjonelle lingvistiske oppgaver ved hjelp av datamaskin: lemmatisering, konkordansarbeid, morfologisk og syntaktisk analyse, produksjon av ordbøker etc..

Fraværet av mer grunnleggende metodisk arbeid vil imidlertid kunne medføre at vi i Norge ikke utvider interessefeltet nok til å kunne ta i bruk den kunnskap som i dag er tilgjengelig på mer avanserte

områder for datalingvistisk forskning. Som gjennomgåelsen ovenfor skulle kunne antyde, er det en omfattende meny å forsyne seg av om en ønsker en appetittvekker.

KONFERANSE OM LITTERÆR OG LINGVISTISK DATABEHANDLING
TEL AVIV, 22. - 27. APRIL 1979

Roald Skarsten

ALLC (Association for Literary and Linguistic Computing) avholdt 22. - 27. april d.å. organisasjonens 6. internasjonale konferanse i litterær og lingvistisk databehandling. Flere universitet i Israel sto sammen om arrangementet, som også hadde økonomisk støtte fra IBM. Det organisatoriske sentrum var The Katz Institute for Hebrew Literature, Tel Aviv University. Ca. 70 deltakere, vesentlig fra Europa og Israel, deltok. På bakgrunn av den politisk vanskelige situasjon og avstanden fra Nord-Europa må dette sies å være bra. Tidligere har denne konferansen alltid vært avholdt i England.

Konferansens tema var stort sett de samme som har vært med tidligere. Her kan nevnes konkordansproduksjon, tekstedítering, leksikografi, språklige databanker, informasjonssøking og språkstatistikk. Det virket som om språkstatistikken denne gang hadde fått ekstra oppmerksomhet.

Som man kanskje kunne vente,ble det en sterk dominans av forelesninger som belyste sider ved det hebraiske språk. Kjennskapet til dette språket var en viktig forutsetning for min deltakelse på denne konferansen, som ga et imponerende uttrykk for datamaskinens anvendelsesmuligheter i språklige og litterære studier. Som naturlig kunne være, kom flere israelske EDB-prosjekter til å stå sentralt på konferansen, og jeg vil nedenfor si litt om to av dem.

Et besøk på the Academy of the Hebrew Language og ved The Historical Dictionary Project i Jerusalem ga oss et innblikk i prosjekter med et svært omfattende datamateriale, hvor maskinen ble brukt til å holde orden på data, sortere, lage korrekturlister og produsere konkordanser og frekvenslister. Det hebraiske språk har en meget "sammensatt karakter". Til ordrøttene f.eks. knyttes i de forskjellige sammenhenger partikler, prefikser, suffikser og pronominale

elementer. I tillegg er ikke vokalene en del av alfabetet, men noe som enten blir "lest til" av leseren, eller antydet ved et prikk-system over og under konsonantene.
Enhver kan ut fra dette tenke seg hvilke problemer man står overfor i forbindelse med automatisk tekstbehandling. I Jerusalem foretok man derfor først en fullstendig manuell analyse av tekstordene med merking (tagging) og deretter ble materialet databehandlet. I neste omgang lot man maskinen komme med forslag til lemmatisering av en tekst på grunnlag av den informasjon den hadde fått i første omgang. Resultatene ble imidlertid dobbeltsjekket, d.v.s. en leksikograf kontrollerte og supplerte resultatene av den automatiske behandling. Gradvis bygget man opp en språkbank med frekvenser og lemma,og man hadde forskjellige databaser for forskjellige tidsperioder. Ved hjelp av materialet herfra presenterte datamaskinen forslag til lemmatisering av nye tekster, idet den startet med den formen som hadde høyest frekvens,og så videre i fallende frekvensrekkefølge. Resultatene var etterhvert imponerende, bare i 10% av tilfellene måtte leksikografen korrigere og/eller supplere, og da vesentlig i forbindelse med egennavn, stedsnavn og fremmedord.

Generelt fikk vi inntrykk av at man var svært stolt over språket, og at man fikk midler til arbeidet. En del av forklaringen på det er vel det ekstra store behovet for nydannelse av et vokabular som kan dekke den nye tids behov. Hebraisk var nemlig lenge, liksom latin i dag, levende p.g.a. den religiøse bruk.

Det andre store prosjektet, The Responsa Project, var et eksempel på at man ville forsøke mest mulig automatisering i arbeidet. Materialet bestod av rabbinske domsavgjørelser fra hele verden fra 1100-tallet og til vår tid. I dette prosjektet bruker man morfologisk analyse for å kunne finne frem til alle mulige former av det ord som brukes som "søkeord". Materialet består av ca. 28 millioner ord, altså en kjempedatabase for et fulltekst "text-retrieval system". Det er et omfattende automatisk grammatisk arbeid som maskinen må gjøre for å kunne finne frem alle, språklig sett, relevante tekststeder til et oppgitt ord. Maskinen har derfor adgang til en ordbok, en grammatikk og diverse "unntaks-banker". Man prøvde å utvikle algoritmer for automatisk løsning av tilfellene

hvor man har å gjøre med flertydige ord. Hittil var man ikke fornøyd med resultatene, men mente at fortsatt arbeid vil gi øket gevinst.

Dette prosjektet var lokalisert til Bar Ilan University i Tel Aviv, et "religiøst" universitet. Selve søkesystemet hadde de utviklet selv, og de var for tiden opptatt med overgang fra satsvis (batch) til direkte (on-line) behandling.

Av de mer individuelt betonte prosjekter som ble presentert, var det særlig på det språkstatistiske område at det ble presentert nye ting. Noen mente de hadde funnet en formel som kunne erstatte Zipf's lov som et mål på vokabularrikdom. Zipf's lov har blitt "forbedret" mange ganger, og problemet har alltid vært at tekstutvalgets størrelse har influert på forholdet. Den nye "lov", som ble påstått å være uavhengig av tekstutvalgets størrelse, var Sichel's fordeling, hvor de to parametre α og ω anga henholdsvis vokabularrikdom og vokabularkonsentrasjon. Når konferanserapporten foreligger, vil nok det empiriske materialet som ligger til grunn for påstanden, bli studert grundig.

Et annet interessant resultat, som riktignok ikke var en nyhet resultatmessig sett, var den statistisk-lingvistiske bekreftelse på kildesammensetningen i Mosebøkene, som protestantiske teologer påviste for over 100 år siden. Metoden som var benyttet var faktoranalyse utført ved hjelp av programpakken SPSS. Kriteriene var vesentlig frekvenser på ordklasseoverganger, d.v.s. hvor ofte f.eks. et substantiv ble fulgt av et pronomen.

En ekthetsundersøkelse av Aristoteles' etiske skrifter ble i en forelesning sterkt angrepet ved påvisning av at en utvilsomt ekte bok om metafysikk ville måtte erklæres uekte ved bruk av de anvendte kriterier. En fruktbar metodisk debatt må kunne foventes her, hvis ikke det kan påvises at metafysikk og etikk stilistisk sett er to forskjellige genre.

Av andre prosjekter som involverer metodologiske spørsmål av mer almen interesse kan nevnes et kanadisk ordboksprosjekt som tok klart

avstand fra at grammatikken ga den beste lingvistiske beskrivelse av et språk. Tvertimot var det ordboken som både logisk og kronologisk gikk foran grammatikken. Grammatikken var en nyttig generalisering og forkortelse av ordboken. Som man kan forstå måtte denne ordboken være adskillig rikholdigere enn tradisjonelle ordbøker.
Det vil føre for langt å gå inn på de mange prosjekter som ble omtalt. Interesserte kan få kopi av den "preprint"-katalog som ble utarbeidet til konferansen.

Undertiden får en inntrykk av den faglige utvikling skjer raskt innenfor denne disiplinen som f.eks. når en foreleser innleder med å si at det som står i "preprint" må glemmes fortest mulig fordi han i mellomtiden har fått nye resultater som har fremtvunget helt nye konklusjoner!

En viss dramatikk var det også på konferansen både saklig og fysisk. I saklig henseende gjaldt det en tysker som stolt presenterte sin nye algoritme for automatisk kontekst-avgrensning ved konkordansproduksjon, hvoretter en eldre professor, som bl.a. hadde vært med å lage den store Shakespeare konkordansen, står opp og forteller at han hadde utviklet og beskrevet den samme algoritme for 10 år siden.
På det fysiske plan besto dramatikken i et par mindre jordskjelv som fikk auditoriet til å riste kraftig.

Et negativt trekk som kan nevnes var det stramme tidsskjemaet som var satt opp. Man kjørte stramt på de oppsatte tider, og det resulterte ofte i at ikke alle ble ferdige med forelesningene sine, eller at de foretok en rask høytlesing av manuskriptet for å få sagt mest mulig.

Sammenfattende vil jeg si at konferansen var nyttig, fordi den ga impulser og bekreftelser på tendenser vi har merket hos oss. Her tenker jeg særlig på den vekt som flere la på bruk av programpakker for statistikk og konstateringen av den økende betydning som statistisk argumentasjon får i humanistiske miljøer.

DE NORDISKE DATALINGVISTIKKDAGER 1979.

Det andre nordiske møte for datalingvister ble holdt ved Institut for anvendt og matematisk lingvistik, Københavns Universitet, Amager i dagene 9. og 10. oktober. Konferanseleder var *Bente Maegaard*.

Det sier litt om veksten innenfor dette arbeidsfelt at møtet samlet ca. 70 deltakere, herav 15 fra Norge.

Denne gang hadde arrangøren, Den nordiske samarbeidsgruppe for datamaskinell språkbehandling, satt opp lemmatisering som et spesielt tema for konferansen, og en del av konferansens 2. dag var derfor konsentrert om foredrag viet dette emne.

Programmet var satt opp uten gruppeinndeling, så alle fikk anledning til å følge alt. Om aftenen 1. dag var det satt av tid til demonstrasjoner av EDB-opplegg og programutrustning gjennom bruk av Universitetets dataanlegg (RECKU).

I likhet med den foregående konferanse i Gøteborg i 1977 var det lagt opp til et arrangements-teknisk enkelt møte - også denne gang svarte denne formen godt til formålet: å gi informasjon om og synspunkter på arbeid som er i gang i de nordiske land.

Fra norsk side bør det noteres som gledelig at den aktive medvirkning herfra øker: På konferansen ble det gitt 3 norske foredrag og 2 presentasjoner av pågående EDB-arbeid. *Geir Berge* (Institutt for informasjonsvitenskap, Universitetet i Bergen) orienterte om konstruksjon av og oppslag i ordbøker på "små" datamaskiner. *Jostein H. Hauge* (NAVF's EDB-senter) presenterte planene om "Norsk tekstarkiv", *Knut Hofland* (NAVF's EDB-senter) presenterte "Lemmatiseringsmetoder i Ibsen-prosjektet", *Kolbjørn Heggstad* (PDS, Nordisk institutt, Universitetet i Bergen) demonstrerte sammen med *Bjørn Eide* og samarbeidspartnerne *Per-Bjørn Pedersen* og *Michael Gillow* program for tekstlagring og tekstbehandling utviklet i forbindelse med samarbeid med Det norske bibelselskap, *Eirik Lien* (EDB-tjenesten for humanistiske

fag, NLH, Trondheim) gjennomgikk og demonstrerte den programpakke for tekstbehandling som han har utviklet i Trondheim (PPTT).

De fleste av presentasjonene på konferansen refererte EDB-prosjekter med et praktisk tilsnitt, men enkelte tok også opp mer grunnleggende metodiske emner.

Peter Bøgh Andersen orienterte om FANGORN-prosjektet som har til mål å formalisere tekstbeskrivelser og lage en abstrakt beskrivelse av strukturerte handlingssekvenser ("plots"). Systemet bygger på SIMULA og LISP.

Gert Schmelts Pedersen ga en oversikt over sitt arbeid med konseptuelle grafer. Arbeidet er basert på databaseteori og legger vekt på utvikling av en datamodell for å forklare forholdet mellom objektene (det som skal beskrives),deres avbildning og deres representasjon i en database. Det ble gitt eksempler på bruk av konseptuelle grafer ved meningsanalyse av enkle setninger.

Bjarner Svejgaard,Hasse Hansson og Gustav Leunbach ga presentasjoner av metoder for ord-deling og måter å evaluere ord-delingsalgoritmer på. Ved å legge inn i datamaskin de vanlige reglene for ord-deling kan en oppnå relativt gode resultater med automatisk ord-deling. De mer faglig interessante problemer oppstår når en skal formalisere ord-delingsprinsipper for de tilfeller som ikke fanges opp av de klassiske regler og utvikle algoritmer for behandling av sammensatte ord. Blant annet brukes statistisk informasjon om vokal- og konsonantkombinasjoner, prefikser, suffikser o.l. som hjelpemiddel til å forbedre resultatene, eventuelt sammen med bruk av ordbok over "vanskelige tilfeller".

Eric Grinstead orienterte om EDB-arbeid ved Østasiatisk Institut, Københavns Universitet. Aktivitetene omfatter arbeid med japansk poesi, dokumentasjon av kinesisk malerikunst 1849-1979 og en ordbok over kinesiske binomer (tegn som består av 2 elementer) med japansk oversettelse. EDB-arbeidet blir drevet i nær kontakt med beslektede institusjoner i Tokyo som på sin side nyter godt av at Japan er langt fremme i den teknologiske utvikling på feltet.

Cecilia Thavenius beskrev status for prosjektet Engelskt Talspråk hvor et utvalg talespråkstekster legges til rette for data-maskinell behandling. Tekstutvalget er hentet fra The Survey of English Usage ved University College London. Tekstmassen består av 34 tekster á 500 ord hentet fra opptak av spontan tale. Tekstene er i ortografisk transkripsjon og i prosodisk analysert form. De vil være tilgjengelige både i bokform og på magnetbånd for bruk i datamaskin. Nærmere opplysninger om datamaskinversjonen finnes forøvrig i ICAME NEWS No. 3, Oct. 1979. Prosjektet vil nå ta fatt på arbeidet med grammatisk merking av dette materialet.

Suzanne Hanon hadde stilt seg oppgaven å finne fram til de prinsipper som brukes i konkordanser og ordbøker når arbeidet består i å lemmatisere ordstoff. Det synes å være lite overrensstemmelse mellom de fremgangsmåter som velges. Lemmatisering består i å klassifisere og redusere ordformer til mindre kompliserte grunnformer, og arbeidet synes ofte å bygge på visse konvensjoner av morfologisk og semantisk art innen et språkområde. Men detaljstudium viser at mange forhold overlates til skjønn og det som er praktisk hensiktsmessig i språklig forstand.

Benny Brodda redegjorde for sitt programsystem BETA, som er et programverktøy for lingvistiske eksperimenter. Han tok særlig opp bruk av BETA-systemet ved morfologisk segmentering av ord i en tekst. Et hovedproblem i systemarbeidet er å finne et kompromiss mellom lingvistens ønske om generelle regler og EDB-ekspertens ønske om enkle regler. Det ble gitt instruktive eksempler på bruk av systemet på svensk, finsk og latinsk ordmateriale.

Rolf Gavare gjorde opp status for lemmatiseringsarbeidet de siste 10 år ved Språkdata, Gøteborgs Universitet. Han gjennomgikk de metoder for lemmatisering uten leksikon som er basert på analyse av store tekstmengder og de prinsipper en i dag ønsker å arbeide etter i slikt arbeid ved Språkdata. Til forskjell fra tidligere tenker en seg nå at homografsepareringen inngår i lemmatiseringsfasen. Arbeidet skjer interaktivt slik at maskinens forslag til lemmamarkering basert på bruk av endelses- og bøyningslister, enkelt kan korrigeres på dataskjerm. Å etablere grunnformer til

ord som bare forekommer en gang i et materiale, må derimot skje ved bruk av morfem- og/eller stammeleksikon. Det foreligger litteratur om emnet utarbeidet ved Språkdata.

Bjørn Ellertson refererte fra forsøk med automatisk lemmatisering av islandsk. Arbeidet bygger på bruk av Martin Kays program for morfologisk analyse presentert i Bergen i 1973. Den islandske variant av dette (Basic) program og bruk av ordbok ga gode resultater ved forsøkene som ble gjort på en PDP 11/60.

Henrik Holmboe var opptatt av begrepet lemmatisering og konstaterte at det ikke er opptatt i lingvistisk terminologi. Den prosess det her gjelder er derimot velkjent i språklig arbeid. Lemmatisering kan enten være å henføre en ordform til en annen (overordnet) form i en tekst eller gjennom utnyttelse av annen grammatisk og semantisk viten bestemme tilhørigheten (mindre - liten). Problemene i nordiske språk er enkle sammenlignet med hva de er i agglutinerende språk der langt flere grammatiske og semantiske forhold uttrykkes morfologisk.

Hanne Ruus tok for seg de problemer som oppstår ved automatisk lemmatisering og kom inn på hvordan semantiske opplysninger burde kunne anvendes for å separere homografer. I datalingvistisk forskning er det gjort en del teoretisk arbeid med bruk av semantiske modeller i språkanalyse. De semantiske strategier som er utviklet innen området kunstig intelligens (artificial intelligence), er i dag vanskelige å utnytte i praktisk arbeid på dette feltet.

Bo Ralph drøftet leksikologi som disiplin i datalingvistikken. Som oftest blir datamaskinelt leksikonarbeid bestemt av de umiddelbare, praktiske forhold man har med arbeidet. Men leksikologien er en så sentral del av (data)lingvistikk at det er behov for å gjennomtenke de språkmodeller som vår praksis reflekterer og å sette opp alternative modeller med et utvidet siktemål: å avbilde hvordan mennesker fungerer rent språklig. Da må ikke minst psykologiske realiteter trekkes inn i vurderingen. Foredragsholderen tenkte seg det "menneskelige" leksikon som et flerdimensjonalt nett-

verk som inneholder mange sorter informasjon: betydning, form, funksjon, synonymer, antonymer, eksempler på språkbruk o.s.v. Å skape en slik ordbok er en kjempeoppgave, men vi kan starte med en enklere utgave uten å gi slipp på kravet at det skal være en psykologisk rimelig leksikonmodell.

I programmets siste foredrag orienterte *Hanne Ruus* og *Bente Maae-gaard* om planleggingen av det europeiske maskinoversettelses-system (EUROTRA) som drives av forskere innen datalingvistikk og maskinoversettelse i EF-landene. Arbeidet startet i februar 1978 og har som mål et oversettelsessystem som skal kunne oversette mellom alle de 6 EF-språk og også utvides til å omfatte flere. Tanken er å utvikle mest mulig av systemet separat og lokalt for det enkelte språk. Oversettelsesprosessen er delt i 3 deler: analyse, overføring og generering, hvorav de to første utvikles separat for hvert enkelt utgangsspråk. Prinsippene og programmene for selve oversettelsen, overføringen, må derimot utarbeides for hvert språkpar for seg. I prosjektet EUROTRA vil en særlig basere arbeidet på den virksomhet som til nå har foregått i oversettelses-prosjekter i Grenoble og Saarbrücken.

På konferansen var det også drøfting av datalingvistikken som studieemne og gjort forsøk på å konkretisere de ulike delområder som faller inn under denne betegnelse. Det ble her en nyttig drøfting av datalingvistisk arbeid uten at drøftingen egentlig munnet ut i forslag til en autoritativ definisjon. Flere mente at det karakteristiske for datalingvistikk er at den er en databehandlende språkvitenskap: målet er å vinne lingvistisk innsikt ved hjelp av det forskningsverktøy og de forskningsmetoder som moderne data-teknikk gir.

Det ble på konferansen besluttet at neste nordiske datalingvistikk-møte skal legges til Trondheim hvor EDB-konsulent Eirik Lien vil være vert. De fleste deltakere ønsket seg om lag samme tidsrom for neste møte. Deltakerne og andre interesserte ble oppfordret til å gi forslag til samlende temaer for neste konferanse. Til slutt ble det opplyst at den nordiske samarbeidsgruppen for data-maskinell språkbehandling vil søke om å få arrangere et nordisk

forskerkurs på Island i 1981.

Det er planen å utarbeide en konferanserapport med alle foredragene. Rapporten kan bestilles hos *Bente Maegaard, Institut for anvendt og matematisk lingvistik, Njålsgade 96, DK2300, København S.*

LA JOLLA CONFERENCE ON COGNITIVE SCIENCE, SAN DIEGO, 13.-16. AUG.

Denne internasjonale konferansen i kognitiv vitenskap (cognitive science) var den første i sitt slag. Den ble arrangert på University of California, San Diego av det nystartede Cognitive Science Society. Denne forening har til hensikt å være et tverrvitenskapelig faglig forum for forskere som arbeider innen ulike områder av kognitiv vitenskap, blant annet områder som kunstig intelligens (artificial intelligence) kognitiv psykologi, almen og datamaskinell lingvistikk og filosofi. På den første konferansen hadde arrangørene klart å få flere av de fremste forskere på feltet til å gi foredrag. Totalt deltok det ca. 600 personer, de fleste fra USA.

Av foredragsholderne kan nevnes *George Lakoff, Marvin Minsky, Roger Schank, John Searle* og *Terry Winograd.*

Det ble gitt en serie hovedforedrag og bortimot 30 foredrag innenfor 6 synposier hvor målet var å gi et bilde av stillingen på forskningsfeltet i dag. Symposiene omfattet emner som

Belief Systems
Cognitive Science and Education
Discourse
Human Development
Language Processing
Psychology of Categorization

Det ble opplyst på konferansen at foredragene senere vil komme i det nystartede tidskriftet Cognitive Science.

En rekke av foredragene omhandlet simulering i datamaskin av

kognitive prosesser og generelle metoder for studiet av symbolsystemer. Videre ble det gitt referat fra en serie prosjekter der målet er å legge til rette i datamaskin semantiske opplysninger og strukturerte kunnskapsfelt (knowledge representation) som hjelpemiddel ved eksempelvis automatisk tekstanalyse ("story understanding systems") og spørsmål/svar-systemer i naturlig språk. Flere foredrag tok for seg utvikling av CAI-systemer (Computer Assisted Instruction) og bruken av CAI-systemer ved studiet av kognitive prosesser.

Ikke alle foredragene var basert på arbeid der datamaskin ble benyttet. Noen av de mest interessante tok for seg utvikling av kognitive prosesser hos mennesket, bl. a. hos barn i løpet av det første leveåret.

For den lingvistisk interesserte var det to høydepunkter: *George Lakoff* ga en inspirerende forelesning om språkbruk som metaforbruk og *Teun A. van Dijk* ga en omfattende presentasjon av tekst-grammatikk.

Den spesielt interesserte kan ved henvendelse til NAVF's EDB-senter få kopi av en mer omfattende reiserapport utarbeidet av *Jostein H. Hauge* og *Knut Hofland*.

Sixth International ALLC Symposium

Computers in Literary and Linguistic Research

28 March - 3 April 1980 University of Cambridge, England

Final Call for Papers

The Sixth International ALLC Symposium on Computers in Literary and Linguistic Research will be held at the University of Cambridge, England, from Friday 28 March to Thursday 3 April 1980. Papers are invited for presentation at the Symposium. It is anticipated that they will be in the following categories:

authorship studies, concordances, data bases, education, input/output, language-oriented studies, lexicography, literary statistics, metrics, quantitative linguistics, software, stylistic analysis, textual criticism.

It will be possible to arrange discussion groups on the evenings of Monday 31 March and Tuesday 1 April. Please notify the Secretary accordingly.

Anyone wishing to present a paper should send three copies of a typed abstract of approximately 300 words to the Secretary to be received before 1 October 1979. The Organizing Committee will then select suitable papers from the abstracts offered, and request the submission of complete draft papers to be received before 1 January 1980. Speakers will be informed by 1 February 1980 of the acceptance of their papers.

Papers should be original. The Organizing Committee reserves the right to consider papers presented at the Symposium for publication. Abstracts (three copies) should be sent to Dr J. L. Dawson, Secretary, 1980 Symposium, Literary and Linguistic Computing Centre, Sidgwick Site, Cambridge CB3 9DA, England (telephone 0223 356411 extension 37).

Introduction to Computing in the Humanities

The next ALLC Summer School, in the successful series organized by Mrs Susan Hockey, has been scheduled to take place at the University College of Wales, Aberystwyth, from Monday 14 April to Saturday 19 April 1980. The subject of the Summer School will be 'Introduction to Computing in the Humanities'. The major element of the course is programming in SNOBOL4, together with the use of concordance and information retrieval packages. Information may be obtained from Mr G. V. Appleton, Computer Unit, Llandinam Building, University College of Wales, Aberystwyth, Dyfed SY23 2AX, Wales, UK.

University of Cambridge

The University of Cambridge is one of the oldest universities in the world. Although lectures are provided by the university, accommodation and individual tuition are provided by the colleges, of which there are at present thirty-one. The oldest college dates from about 1284, and buildings such as King's College Chapel, the Wren Library of Trinity College, and the Pepys Library of Magdalene College, are world-famous.

Literary and Linguistic Computing Centre

The Literary and Linguistic Computing Centre was founded in 1963 on the initiative of Professor Roy Wisbey, to provide a data preparation and computing service for all humanities faculties within the university. The present Director is Dr Ronald Popperwell, Dr John Dawson being responsible for the day-to-day running of the Centre.

Association for Literary and Linguistic Computing

Co-founded in 1973 by Professor Roy Wisbey and Mrs Joan Smith, who became its first Chairman and Secretary, respectively. The Officers are currently Mrs Joan Smith (Chairman), Dr John Dawson (Secretary), and Dr Rex Last (Treasurer).

Programme

There will be an opening session after tea on the day of arrival, Friday 28 March 1980, followed by a reception in the University Combination Room at which the Vice-Chancellor of Cambridge University will be present. All sessions will take place in the Little Hall, Sidgwick Site, Cambridge, opposite Newnham College. There will be no parallel sessions.

Sunday 30 March will be free for sight-seeing. A guided walking tour of Cambridge will be arranged for the morning. In the afternoon there will be an excursion to the city of Ely when it will be possible to visit Ely Cathedral. Afternoon tea will be served. A small additional fee will be charged to those participating in these events, both of which are optional.

An Elizabethan Feast will be held in King's College hall on the last evening of the Symposium, Wednesday 2 April. This is included in the Symposium fee.

Demonstrations of text processing will be given in the Literary and Linguistic Computing Centre (LLCC), and participants will be able to visit the Computer Laboratory. There will be a display of books and examples of work to which those attending are invited to contribute. The ALLC Archives, housed at the LLCC, will be available for consultation. In the foyer of the Little Hall there will be an on-line terminal display from which a data base of Symposium and local information may be accessed.

Accommodation

Accommodation will be in single study-bedrooms in Newnham College, across the road from the Little Hall and the Literary and Linguistic Computing Centre, and a few minutes' walk from the centre of the city. (Please note that children cannot be accommodated in Newnham College.) Each room has a wash basin; bathrooms and/or showers and WC facilities are near the rooms. The College bar will be open during normal licensing hours, and the lounge is available for informal gatherings and conversations.

Hotel accommodation is scarce and expensive in Cambridge, and the organizers of the Symposium regret that they cannot make hotel reservations. However, a leaflet giving details of hotels in Cambridge may be obtained from the Secretary. Delegates are advised to book accommodation for the Symposium well in advance.

Fees

The Symposium fee will be £20 per person, reduced to £17 for members of ALLC or ACH. This is inclusive of the reception, the Elizabethan Feast, and tea and coffee throughout the Symposium.

The cost of accommodation and meals in Newnham College (breakfast, lunch, and dinner) will be approximately £14 per person each day. A deposit of £30 per person is required, irrespective of accommodation or attendance at the Symposium sessions.

Intending participants are asked to complete the tear-off slip below, and to send it with the correct deposit to Dr J. L. Dawson, Secretary, 1980 Symposium, Literary and Linguistic Computing Centre, Sidgwick Site, Cambridge CB3 9DA, England. Cheques drawn on a British bank, and international money orders, should be in sterling, and made payable to 'Cambridge Computing Symposium'.

I/We wish to atteno the Symposium to be held at Cambridge from 28 March to 3 April 1980.

Names: ....

....

Mailing address (including post code or Zip code):

.....

.....

Accommodation required:

number of rooms ..... number of nights.....

from..... to .....(inclusive)

Comments (e.g. special diet):

.....

.....

A deposit of £ ..... is enclosed (£30 per person), made payable to 'Cambridge Computing Symposium'.

10. DATALINGVISTIKKMØTE I DANMARK.

Det 10. møte for danske forskere innenfor datamaskinell lingvistikk ble holdt 22. og 23. november på Odense Universitet. Arrangør er *Suzanne Hanon*, Romansk institutt.

Jubileumsmøtet ble konsenstrert om temaer i forbindelse med oversettelse: oversettelsesteori, oversettelsespraksis og datamaskinell oversettelse.

Det er meningen å utgi foredragene i en rapport. Den vil koste ca. d.kr. 100,- og kan forhåndsbestilles hos Suzanne Hanon, Romansk institutt, Odense Universitet, Niels Bohrs Allé, 5230 Odense M.

EDB-TJENESTENE FOR HUMANISTER I BERGEN OG TROMSØ.

BERGEN.

EDB-seksjonen ved HF-fakultetet, Universitetet i Bergen er i høst bemannet med egen EDB-konsulent. Cand. theol. *Roald Skarsten* er ansatt i stillingen. Han har tidligere bl.a. arbeidet som førstekonsulent ved NAVF's EDB-senter for humanistisk forskning.

TROMSØ.

NAVF's EDB-senter har i høst opprettet et korttidsengasjement for en EDB-konsulent ved Institutt for språk og litteratur. Cand. philol. *Gunnar Thorvaldsen*, som er engasjert i stillingen, vil gi veiledning til humanister i bruk av EDB, og vil i høst særlig arbeide med tilretteleggingsarbeid for bruk av EDB i målførearbeid og i tekstanalyse.

Det vil dessuten bli holdt kurs om generelle tekstbehandlingsprogrammer.

COMPILING

MEDLEMSBLAD FRA NORDISK SAMARBEIDSGRUPPE FOR DATAMASKINELL SPRÅKBEHANDLING.

På det nordiske datalingvistikkmøte i København i oktober ble *Henrik Holmboe*, Institut for lingvistik, Århus Universitet, valgt til redaktør for de 2 neste år.

Den avtroppende redaktøren *Bente Maegaard*, Institut for anvendt og matematisk lingvistik, Københavns Universitet meldte om god interesse for bladet (som er gratis), men skuffende liten mengde faglige bidrag. Dette kan rettes på ved å sende bidrag til den nye redaktøren, som har adresse: Otto Ruds gade 67-69, 8200 Århus N.

COLING 1980

INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS.

COLING 1980 vil bli holdt i Tokyo i tiden 30.9.-4.10.

Det foreligger pr. dato få opplysninger om konferansen hvor prof. *M. Nagao* er leder for den lokale arrangementskomite. Forlydender forteller at det vil bli satt opp spesielle charterreiser for deltakere fra Nord-Europa for å få reisekostnadene ned på et akseptabelt nivå for dem som er avhengig av bevilgninger fra universiteter og forskningsfond.

Nærmere opplysninger vil komme i Humanistiske Data.

SENTERETS RAPPORTSERIE

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

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

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

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

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

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

Rapport nr. 7 *Ivar Fonnes: Norsk landbruksordbok. Prosjektrapport om databehandling og tilrettelegging for trykking.*
Utkommer desember 1979.
Pris kr. 20,-

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

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

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

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

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

S U M M A R Y

HUMANISTISKE DATA PÅ NY

In his editorial note, *Jostein H. Hauge* expresses hopes that the publication will be a useful meeting-point for those interested in computing in the humanities in Norway. After a 2-years' break, the aim is to issue Humanistiske Data 2-4 times a year. An invitation is extended to the research community to publish articles, reports and commentaries in the publication, in any of the many fields of application in the humanities. In particular, the editor urges readers to fill in the attached forms for project registration, which will be used as source material for an overview in the next issue on the "state of the art".
Humanistiske Data is issued free of charge and subscriptions are invited.

STATUTTER FOR NAVF'S EDB-SENTER FOR HUMANISTISK FORSKNING

The Norwegian Computing Centre for the Humanities in Bergen was established in 1972 as a 5-year programme to promote the use of computers in the humanities in Norway. At the end of the first period the Norwegian Research Council for Science and the Humanities found that there would be a long-range need for assistance and research in the field and accordingly established the Centre on a permanent basis. An overview of the background, aims and main fields of interests of the Centre is given.

IBSEN-KONKORDANS

In 1978 work started on a complete lemmatized concordance of Henrik Ibsen's plays and poems. The project is being carried out at the Norwegian Computing Centre for the Humanities under the leadership of Professor *Harald Noreng*. The concordance will be finished in 1981 and the main publication form will be on microfiche.

NORSK TEKSTARKIV

Norsk Tekstarkiv is a project aiming at collecting and distributing modern Norwegian texts in computer-readable form. The text archive will be located in Bergen and will be in operation from 1980 onwards. The bulk of the material will be texts available at publishing houses on paper tape or magnetic tape. In addition, the text archive will prepare texts of different kinds of specialized research purposes.

NORSK LANDBRUKSORDBOK - DEFINISJONSORDBOK OG DATABANK

As a result of more than 20 years' work a comprehensive new dictionary of agriculture was issued in February 1979. The dictionary in 2 volumes, covering 980 pages, contains scientific definitions and synonyms in 6 languages (Danish, English, Finnish, German, Icelandic, Lappish and Swedish). To improve the practical use of the dictionary there are separate indices for each language giving references to the entries in Norwegian. In his article *Ivar Fonnes*, EDP consultant at the Faculty of Arts, University of Oslo, gives an account of the computer work he did in the composition and publishing of the dictionary, as the result of which the manual archive was transformed into a computer data base for photocomposition. Future plans include collaboration and exchange of material with relevant dictionary projects of the European Community in Luxemburg.

EDB SOM HJELPEMIDDEL VED DE ARKEOLOGISKE UTGRAVINGER I GAMLEBYEN, OSLO

In 1970 archaeological excavations commenced in Gamlebyen, a part of Oslo dating from the Middle Ages. Since all the descriptions of the artefacts had to be typed, it was decided to make the typing format compatible with the requirements of OCR so that the material could be computerized. The project co-operates with the Norwegian Computing Centre for the Humanities. At present a database of the material has been established for the use of the NOVA*STATUS Information Retrieval System. A series of individual research projects

based on the findings are underway making use of the computerized material.

REGISTRERINGSSENTRAL FOR HISTORISKE DATA, UNIVERSITETET I TROMSØ

During the last few years a group of historians at the Institute for Social Sciences, University of Tromsø has designed the organization and services of a national historical data archive in Tromsø.

Based on inquiries addressed to the teaching and research staff in history at the Norwegian universities and colleges, a detailed scheme has been devised for the selection of primary historical materials from the main nation-wide censuses and the clerical registers covering the past 100 years.

The aim, says *Gunnar Thorvaldsen*, EDP-consultant of the project, is to provide researchers of Norwegian history with adequate computer tools for the study of micro-history, a method which focuses on individuals for the description of the behaviour and living conditions of groups and classes of previous ages.

Co-operation has also been developed with state and county authorities and plans are currently being considered for the use of surplus labour from the automatization of long-distance telephone operation in the county of Troms in the data-preparation stage of the project.

ICAME

Information is given on the International Computer Archive of Modern English (ICAME). The aims of the organization are 1) to collect and distribute information on English language material available for computer processing, 2) to collect and distribute information on linguistic research, completed or in progress, on the material, 3) compile an archive of text corpora for distribution to research scholars.

ICAME co-operates with the Norwegian Computing Centre for the Humanities. At present a number of text corpora are available. More information is given in the ICAME NEWS, editor Dr. *Stig Johansson*, Department of English, University of Oslo.

REPORT FROM A SYMPOSIUM ON GRAMMATICAL TAGGING OF ENGLISH TEXT CORPORA

Under the joint auspices of The International Computer Archive of Modern English (ICAME) and the Norwegian Computing Centre for the Humanities, an international symposium on tagging techniques was held in Bergen on March 29-30, 1979. The meeting was attended by 37 participants from 10 countries. In his report Dr. Stig Johansson, the chairman of ICAME, reviews the projects presented at the symposium and gives references to publications giving more detailed information on each of the particular projects.

COMPUTER-SENTER FOR GRESK FILOLOGI (THESAURUS LINGUAE GRAECAE)

Professor *Knut Kleve*, Classics Department, University of Oslo, gives a report on a visit to the project Thesaurus Linguae Graecae at the University of California. The aim of the project is to convert the whole of the Greek literature from classic antiquity to computer readable form.

KONFERANSE OM DATATENESTE FOR OG DATASAMARBEID MELLOM DEI KUNST- OG KULTURHISTORISKE MUSÉ, USTAOSET HØYFJELLSHOTELL 23-24. OKTOBER 1979

A report is given of a national conference at Ustaoset. The background of the conference was that several cultural museums and other institutions during the last few years have carried out pilot projects testing various computational methods in collecting, analysing and presenting museum material for external use and in-house research projects and developmental work. As one result a

standard card for object registration that also allows OCR- reading has been designed. At the conference at Ustaoset, an overview was given of the computational work carried out so far. On this basis the establishment of a service institution for computing in the museums was discussed.

THE 4TH INTERNATIONAL CONFERENCE ON COMPUTERS AND THE HUMANITIES (ICCH/4) AND DATA BASES IN THE HUMANITIES AND SOCIAL SCIENCES, HANOVER, NEW HAMPSHIRE.

The two conferences were held at Dartmouth College, Hanover, New Hampshire August 19-24. A brief description is given of the range of topics covered by the conferences. Norwegian readers are referred to a more comprehensive conference report.

17TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, SAN DIEGO, AUGUST 11-12, 1979.

A brief report is given on the 17th meeting held at UCSD, covering fields such as language structure and parsing, knowledge organization and application, dialogue, applications. The conference was very well organized and the proceedings were available from the outset.

LA JOLLA CONFERENCE ON COGNITIVE SCIENCE, SAN DIEGO, AUGUST 13-16

A report is given on the first international conference in cognitive science, where about 600 participants gathered at the University of California, San Diego. The main themes were: belief systems, cognitive science and education, discourse, human development, language processing, psychology of categorization. The speeches will later be published in the journal Cognitive Science.

KONFERANSE OM LITTERÆR OG LINGVISTISK DATABEHANDLING, TEL AVIV 22-27 APRIL 1979

A report is given on the Conference on Computers in Literary and Linguistic Research, Tel Aviv April 22-27.

DE NORDISKE DATALINGVISTIKKDAGER 1979

The Second Nordic Meeting on Computational Linguistics was held at the University of Copenhagen on October 9-10. Interest in the field of datalinguistics is growing and this time about 70 participants gathered from Denmark, Iceland, Finland, Norway and Sweden. A variety of individual and group research projects were presented and on the first night a number of systems for linguistic and literary computing were demonstrated. The main topic at this year's meeting was the methodology of lemmatization to which the second day of the meeting was devoted.

6TH INTERNATIONAL ALLC SYMPOSIUM ON COMPUTERS IN LITERARY AND LINGUISTIC COMPUTING, CAMBRIDGE, ENGLAND

Call for papers and general information.

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

TRYKK OG PAPIR . 5095 ULSET, BERGEN