humanistiske data

Meldingsblad for
NAVF's EDB-senter
for humanistisk forskning

Norges Almenvitenskapelige Forskningsråd

NR.
1-2
1976

INNHOLD:

MEDARBEIDERE I DETTE NUMMER:

ESKIL HANSEN, forskningsstipendiat ved Institutt for nordisk språk og litteratur, leder for talemålsundersøkelsen i Oslo 1971–76.

SVEIN LIE, forskningsstipendiat ved Institutt for nordisk språk og litteratur, Universitetet i Oslo.

DIETER WICKMANN, Dr., Institutt fur Matematisch Empirische Systemforschung, Aachen.

HUMANISTISKE DATA
blir utgitt av
NAVF's EDB-senter for humanistisk forskning
i Bergen.

Senterets leder,
Jostein Hauge, har det redaksjonelle ansvar for meldingsbladet.

De som ønsker å få bladet tilsendt,
kan bestille det ved henvendelse
til senterets adresse:
Villavei 10,
Boks 53,
5014 Bergen–Universitetet.

Innlegg kan sendes til samme adresse.

Sats og trykk:
Universitetets trykkeri,
Bergen

REDAKTØRENS SPALTE

Virksomheten innenfor humanistisk databehandling øker år for år. Stadig flere tar i bruk datamaskin som et hjelpemiddel i sitt arbeid samtidig som datamaskinen søkes anvendt på nye oppgavefelt.

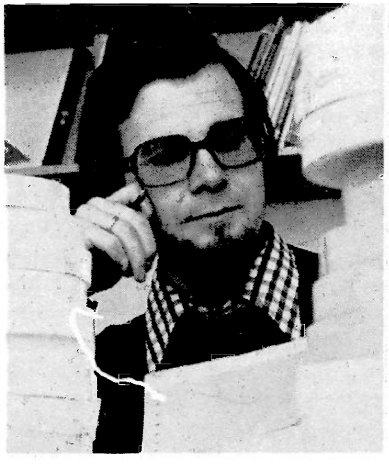
I den siste tiden har en også sett klare tegn på at ikke bare NAVF, men også universitetenes styrende organer for alvor har akseptert datamaskinbruk blant humanistene.

Representanter for humanistene har riktignok allerede i noen år vært med i styrene for universitetenes EDB-sentra. Nå kommer humanistene også stadig sterkere inn når utrustningen av nye dataanlegg skal fastlegges. HF-avdelingene ved alle universitetene har i dag utformet planer om konsulentordninger for humanister som ønsker å ta i bruk datamaskinelle metoder. Det er grunn til å tro at det i løpet av de to nærmeste år kan etableres et par faste konsulentstillinger ved universitetene. Utgifter i forbindelse med datautstyr og datamaskinbruk kommer sterkere inn på de ordinære budsjettene.

De fleste av de humanistiske fag arbeider med tekster. Det er derfor viktig å vite at datamaskinene i lengre tid bl. a. har undergått en utvikling fra elektroniske regnemaskiner til elektroniske tekstbehandlingsmaskiner. Som det er nevnt annetsteds i dette nummer, regner EDB-eksperter med at i 80-årene vil over 50% av den totale EDB-kapasitet (i verden) brukes til »word processing». Hvis det er sant at » når det regner på presten, så drypper det på klokkeren», vil humanistene gå løfterike tider i møte fordi det vil bli stadig mer å hente – metodisk og teknisk – fra andre EDB-felt.

Som det går fram av dette nummeret av Humanistiske Data, finnes det ved alle universitetsanleggene i dag programpakker for tekstbehandling som brukerne kan nytte selv etter kort opplæring. Er de humanistiske grunnbehov dermed dekket? Nei, det vil svært ofte oppstå spesialiserte behov som må få egne løsninger, helst i form av spesialutviklet programutrustning. Utenfor tekstbehandlingsfeltet er det store oppgaver som utfordrer, ikke minst innenfor gjenstandsfagene og i arkivinstitusjonene. Disse arbeider med svære datamengder, til dels som et resultat av vitenskapelig tilrettelleggingsarbeid. Databaseteknikk kan her stå som et stikkord for de løsninger EDB-ekspertene gir på utfordringene knyttet til moderne informasjonsbehandling.

Utviklingsarbeidet innenfor humanistisk databehandling vil stadig gå videre. Det er grunn til å tro at utviklingen har en spiralform: Opparbeidelse av kompetanse på ett nivå i de humanistiske EDB-miljøer, vil utløse behov for utviklingsarbeid og assistanse på et høyere nivå. Om det finnes et »høyeste nivå», er vel heller tvilsomt.



DATABEHANDLING I TALEMÅLSFORSKNING

Erfaringer fra et prosjekt

AV ESKIL HANSSEN

I tidsrommet 1971–76 foregikk et større språksosiologisk forskningsprosjekt, Talemålsundersøkelsen i Oslo, ved Institutt for nordisk språk og litteratur ved Universitetet i Oslo. I dette prosjektet har en brukt EDB både til lagring og behandling av språklige tekstdata, og til kodete data. Jeg vil her gjøre rede for hvordan databehandling ble brukt i dette prosjektet, i grove trekk.

Formålet med prosjektet har vært å granske språksosial variasjon i talemålet til noen utvalgte, innfødte Oslo-boere (i alt 48 personer). Materialet består av sammenhengende språktekster som er skrevet av etter lydbandinnspilte intervjuer. Den språklige analysen gjelder i første rekke syntaks, men også ulike aspekt ved ordbøyning, lydsystem og ordforråd har vært tatt opp. Databehandling har vært nytta i alle disse typene av undersøkinger, så det er tale om nokså ulike og varierte former for arbeidsoppgaver. Vi har brukt programsystemet TEXT, som EDB-konsulent Ivar Fonnes er ansvarlig for. Databehandlinga har stort sett vært basert på de programmene som var tilgjengelige da prosjektet begynte, men det er også laget noen spesielle programmer.

En fordel med datamaskinell lagring av tekst er at en kan bruke programmer som gir ulike typer utskrifter, for bestemte formål. En maskinskrevet eller trykt tekst er som regel ikke særlig hendig å bruke til analyse, det er fordelaktig å ha spesielle typografiske arrangement av teksten. F.eks. at hver analyseenhet settes på egen linje, slik at en kan skrive ut analyseresultatet på utskriftarket. Eks.:

ANALYSE			
Tekst	Helhetstype	Funksjon	Ordklasse
DET	Setning	Subjekt	Pronomen
STOD		Predikat	Verb/Sterk/Pret
I		Adverbial	Preposisjon
AVISEN			Subst/Best/Ent

Dette kan synes nokså trivielt, men er likevel viktig i praksis, fordi det letter arbeidet i stor grad og minsker mulighetene for feil og uklarheter i analysen. Det blir også lettere å bearbeide analyseresultatene videre, f.eks. gjennom omkoding og deretter behandling av kodete data.

Den syntaktiske analysen av TAUS-materialet er en klassifikasjon av enheter på setningsnivå, såkalte makrosyntagmer. De fleste makrosyntagmene er konstruksjoner av fleire ord, og setninger er den vanligste typen i et materiale som TAUS. Under analysen har vi også tatt omsyn til bestemte aspekter ved makrosyntagmenes konstruksjon. På grunnlag av denne analysen har vi så foretatt en kvantitativ undersøking for å finne ut hvorvidt visse syntagme- og konstruksjonstyper blir brukt i forskjellig grad av personer med ulik sosial bakgrunn. Dette foregår i fleire trinn, og som kombinasjon av manuell og maskinell analyse. Den syntaktiske klassifikasjonen omfatter tre nivåer, og klassifikasjonen av hvert makrosyntagme blir kodet om til et tresifret tall. Hvert siffer svarer til et bestemt nivå, og ulike tallverdier uttrykker

de aktuelle avariablene. F.eks. er tallet 111 kode for rettkonstruert setning, og når koden forsynes med et bestemt nummer, veit en hvilken setning det er tale om i vedkommende tekst. (I prosjektet har vi også kodet andre ting, men det trenger jeg ikke komme inn på her.) Jfr. modell for klassifikasjon side 6. De kodete dataene er så lagt inn på en egen fil, i form av en matrise, der enhet svarer til den fortløpende nummerering av makrosyntagmene i tekstene, og variablene er fordelt med et visst antall til hver tekst. Analysedataene er altså holdt atskilt fra tekstdata, noe som vi har sett på som en fordel. Det er likevel praktisk å ha et samband mellom tekstdata og kodete data, og dette er mulig takket være et spesielt program. Det gjør det mulig å gå fra kodete dataenheter til de tilsvarende tekstenheter. På denne måten kan en få utskrift av teksten sammen med den klassifikasjon som hver analyseenhet har fått. En har også muligheter for å bruke ulike filtre på kodete data, og dermed blir det mulig å få utskrevet makrosyntagmer med en bestemt struktur. Vi er f.eks. interessert i å få utskrevet alle ufullførte setninger hos en bestemt intervjuperson. Vi angir kode for denne syntagmetypen og for person, spesifiserer hvilke variable vedkommende tekst har, og får da en utskrift som vi her viser et utsnitt av.

```
MSnr.Kode Tekst
  25  2125 DE Æ– JO VARMT VANN Å:
           KJØKKEN ME ÅPPVASKBENK
           Å SÅNT SÅM JÆ:==/
  43  2125 DE: = VA– VEL:/
  64  2125 D– Æ–: =/
 109  2125 SÅ DE VAR =E=:/
 121  2125 FRA  UNNER  KRIGEN,  SÅ
           KAN  JÆI HU– :/
```

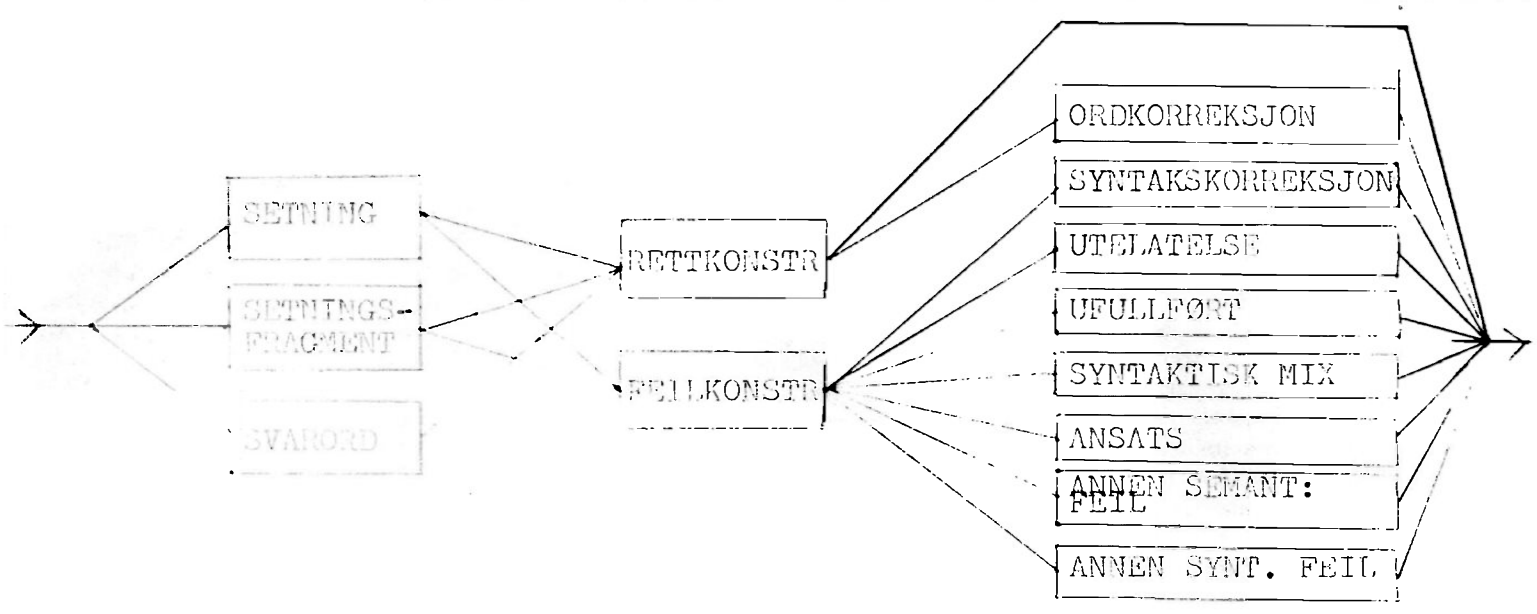
(Eksemplene er her gjengitt i den skrivemåten som er brukt i prosjektet, med spesielle tegn for tonegang, pauser o. a.)

Den kvantitative undersøkinga skal gi svar på spørsmålet om det er noen skilnader i bruksmønsteret eller frekvensen av visse syntaktiske kategorier i talen til de enkelte personene i utvalget. Det foregår trinnvis, og første trinn er opptelling av frekvenser av bestemte syntaktiske kategorier, og beregning av relative frekvenser. Det er blitt gjort med et statistikkprogram som opererer på kodefilen, og som er uavhengig av tekstdata. I neste omgang har vi regnet ut bruksmønstre for grupper av personer og på andre måter bearbeidd det statistiske materialet (beregnet gjennomsnittlige frekvenser, gjort rangeringer osv.). Disse siste arbeidsoperasjonene kunne også vært gjort automatisk, og i dag er det lett å se at en ville ha vunnet mye ved å overlate det til maskinen. Men vi har i alle fall blitt spart for mye enerverende og tidkrevende arbeid.

Den framgangsmåten som jeg har skissert her, er også blitt brukt i et tilsvarende prosjekt, som ligger til grunn for Geirr Wiggens hovedfagsavhandling, og i et delprosjekt om tilhøvet mellom syntaksen i formell og uformell kontekst, ved Olaug Rekdal.[3]

Informasjonssøking er som kjent et felt hvor datamaskinen er overlegen i forhold til alle andre muligheter, dette gjelder også språklig tekstbehandling. Men det er likevel ei begrensning i at maskinen må operere på formelle strukturer, i og med at det ofte er forhold som helt eller delvis har med språkets underliggende strukturer som er av interesse for språkforskeren. Trass i dette forholdet kan en ha mye nytte av datateknikk i språklig tekstgransking. I denne sammenheng vil jeg trekke fram noen eksempler fra TAUS.

Et av de større delprosjekta er en analyse av personlige pronomen, ut fra syntaktiske, morfologiske (bøynings-) og semantiske aspekter.[4] Ettersom det ikke fins noe DB-program som kan hente fram personlige pronomen fra en tekst, når det vel å merke ikke er lagt inn noen markering på de enkelte orda, må vi gå fram på en annen måte. Vi må ta utgangspunkt i de aktuelle ordformene og bruke et kontekstprogram som kan gi utskrift av forekomstene. På forhånd veit vi at det fins fleire ulike former av hver kategori (ulike transkripsjoner)



De fleste termene skulle være sjølforklarende. Ansats vil si begynnelse til nytt makrosyntagme, mens syntaktisk mix (ellers kalt anakoluti) vil si ammenblanding av to syntaktiske konstruksjonstyper

på grunn av forskjellig uttale: f.eks. JÆI, JÆ, JÆI* osv. En går til fullstendige ordlister (for hver tekst eller for prosjektet som helhet) og finner de formene som en trur kan være personlige pronomen, og får utlistet kontekster på dem. På denne måten får en ei nokså pålitelig registrering av forekomster, og samtidig kan en få ei tilfredsstillende ordning av materialet: forekomster av 1. person entall for seg, de andre kategoriene for seg, og ikke hulter til bulter slik de opptrer i en tekst. I den aktuelle undersøkinga er det snakk om ca. 15 000 forekomster av personlige pronomen, og det er trulig søkt på ca. 20 000 kontekster, av en tekstmasse på ca. 293 000 maskinord.[5]
Det er ingen tvil om at denne undersøkinga ville vært praktisk umulig uten EDB. Dvs. den kunne nok ha vært gjennomført, men ikke uten en svær innsats av menneskelig arbeidskraft for nokså trivielle arbeidsoppgaver. I TAUS er det utført fleire delprosjekt med noenlunde samme framgangsmåte som i den omtalte undersøkinga. Her kan nevnes et større arbeid om a-endinger i substantivkategorier (f.eks. gaten/gata, bilene/bila), et arbeid om l-fonemet i Oslomålet og fleire hovedfagsavhandlinger som er under arbeid.[6]

I alle de arbeida som er nevnt er det brukt en kombinasjon av datamaskinell og mer alminnelige, manuelle prosedyrer. Dette må en nok si har vært fornuftig, ofte er det vel så hensiktsmessig å la datamaskinen gjøre »grovarbeid» og så gjøre en del manuelt arbeid, framfor å arbeide mye med å finne høgt utviklete, raffinerte, automatiske prosedyrer. Jeg er likevel ikke i tvil om at det hadde vært det beste å la EDB-teknikken arbeide med fleire av de kvantitative og statistiske oppgavene som jeg har vært inne på, f.eks. i samband med pronomenanalysen. Men det er ofte slik at den tid og det merarbeid som det vil ta å arbeide fram mot en mer avansert rutine, må veies mot en enklere men mer tidkrevende framgangsmåte.

Ut fra en samlet vurdering vil jeg nok si at bruken av EDB har betydd mange fordeler for et prosjekt som TAUS, som tids- og vel også arbeidssparende faktor. I hvert fall har det fritatt medarbeiderne for mange trivielle og tidkrevende arbeidsoppgaver. Jeg tror jeg vil driste meg til den påstand at bruk av EDB er hensiktsmessig i praktisk talt alle språk-forskningsprosjekt som er basert på tekstmateriale, og vel å merke uten fare for den »teknifisering» som enkelte er redd for. Jeg bør vel også føye til at med det nye EDB-anlegget på Blindern er mulighetene langt bedre enn de var mens TAUS pågikk; ut fra dagens forhold ville vi fått en betydelig innsparing av tid og utgifter.

Etter denne lille presentasjonen er det kanskje noen av leserne som lurer på om det er kommet noen språksosiologiske resultat ut av arbeidet. For å besvare det svært kort: jo, det har det. Men det er ikke mulig å generalisere fra det lille utvalget av personer, og vi må si at det egentlig bare er nokså små og ubetydelige tendenser til sosiale »gruppespråk». Derimot er det svært store individuelle skilnader i frekvens og bruksmønster for mange syntaktiske og morfologiske trekk. Dette gjelder de språktrekk som er undersøkt, så vi må foreløpig kunne gå ut fra at de viktige språksosiale skilnader må ligge på andre felter i språkbruken.

Noter

1 Opplysninger om dokumentasjon kan fåes ved henvendelse til EDB-konsulent Ivar Fonnes, Universitetet i Oslo.

2 En mer fullstendig redegjørelse for opplegget av prosjektet og av databehandlinga av materialet fins i Talemålsundersøkelsen Hovedrapport, som skal utgis med det første.

3 Jfr. Geirr Wiggen: Sosio-syntaktisk undersøking av talemålet til utvalgte grupper Oslo-ungdom. 1974. Stensil. En revidert utgave av arbeidet vil seinere bli utgitt. Arbeidet til Olaug Rekdal er foreløpig ikke ferdig.

4 Resultatet av denne delundersøkinga blir lagt fram i Eskil Hanssen: Personlige pronomen i kontekst, som vil bli utgitt seinere.

5 Jfr. Eskil Hanssen: »Ordforrådet i naturlig talespråk» i Norskrift 1, 1975.

6 Jfr. Knut Western: a-endinger i substantivkategorier i Oslomål. (I trykk). Ernst Håkon Jahr: »l-fonemet i Oslo bymål.» i Norskrift nr. 1, 1975.

melding om senterets virksomhet

VÅREN OG HØSTEN 1976

Arbeidet i senteret ble i 1976 ført videre i overensstemmelse med den langtidsplanen som Rådet for humanistisk forskning har vedtatt for perioden 1974-1977. Viktige trekk ved virksomheten i fjor vil bli omtalt nedenfor.

1. Senterets fremtidige status:

I løpet av våren 1976 ble det utført et omfattende planleggingsarbeid i senteret som et forarbeid for den prinsippdiskusjon om senterets fremtidige status som Rådet for humanistisk forskning skulle ha ved slutten av vårsemesteret. Styret for NAVF's EDB-senter utarbeidet bl.a. et perspektivnotat om EDB i humaniora. Bl.a. med grunnlag i dette dokument behandlet Rådet i mai 1976 NAVF's videre engasjement på EDB-feltet, og vedtok at NAVF fortsatt skulle påta seg å opprettholde en nasjonal EDB-tjeneste. Det ble dessuten bestemt at denne nasjonale EDB-tjeneste skulle beholde noenlunde det samme organisasjonsmønster som NAVF's EDB-senter for humanistisk forskning har i dag.

På ettersommeren ble det ført forhandlinger med Universitetet i Bergen om en samarbeidsavtale for EDB-senteret etter 31.12 1977. Forhandlingene resulterte i at gjeldende samarbeidsavtale ble forlenget fram til 31.12.1980. Avtalen gir EDB-senteret om lag de samme arbeidsvilkår ved Universitetet i Bergen i den nye avtaleperioden som i den forrige.

Med bakgrunn i denne avtalen og vedtakene i NAVF planlegger senteret nå virksomheten for en ny 3-års periode.

Det er ventet at NAVF i løpet av denne perioden vil få overført til universitetene det økonomiske ansvar for de lokale konsulenttjenester. Det er grunn til å tro at den første universitetsansatte EDB-konsulent for de humanistiske fag kan starte sitt sitt arbeid fra sommeren av - ved Universitetet i Trondheim.

Når de lokale servicefunksjoner er overtatt av universitetene, kan NAVF's EDB-senter i sterkere grad konsentrere seg om de nasjonale fellesoppgaver.

2. Konsulentassistanse og prosjektsamarbeid.

NAVF's EDB-senter har i løpet av 1976 fortsatt sin konsulentassistanse til enkeltpersoner og gruppeprosjekter som ønsker å nytte EDB. Senterets konsulenter står i dag i et samarbeidsforhold til svært mange av de EDB-tiltak som drives innenfor de humanistiske fag. Noen ganger er det bare tale om innledende veiledning ved starten av et prosjekt eller instruksjon i bruk av standard programutrustning. Andre ganger kan det være tale om oppfølgende konsulentassistanse gjennom hele prosjekttiden. Senteret har også ansvar for konsulentassistanse til institusjoner utenfor universitetene.

En viktig del av senterets servicetiltak ytes i form av puncheassistanse til nye EDB-brukere. Puncheservicen blir utført i Bergen og i 1976 ble det tatt hånd om oppgaver fra alle universitetsbyene. Det vises for øvrig til kunngjøring om konsulent- og puncheassistanse i dette nummer.

Nedenfor omtales noen av senterets samarbeidsprosjekter:

Prøveprosjektet »EDB og manuskriptregistraturer» er et samarbeidstiltak mellom NAVF og Riksbibliotektjenesten. Prøveprosjektet har som mål å vise hvordan databehandling kan tas i bruk ved etablering av en samkatalog for håndskriftmateriale. En del av prosjektet har gått ut på å undersøke fordeler og ulemper ved å punche håndskriftkataloger og håndskriftregistre slik de

er uten forutgående tilrettelegging. Målet har vært at også forsøksarbeidet skulle gi et katalogprodukt av varig verdi. Det vil derfor som resultat av prøveprosjektet bli etablert et fullstendig EDB-register over privatbrev i Riksarkivet. Prosjektet som nå er om lag ferdig, vil bli nærmere omtalt i et senere nummer av Humanistiske Data.

NAVF's EDB-senter har i de siste par årene stått for ledelsen av EDB-arbeidet i forbindelse med tilrettelegging av Norsk Landbruksordbok for trykking (jfr. Humanistiske Data nr. 1-2 1975). EDB-arbeidet har i 1976 vært noe komplisert på grunn av skifte av maskinanlegg ved Universitetet i Oslo. Høsten 1976 er databehandlingen blitt utført på Studentsamskipnadens anlegg.

Ved utgangen av 1976 var det meste av materialet lagt til rette i maskinleselig form. Materialet, som inneholder ca. 18.000 ordartikler med synonymer for oppslagsordene på 7 språk, vil foreligge på trykkeklart magnetbånd (drivetape) i løpet av sommeren 1977. En første trykkprøve ble kjørt ut høsten 1976.

I 1976 har senteret deltatt i et prosjektsamarbeid om automatisk syntaktisk analyse Arbeidet bygger på et programutkast utarbeidet av dr. Martin Kay, USA. Med utgangspunkt i et fragment av en formalisert grammatikk for moderne norsk, utarbeidet av Svein Lie, Universitetet i Oslo (jfr. Humanistiske Data nr. 1-2 1975), har senteret laget en programversjon av grammatikken som virker på enkle norske setninger. For tiden arbeides det med å utvide grammatikkens analysekapasitet og utvikling av program for automatisk ordklassebestemmelse. Analysesystemet vil i løpet av vinteren 1977 bli implementert på DEC 10 i Oslo.

I samarbeid med Nordisk institutt, Universitetet i Trondheim er det satt i gang en forstudie over moderne, norske leseverk. Målet er ved hjelp av datamaskinell tekstbehandling å gi en analyse av språk og innhold i tekstene i leseverket. Datagrunnlaget for prøvearbeidet er Gyldendals leseverk for 7. skoleår.

3. Utdannings- og informasjonstiltak.

Senteret har i 1976 videreført ordningen med ordinære brukerkurs innenfor ulike deler av humanistisk databehandling. Det er blitt holdt kurs i Oslo, Trondheim, Bergen og Tromsø. Dessuten er det gitt individuell instruksjon i forbindelse med bruken av nytt utstyr.

Gjennom informasjonsmøter og oppsøkende konsulentvirksomhet blir senteret kjent med nye, potensielle brukermiljøer. I Oslo har det vært en viktig oppgave å gjøre brukerne fortrolig med det nye data-anlegget, som kom i drift høsten 1976.

Det er holdt spesialseminar i Bergen om bruk av datamaskin i arkeologisk forskning, og om typografiske teknikker og deres betydning for datamaskinell tekstbehandling.

I 1976 ble en ordning med korttidstipend for humanister introdusert. Stipendordningen, som er nærmere beskrevet i en egen melding, bygger på et individualisert 4-ukers studie- og instruksjonsprogram. Stipendiatene kom i 1976 fra Oslo, Trondheim og Tromsø.

Senteret arrangerte i november den første norske konferanse om humanistisk databehandling med representanter fra alle universitetene (se egen melding). Det kom på konferansen klart til utrykk et behov for regelmessig avholdte nasjonale sammenkomster. Deltakerne uttrykte ønske om en veksling mellom mindre, temaorienterte samlinger og bredere anlagte tverrfaglige konferanser.

4. Programutvikling.

En god del av programutviklingsarbeidet i 1976 har bestått i å utbygge standard programtilbud for humanister. I Oslo har det vært lagt ned arbeid i å overføre programpakker til bruk på det nye data-anlegget. I Bergen og Trondheim har det vært en

videre utbygging av tekstbehandlingspakkene PPTT og KVIKKIS. Det vises til egne meldinger om programpakkene.

I Bergen har det vært et samarbeid med Universitetet i Tromsø og Statens Rasjonaliseringsdirektorat om en implementering av tekstsøkesystemet NOVA-STATUS. Senteret vil i januar 1977 implementere NOVA-STATUS på universitetsanlegget ved Universitetet i Trondheim. Dette tekstsøkesystemet, som primært er utviklet for behandling av dokumenter, kan også brukes til søking i fast formaterte data (jfr. seperat omtale av systemet).

Høsten 1976 avsluttet senteret det lingvistiske forarbeid for et datamaskinelt lemmatiseringsprosjekt. Målet er å utarbeide et programsystem for automatisk sammenføring av ordformer under de respektive oppslagsord (»huset», »husene», »husenes» under oppslagsordet »hus»). Metoden går ut på at datamaskinen gir forslag til lemmatisering av alle ord, og at brukeren via terminal foretar de nødvendige korreksjoner. Første versjon av lemmatiseringsopplegget vil være klar i løpet av 1. halvår 1977.

FIFTH INTERNATIONAL SYMPOSIUM ON THE USE OF COMPUTERS IN LINGUISTIC AND LITERARY RESEARCH

(SPONSORED BY THE ALLC)

3-7 APRIL 1978
UNIVERSITY OF ASTON
IN BIRMINGHAM

Themes at the symposium will be:
authorship studies
concordances
classical studies
input/output
oriental studies
software
stylistic analysis
syntactic analysis
text editing
language-oriented groups
education
lexicography
literary statistics

Correspondence address:
Professor D. E. Ager
Department of Modern Languages
University of Aston in Birmingham
Gosta Green
Birmingham B4 7ET
England

PISA

Den 4. internasjonale sommerskole i datamaskinell og matematisk lingvistikk blir holdt i august/september i Pisa, Italia. Det endelige programmet er ennå ikke fastsatt. Under forberedelsene har en vurdert to ulike opplegg:

a) »Natural Language Understanding by Computer»

b) »Text processing in the humanities».

De som er interessert i å delta på sommerskolen bør ta kontakt med professor A. Zampolli

Director of the International Summer School
C N U C E - Via S. Maria, 36
56100 PISA (Italy) Tel. (050) 45245
Telex 50371 - CNUCE

Dieter Wickmann:

Statistics in the Humanities. Some Epistemological Remarks

Dr. Dieter Wickmann, Institut für Matematisch Empirische Systemforschung (MESY), Aachen holdt i november 1975, på invitasjon av NAVF's EDB-senter, tre forelesninger i Bergen. I en lett forkortet utgave presenterer vi her ett av foredragene. Det andre foredraget er trykket i ALLC bulletin 1976 Vol. 4 Nr. 1 »On Disputed Authorship, Statistically».

The key word of the activities in the Humanities is Hermeneutics; Hermeneutics is the art of interpretation. Statistics, on the other hand, as a special branch of mathematics, belong to the Sciences. The application of statistical methods on problems in the Humanities may be meaningful and possible in some special cases; so far, you will certainly agree. But perhaps you will reply that the essential problems in the Humanities are inaccessible to mathematics and statistics in particular. I wouldn't deny that, but I cannot consent entirely, either. Most of you are working in the field of the Humanities; so, I may suppose that you are less familiar with the other side. I mean the activities in the Sciences. I shall, therefore, concentrate most of my talk on the methodology inherent in scientific activity.

I start by quoting K. R. Popper; he writes in his book »Logic of Scientific Discovery»: »Science does not rest upon rock-bottom. The bold structure of its theories rises, as it were, above a svamp. It is like a building erected on piles. The piles are driven down from above into the svamp, but not down to any natural or 'given' base; and when we cease our attempts to drive our piles into a deeper layer, it is not because we have reached firm ground. We simply stop when we are satisfied that they are firm enough to carry the structure, at least for the time being».

The comparision of scientific activity with a pile-construction above a swamp is an excellent characterization of the general situation.

One is essentially not concerned with exact matters about which, in the end, there is no difference of opinion in as much as only logic is acknowledged. The fact that nowadays the benefits of the Sciences are completely integrated in our society, veils its true methodological traits which may be circumscribed by the question: Since we cannot know what the world actually is, according to which criterial have we to behave in the face of our objects? A review of the history of the Sciences teaches us that, till our days, theories and pictures about the world replace one another. No theory is true in the logical sense of the word; all you can say is that it has stood several tests. Numerous examples in the fields of physics, chemistry, biology, sociology, psychology bear witness to the permanent replacing of theories by new ones. We are not dealing with unassailable truth, but with hypotheses about the world, each having its own historical development and

lifetime. The American historian of Science, Thomas S. Muhn, calls it a permanent change of paradigmata.

Now, let us have a closer look at the logic of these changing theories. Theories are hypotheses; they are assertions valid as long as no serious arguments contradict them. What logical form do such hypotheses have? They are universal statements, and universal statements have the form: »All elements of the class C have the property P». E.g. the class C1 of celestial bodies: »All elements of the class C1 have the property P1 of moving according to Newton's law». Or the class C2 of ravens: »All elements of the class C2 have the property P2 of being black». This is in short: »All ravens are black». Universal statements do not have indications about space and time; they claim to be valid always and everywhere. Universal statements are logically equivalent to the negation of universal existential statements. »All ravens are black» is equivalent to »There is no non-black raven». And the first example may be transformed in »There is no celestial body not moving according to Newton's law». All of you know the fundamental law of thermodynamics: »There is no perpetual motion machine». Here comes out that universal statements might be compared to prohibitions: they don't assert the existence of something, but prohibit the existence of something.

We have to distinguish between universal statements and singular statements, in particular singular existential statements, which do have indications about space and time. Let us put together these two indications to a single one, the so-called space-time-region k. Then, the singular existential statements have the form: »There is a so-and-so in the space-time-region k» or »Such-and-such an event is occuring in the region k». E.g.: »At 10h a black raven was sitting on the roof of this building».

The relation between universal statements and singular existential statements is of great importance for the progress of knowledge. We claim: Singular statements, regardless of how, are not able to prove or to verify a universal statement claiming to be valid for an infinite class of elements and, sticking to our example, who will warrant the next raven, after a thousand observations, to be also black?

On the other hand, a single singular statement as »There is a white raven in the space-time-region k» disproves or falsifies the universal statement »All ravens are black». This asymmetry of verification and falsification is typical for the well-known mode of logical conclusion called modus tollens. I. Kant says: »The modus tollens not only proves rigorously, but also very easily. For, if there is only one false conclusion drawn from a statement (a universal statement that is) then the statement is false». Thus, the singular existential statements are of great importance for the falsification of theories, i.e. hypotheses.

But in reality we are not only dealing with logical relations between statements, but also with their contents, and here the big problems start.

Would you really refuse the universal statement »All ravens are black» when you are told about a white raven now sitting on this roof? Would you do it when you perceived a white raven yourself? Probably not. You wouldn't trust your own perception. Perhaps you would rub your eyes and have another look at the phenomenon. You would walk around it because you might think of a light reflexion, and so on. In any case, you would set up s e v e r a l of those singular statements, for instance: »At 10.30 and at 10.31 and at 10.33 I perceived a white raven on the so-and-so place». How many of those observations would

suffice to convince you that the universal statement was wrong? Or: A planet is never observed exactly in the position deduced from a theory. Would an astronomer reject Newton's law after having observed one single deviation? The answer again is in the negative. What must the observations be like to falsify the Ptolemaic hypothesis of the circular orbits? You will find numerous further examples in all fields of scientific activities. Now, you might reply: Regardless of how you argue, it is the case here in opposition to the Humanities, that the objects of these activities are observable ones. But this is, for the most part, not the case. Of course, sensual perception is of a certain importance, but not essentially more so than in the Humanities. There too, you have to observe a sequence of signs in a text, a sequence of sounds in a speech, gestures, behaviour, attitudes. Observability is not the criterion separating the Sciences from the Humanities. In both, the proportion of sensual perception to their intellectual implications may be compared with the visible part of an iceberg to the rest of it.

We have, however, to postulate that the contents of singular existential statements can be examined; in other words that, at least in principle, intersubjective agreement is attainable. This is one of the most important piles of our pile-construction. Singular existential statements whose content can be examined are called (due to Popper) basic statements. They are so called because they form the base in the specific field an investigator is working in — the base which, as we hope, will carry the construction, at least for the time being.

Statements about the world are always statements about our experience of the world. Thus, basic statements cannot be »true« or »false«; they are accepted or not accepted. Anybody who hesitates to accept a basic statement should have the chance to examine its content for as long and as rigorously as he wants to. If this is not possible, for whatever reason, the investigation is unsuccessful. Here, the communicative foundation of scientific activity comes to light. The postulate of intersubjective testability essentially determines the historical development of the Sciences. Individual experience which cannot be shared by other persons does not pertain to scientific discovery. The postulate of intersubjective testability involves another postulate, the postulate of reproducibility of the phenomena concerned. We shall come back to this point later on.

First, there are obviously different kinds of basic statements. Some you may accept more easily than others. Why would you trust more in your own perception of a white raven than in that of somebody else telling you about his detection? Probably you would trust more in such news reported in a scientific paper than in an issue of the public press. Why do we accept a basic statement about a recent astronomical event more easily than one come down from Ptolemé? Although we cannot go now into a detailed analysis, we may put it as follows: The confidence we have in a basic statement depends on theoretical implications involved in the statement itself. It was mentioned already that theories cannot be proved, but they are more or less corroborated. Popper has shown an approach to measuring the corroboration of a theory; I refer to his book »The Logic of Scientific Discovery«. The theoretical implications I mean are universal statements about the communication chain or the information channel between the observer's brain and his object. There are several possible channels between the observer's brain and, say, a white raven; the direct visual one is the most sure one and this

is in itself a theory, a hypothesis, formulated in a universal statement: »All visual perception is an exact copy of the environment». If, therefore, Mr. N. reports his observation of a white raven and you are inclined to accept it, the following universal statements might be involved: »Mr. N. is an ornithologist; he is able to distinguish ravens from other birds at any time and in any place» etc.

The valuation of basic statements depends on the degree of corroboration of the theories involved. Basic statements are interfused with theories. I quote Popper again: »Observations, and even more so observation statements, are always interpretations in the light of theories». Thus, the base begins to waver. Now, what about theories? How do t h e y come to life? How did Keppler happen to think of an elliptic motion of the planets? What leads us to claim that the sun rises every morning? (This is, indeed, a theory; you may put it into the form of a universal statement). How do we find, in general, the innumerable theories guiding our daily life, more or less unconsciously? A sudden inspiration, a fancy, a new idea, a change of a paradigma, are somewhat spontaneous, speculative. Although those creative processes cannot be systematized, one might say that they are not independent of certain observations or, more precisely, not independent of the content of certain accepted basic statements. That is to say: An inventor of a new idea is guided by a set of basic statements pertaining to the topic he is dealing with. Keppler had to have a substantial amount of information in order to be able to find out the elliptic orbits.

Now, we have seen on the one hand that basic statements are interfused with theories, on the other that theories are established on the base of basic statements. What, in the end, is the basis of scientific discovery — theories, or basic statements being interpretations in the light of theories? They condition each other and, obviously, we are landed in a circle; it is exactly that circle which in the Humanities is called the hermeneutic circle. According to Heidegger, the hermeneutic circle consists in the fact that the understanding individual, by his own experience, must have knowledge about the object of his understanding. But, as you may know, the hermeneutic circle is really not a point of discussion in the Sciences; it is really not a problem there. I would say: it depends. For, the hermeneutic circle is realized to a variable extent. In its most simple, trivial form it is identical with the vicious circle of logigs in which the only reason for an explicans is its explicandum. The explicans being a theory and a theory being a hypothesis, we call it an ad-hoc-hypothesis. This was only established in order to explain the present data which, in turn, were the basis for the hypothesis; it is a pseudo-explication, of course.

Keeping this in mind, it is obvious what we have to do to escape the circle: We have to deduce from the theory in question testable singular statements which have not been used for the establishment of the theory. We are only allowed to speak of a corroborated theory if there exists at least one such statement. Now, if there is only one, we are still close to the circle. But with an increasing number of such statements the situation gets better and better. I would like to quote here the American scholar of Literature, E. D. Hirsch, dealing with this point too, in his book »Validity in Interpretation»: »A useful analogy to the self-verifying ability of interpretations is to decipher an unknown code. The mykenic linear B, for example, was deciphered by Ventris, but Ventris' solution was not generally accepted. Some scholars argued

that such a decoding has the property of verifying itself, because the decoded elements were employed just in order to establish the system. The text infallibly proves the theory, because there was nothing in it that was not born of the theory. Ventris convincingly could refute this objection only when further texts were deciphered not yet considered for the set-up of this system». We shall call singular statements deduced from a theory »prognostic statements»; »pronostic» refers to the future knowledge of the examinator, not to further events. To this effect, a past event deduced from a theory is also a prognosis. We say that the accordance of a prognostic statement with its corresponding basic statement corroborates the theory.

Summing up the first part of my talk: (1) The problem of the mutual conditioning of theories and their objects exists equally in the Humanities and in the Sciences. (2) The hermeneutic circle is realized to a various extent. One escapes it better the more prognostic statements are testable. (For the sake of clarity I have to add that the progress in knowledge not only depends on the number of testable prognostic statements, but also on certain properties of the theories themselves. In view of the limited time, we cannot enter into a discussion about this point now.)

Here the Sciences seem to differ from the Humanities: The objects dealt with in the Sciences, more often than not, may be reproduced as many times as you like; the set of basic statements for comparison with prognostic statements is, in general, unlimited. Therefore, the theories in the Sciences are well corroborated. But this is not always the case, this is not a t y p i c a l trait of the Sciences, which I want to explain now.

We are confronted with the following situation:

Regardless of how brilliant the intuition was that formed a hypothesis, the hypothesis must be tested. This actual test will be based, in any case, on only a limited set of data or information. After the investigation, we are faced with one of the following four possibilities:

(1) The hypothesis is tenable and, after the investigation, it is accepted (right decision)
(2) The hypothesis is tenable and, after the investigation, it is refused (false decision)
(3) The hypothesis is not tenable and, after the investigation, it is accepted (false decision)
(4) The hypothesis is not tenable and, after the investigation, it is refused (right decision).

In view of the fact that human knowledge is principally limited and incomplete, these four outcomes of an investigation are the starting point for all decision-analyses. Now the challenge is to avoid, as much as possible, the decisions of lines (2) and (3). And here statistics comes into play. The challange mentioned is the general impact of statistical test-theory. Robberts and Wallis put it like this: »Statistics is a body for making wise decisions in the face of uncertainty». Certainly is unreachable, thus any hypothesis is a potential candidate for a statistical analysis. Although we don't apply statistical methods on every occasion, we do behave,more or less unconsciously, statistically. Statistics, in the proper sense, as a body of methods, satisfies beyond that the postulate of intersubjective testability and the postulate of eliminating wrong hypotheses as soon as possible.

The word statistics covers several meanings. Besides the simple meaning of a listing, e.g. a statistic about natality or an income distribution, there are two main meanings: testing and estimation. Estimation concerns establishing a theory, which is not our point now. Testing concerns examining a theory.

The main part of what a student in statistics has to learn is to become familiar with the methods involved and, beyond that, to acquire the ability of developing new methods for new problems. The beginning of systematic statistics goes back to the 17th century; but the modern mighty and efficient discipline of mathematical statistics has only been developed during the last 60 years. As it is in general mathematics, so in statistics: the essential objects are not numbers but structures and relations. Models are set up upon which to map the reality; the relations between the entities and variables are studied in the model and, finally, the results are reinterpreted into reality. Chance, randomness, probability are basic concepts in statistics. Randomness is nothing mysterious. An event is called a random event if the outcome of a trial or an experiment is n o t certain. Randomness is closely related to incomplete knowledge, which does not mean the absence of relations between random events; it is just a main statistical business to discover these relations.

If you were to ask me to what extent statistical methods are applicable in the Humanities, I couldn't give you a precise answer. In physics for example, very often special mathematical methods had to be developed to fit the actual needs. The same is likely to be the case for the Humanities. Take, for instance, statistical linguistics. In general, you have to manage an enormous amount of data in order to obtain relevant and useful results. In my opinion this is one of the reasons why this discipline has come into existence only recently, because, for the extensive data processing, large computers are necessary and these are about 10 years old. Note, by the way, that a research worker in the field of language has a lifelong experience about his subject, whereas a computer at the beginning of your job knows as much about that as a new-born baby.

It is more instructive to talk about the logic of statistical inference. For this end, let us have another look at the four possibilities after a decision is made. First, the hypothesis H to be tested is supposed to be true. We call it the null hypothesis. Then, the probability is calculated of the occurrence of such a configuration of data as the one actually observed. For this purpose all the data are combined, according to the given problem, into a single random variable, a so-called statistic. In general, this is the most difficult step. Here I ought to say something about the relation between quality, quantity and measureability, but this is a subject on its own. If the probability of the occurrence of the statistic is sufficiently great, we will conclude that the data do not contradict the hypothesis H; if it is small we are, beyond a certain point, no longer willing to accept the hypothesis H; we will reject it. If, at that time, an alternative hypothesis A is already at hand, we will decide in favour of A; if not, we have to seek for an alternative hypothesis. This is what Keppler did after having discovered that the circles or epicycles did not fit the planetary orbits or, in our terms, after having rejected the Ptolemaic hypothesis. He still needed 9 more years to establish his well-known Keppler laws. The mode of statistic inference is to be compared with the reductio ad absurdum in mathematics; we may call it reductio ad improbabilitatem. In mathematics or logics an hypothesis is rejected if some inference drawn from it is false; in statistics that means dealing with a r e a l deciseon problem - we re ject an hypothesis if the occurrence of some inference drawn from it is improbable. Thus, it comes out that the mode of statistical inference is that of the modus tollens, though in a somewhat reduced strictness. Instead of the s i n g l e statement contradicting and falsifying a universal statement in pure logics,

we have now a certain set of basic statements contradicting their corresponding prognostic statements to a certain, but well-defined, extent, inc[illegible]ng us to reject the hypothesis. This very set of basic statements is called, in statistical terms, a random sample.

We may summarize the situation as follows: By intuition or creativity, theories about the world have to be invented - they are interpretations by way of trial. The second step is to examine the hypothetical theories, that is to check their validity.

After the examination we either find that the data do not contradict the hypothesis and we accept it for the time being, or, if the data contradict the hypothesis in the mentioned sense, we will reject it by virtue of the modus tollens. Therefore, progress in knowledge may be seen in rejecting or refuting invalid hypotheses, which we may put also in this way: progress in knowledge consist in restriction of choice.

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

Now, which scientific disciplines do, in fact, apply statistical methods, that is to say, statistical test methods? Physics and chemistry, for instance, make very little use of it. These are research fields dealing with phenomena reproducible at any time.Although the dilemma mentioned is a fundamental one, in these fields many basic statements are at disposal to enable even the most scrupulous examinator to decide in favour of this hypothesis or the other. A good deal of scientific activity, however, does not dispose of such an exuberance of relevant information. The acquisition of information may be too expensive, may take too much time, may be too perilous as in medical research for instance, or is completely exhausted as in histotical investigations. In short: In all those scientific fields in which only restricted information is at hand, the application of statistical methods is meaningful. In other words: The interrelation of universal statements and basic statements, loosely speaking the interrelation of theory and experience, is embedded between two extreme situations - the vicious circle (the only reason for the explicans being the explicandum) on the one hand: on the other we have the possibility of procuring as much information as we want to. For both, the application of statistics is inadequate - in the first case in a very trivial sense, in the second because here is no critical decision problem. In between, in the range of limited information, is the potential domain of statistical test theory and its application.

Ladies and gentlemen, the aim of my talk is, essentially, to give a contribution to the unity of scientific activities, in a large sense of the word. We must not seek for it on the level of the objects of the various investigations, but on the level of methods by means of which the objects are treated. If we feel bound by some postulates, the most important being the postulate of inter-subjective testability, we have to follow the logic of scientific discovery,this being a methodology for a systematized examination of universal statements. It makes good sense to call statistics a meta-theory.

NORSK KULTURRÅDS *Fotoprosjekt*

Databehandling av opplysninger til et fotografisk materiale

Norsk Kulturråd nedsatte i juni 1972 et utvalg for registrering og bevaring av gamle fotografier (Fotoutvalget) med fotografmester Jacob Brun som formann.

Utvalgets mandat var

1. Utarbeide forslag til hvordan registrering og arkivering av gamle fotografier og platesamlinger kan organiseres lokalt gjennom faglige, regionale, fylkeskommunale og kommunale instanser.
2. Utarbeide forslag til standardisert registreringsmateriell for hele landet.
3. Utarbeide forslag til en felles systematikk for arkivering, foreslå hvordan materialet bør arkiveres og å undersøke de tekniske betingelser for arkivering av gamle fotografier.

Som ledd i sitt arbeid vedtok utvalget å få utført et prøveprosjekt med datamaskinell behandling av opplysninger til et utvalg av fotografier. I den forbindelse ble NAVF's EDB-senter kontaktet (jfr. Humanistiske Data nr. 2, 1974).

I juni 1974 ble det bestemt at senteret skulle inngå samarbeid med Kulturrådet om prøveprosjektet. I første rekke skulle senteret yte konsulenthjelp i alle faser av arbeidet, og dessuten etablere en database for prøvemateriale og stille til disposisjon program for søking i materialet. Senterets oppgaver i forbindelse med prøveprosjektet ble avsluttet i desember 1975.

Fotoutvalgets medlemmer valgte for prøveformål ut ca. 2000 bilder fra ulike typer arkiv og publikasjoner. Til bruk ved tilretteleggingen av materialet for punching, ble det ved senteret utarbeidet et eget registreringsskjema og en utfyllingsinstruks. Registreringsskjemaet inneholdt felt for:

1. Registreringsnummer.
2. Opplysninger om fotografen.
3. Opplysninger om fotografiet.
4. Motivopplysninger (fritekst).
5. Stikkordskildring av motivet (frie stikkord).
6. Faste tilleggsopplysninger av motivet (faste stikkord).
7. Oppbevaringssted for fotografiet.
8. Registreringsopplysninger.

Utfylling av registreringsskjema ble i hovedsak utført av en fast registrator.

Punching og retting ble utført ved senteret. Korrektur- og retteprogram, sorteringsprogram og tekstsøkeprogram (COBOL) ble utviklet.

For en bruker av bildematerialet vil det være av betydning at det er lett å finne fram til de bilder vedkommende har interesse av. Selv etter en vidtgående seleksjon av bevaringsverdige bilder vil det være behov for å ta hånd om og søke opplysninger om mange hundre tusen bilder i et sentralregister. Et EDB-opplegg for et slikt sentralregister vil derfor kreve store planleggings- og etableringskostnader. For de begrensede mål en hadde med prøveprosjektet valgte en å søke fram bilder gjennom ulike typer listeprodukter:

1. Hele arkivet ble skrevet ut fullstendig i en liste ordnet etter registreringsnummer.
2. Et sorteringsprogram gjorde det mulig å sortere materialet etter ett eller flere underfelt i registreringsskjemaet for et bilde og å få skrevet ut en sortert liste

Eksempel på utskrift av konkordans

```
  251          Olo Bull står i midten med hatten i hånden, hans amerikanskfødte fr. Sara, sitter ytterst til venstre; Stabbur fra
 1203    med en reise-kiste i hånden. Trolig ser vi startfasen på Amerikareise. En kone sitter på dørhellene. Gjennom den åpne døren
 1083       Kaptein Smith, sorenskriver Erichsen, kjøpmann Holst, amtmann Collett og jernbanedirektør Pihl.
   44          atelier. Laurentius Urdahl, Wilhelm Holst og Roald Amundsen. 1872-1928. Polarforsker. Kvinne i hel figur, kledd i
   14          med Simon 'rift i bavg''. Laurentius Urdahl, Roald Amundsen, Wilhelm Holst på skf  i atelier. De har ryggsekk med
 1471                  1910. Røker langpipe. Portrett av Roald Amundsen malt av Otto Hjort Portrett av kansellfråd Peter Collett
```

Eksempel på utskrift av alle registreringsopplysninger

```
      302  ****          ****       302  ****          ****       302  ****          ****       302  ****       302

AR-STED          1909                      Oslo                      EKSTERIØR  /PORTRETT/TOPOGRAFI/GJENSTAND
FOTOGRAF         Wilse, Anders Beer        Oslo                      1865       1949       YRKE

FRITEKST         Utenriksminister Wilhelm Christophersen (1832-1913) med kone og datter i slede utenfor ministerboligen i Parkveien
                 45. Sleden er forspent med to nester og de har et hvitt heketeppe over seg.

FASTE-OPPLYSN    N=Christophersen, Wilhelm (1832-1913)! Y=Utenriksminister; S=Parkveien 45; G=Slede; G=Hest; G=Hestedekke; A=Vinter;

STIKKORD         samferdsel, slede

DIV. OPPLYSN     FOTOGRAFI                 ANNET                     S-H        1   13x18      COPYRIGHT
INSTITUSJON      Norsk familiealbum        Grøndahl og Søns forlag, Oslo s.130
EIER             Grøndahl og Søns forlag   Oslo
REG-OPPLYSN      Sofie Rogstad             19750529
INFORMATOR       Norsk familiealbum        Grøndahl og Søns forlag, Oslo
```

Eksempel på utskrift av faste stikkord

```
Y=Utenriksminister          302  Oslo                    1909      PTG
                                                                   G=Hest
                                                                   G=Hestedekke
                                                                   G=Slede
                                                                   N=Christophersen, Wilhelm (183
                                                                   S=Parkveien 45
                                                                   Y=Utenriksminister
                                                                   A=Vinter
```

som inneholdt oppgitte felt. Det ble bl. a. tatt ut lister sortert etter fotograf, år og sted samt at registreringsnummer ble skrevet ut på listen.

3. Et program plukket ut faste og frie stikkord i postene og sorterte disse enkeltvis. For hvert stikkord tok en også med registreringsnummer, sted, årstall, motivkode samt de faste stikkordene under sortering og utskrift. En kunne således kombinere et stikkord med en eller flere av de andre opplysningene under søking.
4. Fritekstdelene av datafilen ble behandlet for seg. Hver fritekst ble plukket ut sammen med registreringsnummeret. Til denne samlingen fritekster ble det laget en konkordans og en indeks ved hjelp av standard tekstbehandlingsprogrammer. Disse listene hadde referanser til registreringsnummer på bildet der ordet forekommer i friteksten.

I store trekk kan en si at Fotoprosjektet svarte til sin hensikt som et prøveprosjekt i forbindelse med et utredningsarbeid. Ved hjelp av ulike listeprodukter fikk Fotoutvalget anledning til å sette seg inn i viktige sider ved databehandlingsarbeid og utifra faglig relevante data vurdere hvordan databehandling kan utnyttes ved arkivering og fremsøking av opplysninger om bilder. For senteret var det interessant å få erfaring for at en KWIC-konkordans (Key Word in Context) av fritekstene ble vurdert som et viktig listeprodukt ved fremsøking av bilder.

Metoden en brukte i arbeidet kan sies å være hensiktsmessig når målet er å gi en rekke ulike listeprodukter med opplysninger om et begrenset antall bilder. Men den er ikke hensiktsmessig dersom ønsket er å demonstrere et EDB-opplegg som kunne operere på en database som omfatter f. eks. 500000 bilder.

Etter at oppdraget fra Norsk Kulturråd ble avsluttet, har senteret for demonstrasjonsformål lagt data fra Fotoprosjektet til rette for bruk av tekstsøkesystemet NOVA—STATUS. Interesserete kan kontakte senteret for nærmere orientering.

Fotoutvalget i Norsk Kulturråd har utarbeidet en innstilling datert 5.5.1976 vedrørende registrering, arkivering og oppbevaring av gamle fotografier i Norge. En del av innstillingen omhandler forslag om opprettelse av et sentralregister for billedmateriale bygd på bruk av EDB. Innstillingen er behandlet av Norsk Kulturråd som i første omgang vedtok bl. a. å styrke innsatsen for fotobevaring og i samarbeid med Riksarkivaren og Riksbibliotektjenesten å arbeide videre med planer om et koordinerende sentralorgan.

ICCH

Third International Conference on Computing in the Humanities

Kongressen holdes 2. - 5. august 1977 i Waterloo, Ontario, Canada. I følge kongressplakaten skal følgende tema behandles:

»Frontiers between language and literature
Semantics
Literary stylistics
Lexicography
Language teaching
Computer aided instruction
Classics
Medieval studies
Historical studies
Philosophy
Public service systems
Information retrieval
Computing techniques and hardware
Music
Archaeology
Sculpture
Graphics»

Påmelding til Professor Paul Bratley
Departement d'informatique
Université de Montréal
C.P. 6128, Succursale A
Montréal, Canada
H3C3J7

NOVA STATUS

TEXT RETRIEVAL SYSTEM

I det forrige nummer av Humanistiske Data (Nr. 1–2 1975) ble det meldt at tekstsøkesystemet STATUS var implementert på UNIVAC 1110, Universitetet i Bergen. Vi vil i denne artikkelen nevne en del egenskaper ved systemet.

Mange vil gjennom sitt arbeid bli stilt overfor oppgaver og spørsmål av typen:

- finn alle dokumenter som er kommet inn i et bestemt tidsrom og som omhandler en spesiell sak.
- lag en oversikt over alle gjenstander innen et museum som oppfyller visse spesifiserte krav.
- finn alle setninger hos en forfatter som inneholder bestemte ord eller ordkombinasjoner.

Det som er felles for disse og analoge problem er at det er en arbeidskrevende prosess å løse oppgaven. Selve fremfinningen kalles ofte for informasjonssøking, og vi er vant til at den foregår i arkiver og biblioteker, eller ved oppslag i registre o.l., og at den utføres manuelt.

En manuell søking i arkiver fungerer ofte utmerket på grunn av et godt utbygd klassifikasjonssystem. Det finnes imidlertid situasjoner hvor man ved hjelp av den tradisjonelle metode ikke får tak i alle relevante dokumenter (ordet »dokument» brukes her i en vid betydning — en hvilken som helst tekstenhet kan utgjøre et dokument, f.eks. en gjenstandsbeskrivelse, »sammendrag» av en artikkel, et kapittel i en bok o.l.). Det kan skje når det letes etter et dokument ut fra en helt annen synsvinkel enn den som ble lagt til grunn da dokumentet ble klassifisert. Grunnen til at det tradisjonelle systemet svikter i slike situasjoner, er at det under arkiveringen blir bestemt en gang for alle hvor et dokument skal ligge.

I de senere år har EDB gjort sitt innpass også på dette området. Man kan i dag ved hjelp av en terminal som en tilknytter en datamaskin, lete i store datasamlinger. Leting som ellers ville tatt timer, kan nå utføres i løpet av noen sekunder.

EDB-systemer for informasjonssøking kan grovt deles i to typer:

- systemer hvor dokumentene er representert i en kodet form
- systemer hvor dokumentene er representert i sin helhet.

Systemer som krever at teksten det skal letes i, er representert i en kodet form, lider ofte av de samme svakheter som den manuelle metoden. Svakheten ligger nettopp i kodingen av informasjonen. Denne kodingen er en ressurskrevende prosess, foruten at den alltid vil resultere i tap av informasjon, informasjon som senere kan bli betraktet som vesentlig. Disse problemene elimineres ved fulltekstsøkesystem. NOVA-STATUS er et fulltekstsøkesystem. Det er en videreutvikling av systemet STATUS I,

som Rasjonaliseringsdirektoratet har kjøpt av den engelske forskningsinstitusjonen A. E. R. E. Harwell. Systemet er som tidligere nevnt implementert på UNIVAC 1110 ved Universitetet i Bergen, og vil i nær framtid også være tilgjengelig på universitetsmaskinene i Oslo og Trondheim.

Under arbeidet med NOVA-STATUS er det lagt stor vekt på at systemet skal være lett å konvertere til nye maskiner, bl.a. er hele systemet programmert i FORTRAN.

Spørsmålet som sikkert mange stiller seg er om systemet er vanskelig å bruke og om det er kostbart i drift. Under utformingen av dialogen bruker—maskin har man hatt i tankene at denne type system er et verktøy som i mange tilfeller skal benyttes av personer som ikke har noe kjennskap til EDB. Systemet er derfor enkelt å betjene. Det eneste brukeren trenger er en terminal som kan kobles til telefonnettet.

Har man overført et helt forfatterskap eller deler av det for datamaskinell behandling, kan man enkelt søke etter bestemte ord eller ordkombinasjoner — gjerne i form av synonymlister dersom man er interessert i tema eller stil.

Når »Musikk fra en blå brønn» av Torborg Nedreaas er datamaskintilgjengelig, er det f.eks. lett å hente fram alle teksteksempler på bruken av »blå» eller »blå brønn». Det samme gjelder bruken av fargeadjektiver hos f.eks. Henrik Wergeland dersom hans litterære produksjon lå klar for automatisk tekstbehandling.

Har man lagt til rette en kulturhistorisk database og på terminalen skriver spørsmålet

kiste?

vil systemet finne alle dokumenter som inneholder ordet »kiste», og skrive resultatet av søkingen ut på terminalen. Man kan da på en enkel måte få vist på terminalen de forskjellige dokumentene, slik at man kan avgjøre om det er et relevant dokument.

Skriver man f.eks.

(kiste eller skap) og rosemalt?

vil systemet gi oss en liste over alle de dokumenter som inneholder ordene »kiste» eller »skap» sammen med ordet »rosemalt».

Ved hjelp av de logiske operatorer OG, ELLER og IKKE, kan man stille så kompliserte spørsmål man bare måtte ønske.

Responstiden på et spørsmål, dvs. tiden fra brukeren stiller et spørsmål til svaret foreligger, er kort uansett hvor kompliser spørsmålet er. Forklaringen på den raske responstiden er at systemet slipper å lete gjennom hele teksten for å finne alle dokumentene som inneholder søkeordene.

Hovedprinsippet som ligger til grunn for at søkingen skal skje effektivt kan kort formuleres som følger:

Når teksten leses inn i maskinen, blir det dannet en ordliste som inneholder alle de forskjellige ordene som er i teksten. Hvert ord i denne listen har henvisninger til de steder i originalteksten hvor ordet forekommer. Enhver søking skjer ved oppslag i denne ordlisten, og oppslagene utføres på en hurtig og effektiv måte. Dette er også hovedgrunnen til at systemet er relativt rimelig i drift.

NOVA-STATUS er et generelt tekstsøkesystem i den forstand at de data som skal inn i systemet ikke behøver å ha en bestemt struktur. Det eneste som kreves er at teksten finnes på et maskinlesbart medium. Data som allerede finnes på et maskinlesbart medium kan derfor lett gjøres tilgjengelig for bruk på NOVA-STATUS. Systemet er blitt presentert for ulike grupper, og reaksjonen på systemet må sies å være positiv. Det er derfor å håpe at dette er et verktøy som også kan være et hjelpemiddel for flere brukere innen de humanistiske fag.

Nedenfor er et eksempel på hvordan en søkesesjon kan arte seg for brukeren.

Databasen inneholder i dette tilfelle data fra Norsk Kulturråds fotoprosjekt.

```
*****************************
*                           *
*   N O V A * S T A T U S   *
*                           *
*   TEXT RETRIEVAL SYSTEM   *
*                           *
*****************************

GI DATABASENAVN :
>foto
GI KOMMANDO
>spørsmål
   S  1
>utsikt fra slottet?
    11 DOKUMENTER FUNNET
GI KOMMANDO
>titler 5
  1 BILDENR:475
  2 BILDENR:752
  3 BILDENR:766
  4 BILDENR:767
  5 BILDENR:768
GI KOMMANDO
>les 4
     711
BILDENR:767
   FOTOGR:HOLMSEN, FOTOGR:JOHANNES
   STED:OSLO
   AAR:1865-CA
   KODE:1
   UTSIKT FRA SLOTTET MOT ØST 1865.
   S=KARL JOHANSGT. OSLO OG FJORDEN;
   OVERSIKTSBILDE


GI KOMMANDO
>spørsmål
   S  2
>fotogr:holmsen?
    54 DOKUMENTER FUNNET
GI KOMMANDO
>spørsmål

   S  3
>sted:oslo?

  1316 DOKUMENTER FUNNET

GI KOMMANDO
>spørsmål

   S  4
>fotogr:holmsen.og.sted:oslo.og.kode:1?

    50 DOKUMENTER FUNNET

GI KOMMANDO
>end

DU FORLATER NÅ  N O V A - S T A T U S .

TAKK FOR OPPDRAGET, VELKOMMEN IGJEN Ø
>
```

NORDISKE FORSKERKURS:

Mikrodemografi med tillämpning inom olika forskningsområden

En systematisk datainsamling främst av kyrkboksmaterial för 1800-talet utförs vid Demografiska databasen i Haparanda-Umeå. Individer, familjer och hushåll följs i sitt sociala och demografiska mönster inom ett antal församlingar från Skåne i söder till Tornedalen i norr. Insamlade data bildar en databank för studier inom olika forskningsområden.

Kursen är således tvärvetenskaplig. Den riktar sig till forskare inom humaniora, samhällsvetenskap och medicin. Utifrån övergripande demografiska och socialhistoriska aspekter skall det omfattande källmaterialet granskas och till-lämpas på skiftande frågeställningar inom skilda fack. Kursen ger praktiska övningar att utnyttja databasens information.

Kursort: Demografiska databasen i Haparanda (5 dagar) och universitetet i Umeå (6 dagar) **Kurstid:** 17-27 augusti 1977

Kursledning: Docent Egil Johansson och forskningsledare Jan Sundin, Demografiska databasen, B 105 Humanisthuset, Umeå universitet, 901 87 Umeå (tel. 090-13 82 63 eller 090-12 56 00) samt Produktionsenheten, Demografiska databasen, Box 85, 953 00 Haparanda (tel. 0922-1 14 50)

Anmälningstid: 1 juni 1977

PROSJEKT FOR DATAMASKINELL SPRÅKBEHANDLING, NORDISK INSTITUTT, UNIVERSITETET I BERGEN

har utgitt et skrift med oversyn over arbeidet ved PDS fra starten i 1967 og fram til 1976. Skriftet kan fåes ved henvendelse til PDS, Harald Hårfagresgt. 29, 5014 Bergen-Universitetet.

Orientering om kortvarige EDB-stipend for humanister

Humanistenes opplæringsmuligheter når det gjelder EDB i humaniora, kan ikke sies å være helt tilfredsstillende. Egenopplæring, som enkelte har brukt, er vanskelig og tidkrevende, og kan neppe sies å være en alment anbefalelsesverdig framgangsmåte. Undervisningstilbudet fra institusjoner utenfor humaniora er av begrenset verdi, særlig hvis det skal vurderes som et stimulerende og initierende tiltak for humanister, og det er da heller ikke deres formål.

Kurstilbudet ved NAVF's EDB-senter har vanligvis vært ettermiddagskurs.

De har vært nyttige, men har samtidig vist seg å ha klare begrensninger. Undervisningen blir oppstykket, deltakerne er ikke alltid like opplagte etter arbeidstid og miljøkontakten blir beskjeden. Videre har disse kursene den begrensning at bare de humanister som bor enten i Oslo, Bergen eller Trondheim, hvor senteret har konsulenter, kan nyte godt av tilbudet.

På denne bakgrunn har senteret funnet det hensiktsmessig å introdusere et nytt tilbud, nemlig kortvarige EDB-stipend for humanister i form av et hospitantopphold ved NAVF's EDB-senter for humanistisk forskning. Stipendene er beregnet på humanistiske fagfolk i faste stillinger. Alle fagområder som hører til Rådet for humanistisk forskning, er aktuelle i denne sammenheng, og også institusjoner som arbeider med humanistisk forskningsmateriale.

Stipendene ble første gang utlyst våren 1976, og det kom inn forholdsvis mange søknader. Fire stipend ble utdelt, og 18. oktober begynte hospitantoppholdet i Bergen for tre av stipendiatene. Den fjerde stipendiaten begynte med 2 uker i Trondheim og får resten av perioden i Bergen.

Senteret håper at den erfaring med humanistisk databehandling som er opparbeidet, skal kunne føre stipendiatene rett inn i de sider ved EDB som er relevant ut fra deres faglige interesse. Som et ledd i dette puncher senteret et prøvemateriale som hver av stipendiatene plukker ut som faglig relevant, og som de ønsker å bearbeide med EDB. Dette materialet blir benyttet i selve opplæringsfasen for å gi realistiske og meningsfylte øvelser. Med anvendelse av dette materialet kan de få erfaring med viktige sider ved et prosjekt hvor EDB er et hjelpemiddel. Vi kan bare nevne sider som puncheforskrifter, korrekturlesing og korrigering, bl.a. ved hjelp av editeringsprogram og definering av utskriftsformat. Ikke minst gir dette muligheter for å prøvekjøre ideer og planer for et eventuelt framtidig prosjekt.

Oppholdet innbefatter også generell innføring i EDB, både ved selvstudium og ved forelesninger. Av aktuelle emner kan nevnes: datamaskinens oppbygning og virkemåte, former for kommunikasjon mellom menneske og maskin og hovedprinsipper for programmering. Studium av tidligere og igangværende prosjekter inngår også i opplegget.

Det pedagogiske opplegg er basert på veksling mellom forelesninger, selvstudium, praktiske øvelser på terminaler og regelmessige repetisjonsdager. Samtidig gir opplegget rom for utstrakt manuduksjon, med mulighet for hurtig oppklaring av vanskeligheter som hindrer framdriften, enten de nå er praktiske eller teoretiske. Hver dag avsluttes med en oppsummering og spørsmål og diskusjon i tilknytning til dagens tema.

Formålet er bl.a. at stipendiatene, som tidligere ikke har brukt EDB, etter stipendperioden skal ha så pass erfaring og innsikt i bruk av EDB at de kan ta stilling til om EDB fornuftig kan benyttes som et hjelpemiddel innenfor deres forsknings- og arbeidsområde. Opplegget for høsten 1976 synes å ha svart til forventningene og opplegget vil bli videreført i år.

The 6th International Conference on Computational Linguistics.

The 6th International Conference on Computational Linguistics (COLING 76) ble holdt i Ottawa, Canada 28. juni - 2. juli. Kongressen samlet vel 200 deltakere fra alle verdensdeler. Det ble holdt ca. 65 foredrag, men de fleste av disse ble holdt i parallelle sesjoner, så det var mulig å være til stede bare ved ca. 20 av foredragene. Ca. halvparten av foredragene ble holdt av nordamerikanere (mest fra USA) og nesten halvparten av europeere (mest franskmenn) og noen få av japanere.

Et stort antall foredrag dreide seg om problemer i forbindelse med informasjonssøking. Mange steder prøver en å bygge opp spørsmål-svar-systemer der brukeren stiller spørsmål til maskinen i vanlig språk, og maskinen så gir et svar på grunnlag av de opplysninger den har. Ett problem er her å å få maskinen til å forstå spørsmålet. Det blir oftest analysert semantisk på grunnlag av en forutgående morfologisk og syntaktisk analyse. Et annet og vanskeligere problem er å få maskinen til å komme med de riktige svarene. Det er ikke så vanskelig der opplysningene er gitt direkte på forhånd. Verre er det når en må slutte seg til opplysninger som ikke eksplisitt er gitt. Ut fra opplysningen »Per hadde hendene i fanget» bør maskinen kunne svare ikke bare på »Hvor var hendene til Per?», men også »Hvor var fingrene til Per?». Det siste forutsetter at en på forhånd har lagret opplysninger om den ytre verden, f.eks. at fingrene er en del av hendene. - Det er også hevdet at maskinen også må fores med beskrivelser av visse vanlige hendinger og situasjoner. Ut fra opplysningen »Per gikk på en restaurant og spiste middag» vil et menneske som regel svare bekreftene på spørsmålet »Satt Per og spiste?». Hvis maskinen skal kunne gjøre det samme, må den »vite» at det er vanlig å sitte når en spiser. Vi må altså gi maskinen en detaljert beskrivelse av hva som vanligvis skjer når en går på restaurant. Slike beskrivelser har vært kalt »frames», »scripts» eller »schemata».

Jeg har her gitt en generell (og forenklet) beskrivelse av en del problemer som ble tatt opp i mange foredrag. Noen diskuterte om det var ønskelig med slike »frames», om en heller skulle ha mer generelle regler for hvordan en skal slutte fra en opplysning til en annen. Hvis en legger inn slike »frames», er det også et problem hvor mye informasjon en skal ha med. Prinsippielt kan en tenke seg at all vår opplagrede kunnskap om verden kan være en forutsetning for å forstå et spørsmål og svare på det riktig. Men å forhåndslagre all vår kunnskap om verden er både teoretisk og praktisk umulig. Hvis en lager slike »frames», er det også et problem hvordan informasjonen skal lagres.

Jeg har gått såpass nøye inn på disse problemstillingene siden de var så sterkt framme i foredragene. Men også andre, mer »tradisjonelle» emner var representert. Jeg skal nevne noen få.

Et par av plenumsforedragene var reint lingvistiske. Ch. Fillmore (Berkeley) hevdet at frames eller schemata var nødvendige også for en vanlig (ikke-maskinell) tolking av setninger. Eva Hajičová (Praha) snakket om forholdet mellom spørsmål og svar. En riktig språklig forståelse for dette er nødvendig for å kunne bygge spørsmål-svar-systemer.

Språktilegnelse hos barn har vært mye diskutert i seinere år. En gruppe fra Edmonton, Canada prøver å simulere barns forståelse og produksjon av språk i ulike faser.

Automatisk språkoversettelse har ikke vært så sterkt i skuddet i seinere år som det var for et par årtier siden. Men det er likevel flere større prosjekter i gang. I det tospråklige Canada er problemet spesielt påtrengende, og en gruppe i Montréal (TAUM) har i flere år arbeidd med et engelsk-fransk oversettelsessystem. I dag har de kommet så langt at en rekke artikler og offisielle dokumenter kan trykkes direkte uten noen etterredigering slik det kommer ut av maskinen.

Et av de få foredragene om fonologi ble holdt av Benny Brodda (Stockholm). Han har (for Patentstyrelsen i Sverige) prøvd å måle kvantitativt den fonologiske ulikheten mellom ord ved hjelp av fonologiske faktorer.

Et prosjekt, som nok er lite, men som kan få stor praktisk verdi, står Andy Tretiakoff (Paris) for. Hun har laget et apparat som overfører tegn på magnettape (vanlige kasetter) til blindeskrift, og som gjør at en blind ved å feste apparatet til kasettspilleren kan lese det som står på tapen i blindeskrift. Apparatet kan også koples til datamaskin og til elektroniske kalkulatorer.

Etb (edb)
Ordbehandling
Textbehandling
Word Processing

– endnu en gang, nye termer og begreber som er med til at forvirre og udbygge vort „præstesprog" og dermed gøre kommunikationen med omverdenen endnu mere besværlig. Bedre bliver det ikke af, at vi ikke selv ved, hvorledes vi skal fordele området på disse begreber, fordi store leverandører af salgsmæssige grunde skaber deres eget sprogbrug.

Men, der er i høj grad tale om begreber, som vil komme til at præge edb-hverdagen i de kommende mange år – hvad enten man i det amerikanske begreb Word Processing lægger den automatiserede sekretærfunktion, eller man går til den anden fløj – tekstproduktionen i forbindelse med avisfremstilling.

Eksperter på edb-området har spået, at området Word Processing vil belægge mere end halvdelen af den samlede edb-kapacitet, når vi når et stykke ind i 80'erne. I dag tænker de fleste på de såkaldte korrekturmaskiner, når man taler om Word Processing, men begrebet får en hel anden valeur, når man prøver at se, hvilke muligheder der opstår, når man tager skærmterminaler, datorer og databaser ind i billedet.

Mange ser i begreberne en revolution indenfor kontorverdenen.

De, der kun har sigte på de automatiserede skrivemaskiner, og brevudskrivning på datamaterne ved en kombination af en række databaser, frygter, at vi her står overfor den helt store papiroversvømmelse, alle vil blive lammet af breve – og postvæsenet, som gennem lang tid har nedsat sit serviceniveau, vil bryde helt sammen.

Men hvorfor ikke se det som George E. Pake, Xerox Corp., som udtaler til Business Week, at han venter, der vil ske en revolution på kontoret i stil med den, der skete i hjemmene i forbindelse med TV's indtog – „jeg vil blive i stand til at kalde dokumenter frem på min skærm, og modtage min post og meddelelser på anden måde"

data

Oppmerksomheten henledes på tidsskriftet »Data» som er et nordisk datatidsskrift for EDB-brukere og spesialister. Tidsskriftet inneholder foruten spesialstoff også mange artikler om allmenne EDB-emner og gir gjennom annonsene et godt innblikk i produktutvikling og nye bruksområder. I DATA 11/76 er det bl.a. en artikkel om word processing som introduseres av en leder som gjengis her (noe forkortet):

TEXT

Programpakken TEXT
status pr. januar 1977.

Programpakken TEXT ble utviklet i Oslo i 1970—71 og presentert for aktuelle brukere som en komplett programpakke i mai 1971. I en femårsperiode fram til sommeren 1976 var programpakken i stadig bruk og representerte det viktigste programtilbud til EDB-brukerne ved Det historisk-filosofiske fakultet ved Universitetet i Oslo. Da universitetet skiftet datamaskin sommeren 1976, ble programmene overført til Studentsamskipnadens anlegg som er av samme type som universitetets gamle. TEXT har derfor også vært tilgjengelig for brukerne høsten 1976 og har vært benyttet av en del prosjekter. Fra 1.1.1977 er det imidlertid lite aktuelt å kjøre på Studentsamskipnadens anlegg, og denne versjonen av TEXT blir derfor nå brukt bare i spesielle tilfeller. Overføringen av TEXT til universitetets nye lokalanlegg, DEC—10, er i full gang. Før denne overføringen er fullført, vil TEXT ikke være tilgjengelig for normal bruk.

I perioden 1971—76 er det foretatt en del justeringer og forbedringer i programmene, og det er dessuten kommet til en del nye programmer og andre utvidelser. Sommeren 1976 omfattet systemet programmer for følgende oppgaver:

1. Innlesning, redigering, korrigering av tekst.
   (Tekstbehandlingsprogrammer på universitetets daværende maskin var ikke særlig velegnet, og det aller nødvendigste ble da lagt inn i TEXT).
2. Omforming av teksten til en bestemt struktur som medfører betydelig effektivisering av tekstsøking, ordlisteproduksjon osv.
3. Produksjon av ordlister med frekvenser, basert på hele eller vilkårlige deler av et materiale.
   Brukeren kan velge hvordan ordlisten skal sorteres (alfabetisk, finalalfabetisk, etter frekvens).
4. Søking etter bestemte ord i teksten og utskrift med kontekst av vilkårlig lengde. Søking etter tegnsekvenser i teksten med vilkårlige intervaller mellom sekvensene. Utskrift med kun tekst.
5. Produksjon av konkordanser på grunnlag av hele eller vilkårlige deler (og kombinasjon av slike) av materialet. Alle ord som ikke er definert som stoppord, angis med antall forekomster, og alle forekomster skrives ut med kontekst og kildereferanse.
6. Diverse statistikk, beregning av ordforråd, LIX osv.
7. Muligheter for klassifisering av brukerdefinerte enheter i teksten (f.eks. setninger), og parallellkjøring av koder og tekstinnhold, f.eks. utskrift av alle setninger som er klassifisert med bestemte koder.
8. Enkel statistikk på kodede data samt overføring til DDPP-format (DDPP – programpakke for statistikk) for mer omfattende statistikk.
9. Diverse spesialprogrammer knyttet til TEXT, f.eks. et program for studier av konsonantallitterasjon.

Pr. 14. januar 1977 er programmene som er beskrevet under punktene 2 og 4 overført til DEC—10 og vil være klar for bruk med det første. Punkt 1 (innlesning redigering osv.) er erstattet med standard

tekstredigeringsprogrammer på DEC—10, og det er laget forbindelses-programmer mellom disse og TEXT. Videre er det laget program for overføring av TEXT-data fra det gamle anlegget.

Av det som gjenstår, vil 3, 5 og 6 bli prioritert og ventes ferdig utpå vinteren. 7 og 8 kommer senere, mens 9 ikke vil bli overført uten at det framkommer spesielle ønsker fra brukerne.

Når punktene 1—6 er klare på DEC-10, vil de aller fleste brukerne kunne utføre sine oppgaver som før. Dessuten vil alle programmer som overføres ha muligheter for interaktiv bruk.

TVERRFAGLIG INTERESSEGRUPPE I LITTERÆR OG SPRÅKLIG STATISTIKK

På den nasjonale konferansen om humanistisk databehandling som ble holdt på Gol 3. - 5. november i fjor, kom det i gruppen for språkstatistikk bl.a. frem forslag om at NAVF's EDB-senter burde ta initiativet til dannelsen av en interessegruppe for distribusjon av relevant språkstatistisk litteratur. Ettersom metodeproblemene ved bruk av statistikk i grove trekk er felles i språk - og litteraturforskningen, er det all grunn til å etablere en slik tverrfaglig interessegruppe. Tanken var at interesserte kunne tegne seg på en adresseliste og derved få tilsendt fra senteret tidsskriftartikler, konferansestoff, opplysninger om egnete bøker o.l., i det hele tatt slikt som kan være av interesse i denne sammenheng.

Det er en forutsetning for en vellykket distribusjonssirkel at hver enkelt deltaker er villig til å informere de øvrige i gruppen om slik litteratur, formidlet via senteret og dets adresseliste. Ved valg av stoff vil det særlig bli lagt vekt på metodiske problemer ved bruk av statistikk i språklig og litterær sammenheng.

På sikt kan det være aktuelt for senteret å arrangere seminar og kurs med utgangspunkt i denne gruppen. Tiltak i tilknytning til senterets oversettelse av en lærebok i språkstatistikk kan også være aktuelle.

Det er anledning til å tegne seg på adresselisten for andre enn de som deltok på Gol. og dette er også et tilbud til de som ikke selv har arbeidet på dette feltet, men ønsker å orientere seg på området.

Interesserte bes sende navn og adresse til

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

COMPILING

(Computers in Linguistics)

Dette er et nystartet meldingsblad for Nordisk samarbeidsgruppe for datamaskinell språkbehandling. Meldingsbladet gir orientering om større og mindre faglige sammenkomster på feltet og bringer videre informasjon om virksomhet på forskningssiden som kan ha interesse for andre.

Av opplysninger som finnes i det siste nummer kan nevnes at Samarbeidsgruppen planlegger en Workshop in Computational Linquistics i California sommeren 1977. Deltakerne, 15 skandinaver og 15 amerikanere, vil få førstehånds erfaring i kjøring av amerikanske systemer for automatisk databehandling av naturlig språk. Arrangementet har fått støtte fra Nordisk kulturfond. Samarbeidsgruppen vil stå for utvelgelsen av deltakerne fra skandinavisk side. Det vil i løpet av våren bli sendt ut mer informasjon om arrangementet.

KVIKKIS er navnet på en serie programmer for grunnleggende tekstbehandling som finnes tilgjengelig ved universitetsanleggene i Bergen og Trondheim. Navnet henspiller på noe av det en kan få gjort ved hjelp av programpakken (KWIC – Indeks – Søking).

Programmene er lagt til rette slik at de skal være enkle å bruke og kreve lite kjennskap til datamaskinens styrespråk. I tillegg skal brukeren kunne angi en del sentrale parametre.

Kvikkis

tekstbehandlingsprogrammer

Oversikt over de enkelte funksjoner i KVIKKIS

A. Produksjon av ordlister til en tekst. Følgende lister kan produseres:

Til alle lister kan en angi et sett med ord som en ikke ønsker å få ut (stoppordliste) eller et sett med de ord en ønsker å få ut (plussordliste).

Ordliste, indeks og KWIC-konkordans kan ordnes forlengs (vanlig alfabetisering) eller baklengs (ordningen skjer fra siste tegn i ord og forover).

For de lister som inneholder referanser til teksten kan en velge mellom to typer:
Referanse til linjenummer i teksten (tekstlinjene nummereres fortløpende).
Referanse til inndeling av tekst i f.eks. bok, kapittel, side.
Opplysninger om kapittelskift, sideskift må da markeres på en spesiell måte i teksten.

Brukeren står selv fritt til å velge hvilke tegn som skal være skille mellom ord og setninger.

B. Statistikk.

I tillegg til ordlistene får en ut noe statistikk over tekstmaterialet.

Antall setninger
gjennomsnittlig setningslengde
fordeling av setninger etter antall ord i setning
antall ord
antall ulike ord
type/token forhold
leselighetsverdi (lix)

C. Søking.

KVIKKIS inneholder også prosedyrer for å søke ut deler av en ordliste eller konkordans utifra oppgitte mønstre som f.eks. prefiks, suffiks og ordkombinasjoner.

TYPE	INNHOLD
ORDLISTE	ord, absolutt frekvens
FREKVENSORDLISTE	ord, absolutt, relativ og kumulert relativ frekvens ordnet etter avtagende frekvens
INDEKS	ord, absolutt frekvens, referanse til tekst
KWIC-konkordans (key-word in context)	ord, absolutt frekvens, referanse, kontekst til ord (fast kontekstlengde før og etter ord).
KWOC-konkordans (key-word out of context)	ord, absolutt frekvens, referanse, kontekst til ord (*setning som kontekst*).
KLIW (key letter in word)	konkordans på tegnnivå innen hvert ord

D. Merking (tagging).
I KVIKKIS finnes det program for merking av ord i en tekst, eksempelvis med ordklasse. Utgangspunktet for ordklassemarkeringen er en KWIC-konkordans til teksten. Ordklassen til de aktuelle ord angis sammen med første og siste linjenummer i konkordansen. Dersom et ord er homograf angir en først den hyppigste ordklasse og etterpå de mindre hyppige som unntak. Programmet setter så merkene inn i teksten for alle forekomster av ordet.

Programtilbudet i KVIKKIS vil bli utvidet etter hvert som behovet melder seg. I første omgang tar en sikte på å innarbeide program for å liste ut ordsammenstillinger (collocations).

Det er utarbeidet en brukerhåndbok til KVIKKIS. Denne kan fås ved henvendelse til NAVF's EDB-senter for humanistisk forskning. I tillegg holdes det regelmessige kurs i bruk av programmene.

KONSULENTHJELP—PUNCHEASSISTANSE

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

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

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

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

Adresser:

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

Oslo
EDB-konsulent Ivar Fonnes
c/o Historisk institutt,
Universitetet i Oslo
Postboks 1102 — Blindern
Oslo 3

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

SPRÅKDATA
Avdelning för språklig databehandling
Göteborgs Universitet

utgir årlig forskningsrapporter som gir et allsidig bilde av virksomheten. Interesserte vil kunne få slike ved henvendelse til Göteborgs Universitet

SPRÅKDATA
Norra Allégatan 6
S 41301 GÖTEBORG

COMPUTER ARCHIVE OF MODERN ENGLISH TEXTS (CAMET)

NAVF's EDB-senter har fått flere henvendelser fra forskere som er interessert i å følge med i oppbyggingen av CAMET, som foregår ved University of Lancaster, England. CAMET er tenkt å være et engelsk-språklig motstykke til det amerikanske Brown-Corpus som inneholder en million ord fra ulike genre. Senteret har nå innledet et samarbeid med dosent Stig Johansson, Britisk institutt, Universitetet i Oslo for å få rede på status for CAMET. Dosent Johansson vil fra 9. januar oppholde seg en uke ved Universitetet i Lancaster for å sette seg inn i forholdene. Nærmere opplysninger om CAMET vil bli gitt i et senere nummer av Humanistiske Data.

PPTT (ProgramPakke for Tekstbehandling, Trondheim)

betegner en samling programmer for tekstbehandling som er organisert i en pakke. Programsystemet er implementert på hoved*dataanlegget ved Universitetet i Trondheim* og tilgjengelig for brukere der. Programpakken kan også kjøres på universitets*anlegget i Bergen. Programmene er stort* sett skrevet i NU ALGOL. Pakken er ment å skulle løse de vanligste oppgavene innen kvantitativ tekstbehandling, og mulighetene vil bli utvidet etterhvert som behovet melder seg eller nye ideer oppstår.

De enkelte programmene henger sammen slik at for å løse en del av dem, må en ha behandlet teksten med andre program tidligere (jfr. skisse i brukerbeskrivelsen for PPTT). Dette er gjort dels for å økonomisere med maskintid, dels for å skape fleksibilitet i systemet og dels for å dele opp problemene i mindre, separate — og dermed mer oversiktlige — kjøringer.

Innmating til programsystemet er vanlig tekst (som er gjort maskinleselig ved hjelp av DATA-prosessoren) hvor en står nokså fritt i å bestemme formatet. Begrensningene er at linjelengden ikke må overstige 132 tegn, og at sidetall, kapitler og avsnitt (dersom det ønskes markert) må merkes på en bestemt måte. Derimot står det brukeren helt fritt å definere hvilke tegn som skal oppfattes som skilletegn i teksten.

Hvert ord blir gitt referanser på to måter: den ene angir sidetall og linjenummer på sida, den andre gir referanse etter kapittel, avsnitt i kapitlet, periodenummer i avsnittet og ordnummer i perioden.

De enkelte programmene og deres formål

ORD	—splitter opp teksten i de enkelte orda, knytter referanse til hvert ord, beregner tekstens lix-verdi, antall ord og perioder i teksten.
SORTB	—lager baklengssortert ordliste med absolutt frekvens, ordlengde og referanser.
SORTF	—lager vanlig alfabetisk ordliste med absolutt og relativ frekvens, ordlengde, referanser og såkalt »type/token»-fordeling.
SORTFREKVENS	—lager frekvensordliste etter absolutt frekvens og med kumulert relativ frekvens angitt.
SORTLENGDE	—lager ordliste sortert etter ordlengde.
STAVINGER	—splitter *opp orda i* den alfabetiske ordlista i de enkelte stavinger (gir et forslag som kan korrigeres manuelt), finner antall stavinger.
KWIC	—lager konkordans av teksten med mulighet for å legge inn stoppordliste.
LINJER	—gjør det mulig ved hjelp av editor å søke etter mønstre som går ut over ordgrenser.

Planlagte utvidelser

KLIC	—»key-letter-in-context», dvs konkordans over tegnene/bokstavene i de enkelte orda, med ordet som kontekst.
TEGNKOMB	—angivelse av hvilke grafemer som opptrer ved siden av hverandre (i første omgang to-grafem-kombinasjon).
FRKVTEGNKOMB	—disse grafemkombinasjonene sortert etter frekvens.

SORTSTAVING —alfabetisk liste over forekommende stavinger i teksten etter at maskinens forslag er korrigert manuelt.

FRKVSTAVING —frekvenssortert liste over de samme stavingene.

LENGDESTAV —liste sortert etter lengden av de samme stavingene.

FONEMENSTAV —enstavingsorda i teksten »oversatt» til fonemisk representasjon (gir et forslag som kan korrigeres manuelt) — gjelder norsk tekst.

SORTFONEM —alfabetisk liste over de »fonemiserte» orda.

Brukerbeskrivelse for PPTT er laget og kan fåes ved henvendelse til
NAVF's edb-konsulent
Universitetet i Trondheim
Norges Lærerhøgskole
7000 TRONDHEIM

Ibsens-konkordans

Under ledelse av dr. R.G. Popperwell er det ved Literary and Linguistic Computing Centre, University of Cambridge satt igang arbeid med å lage konkordanser til Ibsens skuespill og dikt. Til i dag er flere av Ibsens tidlige skuespill og diktene overført til maskinleselig form med grunnlag i 100-årsutgaven.

Til visse verk er det også utarbeidet konkordanser. Målet er å distribuere både tekstene og konkordansene til bruk i litteraturvitenskapelig forskningsarbeid. Det er meningen at en fra norsk side skal inngå et samarbeid med britene, dels i form av punching av tekster og dels i form av utarbeidelse av *homografseparerte, lemmatiserte* konkordanser til alt materialet. Det er utpekt en samarbeidsgruppe på norsk side som har ansvaret for planleggingen. Med støtte fra Universitetet i Bergen og Institutt for nordisk språk og litteratur ved Universitetet i Oslo har George M. Gillow hatt et fire måneders engasjement høsten 1976 for å utarbeide planer for den norske delen av et Ibsen-prosjekt. Nærmere opplysninger om tiltaket kan fåes ved henvendelse til NAVF's EDB-senter.

NYTT STYRE FOR NAVF'S EDB–SENTER FOR HUMANISTISK FORSKNING

NAVF har oppnevnt nytt styre for NAVF's EDB-senter med funksjonstid fra 1.8.1976 - 31.12.1978.

Styret har i dag denne sammensetningen:

Professor Egil Pettersen, Nordisk institutt, Universitetet i Bergen (formann).
Avd.leder Carl Erik Ellingsen, Avdeling for elektronisk databehandling, Universitetet i Bergen.
Professor Ottar Dahl, Historisk institutt, Universitetet i Oslo.
Dosent Ådne Findreng, Tysk institutt, Norges Lærerhøgskole, Universitetet i Trondheim.
Senterets leder Jostein H. Hauge, sekretær for styret.