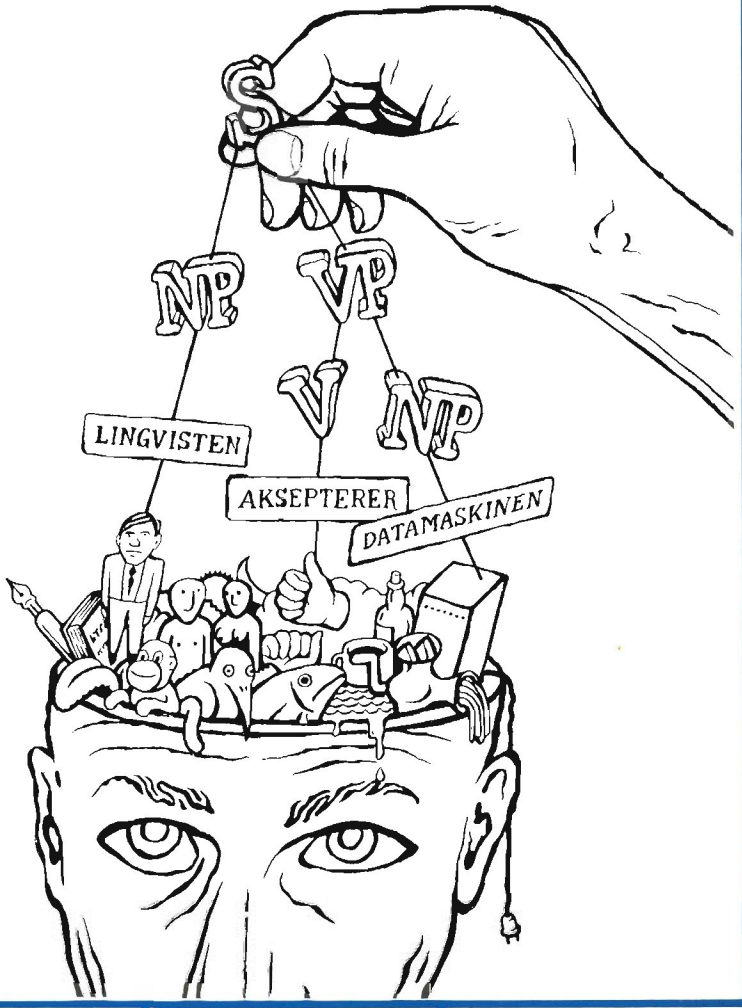
ISSN 0800-6792

humanistiske data 3-84



NAVF

ARTIKLER
RAPPORTER
MELDINGER
SUMMARY

NAVFs EDB-senter for humanistisk forskning

The Norwegian Computing Centre for the Humanities

SENTERETS RAPPORTSERIE

Rapporter utgitt f.o.m. 1980

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

RAPPORT nr. 14. *NOVA*STATUS HÅNDBOK*

Del 1: Søking. Brukerveiledning. 3. opptrykk februar 1983. ISBN 82-7283-011-6 Pris kr. 20.

Del 2: Fil-beskrivelser. Systemdokumentasjon. Utsolgt.

Del 3: Generering og oppdatering av databaser. Utsolgt.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Forts. 3. omslagsside.

humanistiske data 3-84

NAVFs EDB-senter for humanistisk forskning

The Norwegian Computing Centre for the Humanities

NAVF NORGES ALMENVITENSKAPELIGE FORSKNINGSRÅD

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

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

Senteret utgir tidsskriftet Humanistiske Data (3 nr. pr. år) og en rapportserie (34 er utkommet pr. 20.11.84).

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

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

Interesserte kan kostnadsfritt bestille årsmelding og Humanistiske Data (kr. 60,- for institusjoner).

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

Senterets adresse: Harald Hårfagresgt. 31, Boks 53, 5014 Bergen-Universitetet. Tlf. (05) 212954/55/56

Artikler, rapporter, meldinger mottas. Redaksjonen avsluttet 20. november.

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

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

Medarbeidere fra Senteret i dette nummer:
Jostein H. Hauge, Ole Lauvskar, Kristin Natvig, Espen Ore, Øystein Reigem.

Fotosats i kommunikasjon med Univac 1100/82.
Sats: Universitetet i Bergen/NAVFs EDB-senter for humanistisk forskning.
Grafisk design og montasje: Kristin Natvig.
Trykk: John Grieg A/S
Forsidebildet er tegnet av Øystein Reigem.

Innhold

Artikler

Rapporter

Hva er datalingvistikk?

Helge J. Jakhelln Dyvik

Forskning og utviklingsarbeid i skjæringsfeltet mellom lingvistikk og datavitenskap er et område der vi ser en eksploderende aktivitet internasjonalt. Nye og fruktbare forbindelseslinjer utvikler seg mellom fagområder som mange er vant til å plassere på ulike sider av høye gjerder. Slikt samarbeid gjelder ikke bare praktiske applikasjoner; stadig oftere ser vi også spennende teoretiske arbeider fellesforfattet av lingvister og datavitere, og stundom også av kognitive psykologer, matematikere, logikere og/eller filosofer i samarbeid. Dette åpner også nye finansieringskilder for humanistisk forskning, og ikke bare for kortsiktige prosjekter med sikte på å levere et omsettelig produkt. Det er verd å merke seg at større internasjonale datakonserner som f.eks. Xerox ser seg tjent med å fulltidsansette bl.a. lingvister ved sitt Palo Alto Research Center i California, hovedsakelig for å drive grunnforskning og følge med i hva som skjer innenfor teoretisk syntaks og semantikk. Universiteter og andre åpne forskningsinstitusjoner bør kanskje ikke *bare* være tilfreds med at noen av de lingvistiske teorier som debatteres mest i faglitteraturen for tiden, er utviklet ved Xerox PARC.

Hva går så all denne datalingvistiske virksomheten ut på? Det er svært mangfoldig; men skulle man prøve å gi en sammenfattende karakteristikk, måtte den være at det dreier seg om å utvikle programvare for grunnforskningsformål eller for direkte praktisk anvendelse, som *inkorporerer lingvistisk innsikt,* eller, om man vil, *simulerer språklig kompetanse,* på ett eller annet nivå, banalt eller sofistikert. Jeg skal nevne noen hovedtyper av systemer som utvikles for praktiske formål.

Spørsmål-svar-systemer er kanskje den vanligste typen datalingvistiske systemer som er i praktisk bruk. Dette er systemer som tillater en bruker å be om informasjon fra en database, og som inneholder en såkalt «natural language front end» som gjør det mulig å stille spørsmål i et avgrenset utsnitt av et naturlig språk, f.eks. engelsk, og som genererer naturlig-språklige svar innenfor det samme utsnitt. Det er her tale om spørsmål som stilles i skriftlig form fra tastaturet, altså ferdig segmentert naturlig språk - automatisk analyse av talt språk på dette nivå ligger nok ennå i fremtiden. Spørsmål-svar-systemer må kunne tolke de naturlig-språklige spørsmålene, det vil si, de må i en eller annen forstand kunne «forstå» deler av naturlig språk, og de må kunne «resonnere» på grunnlag av det de har forstått. Systemene setter et menneske i stand til å føre en slags naturlig-språklig dialog med en maskin. Det vil riktignok lett kunne bli en lite tilfedsstillende dialog,

særlig hvis systemets språklige «kompetanse» er for rudimentær. Hvis f.eks. kompetansen bare omfatter setningsnivået, kan det fort bli frustrerende – som i et naturlig-språklig grensesnitt som ble utviklet til en database med informasjon om polarekspedisjoner. Der kunne man riktignok – på fransk – stille spørsmål som «hvilken båt deltok i toktet da og da», og få svar. Men hvis man så fortsatte med «hvor mange personer var ombord i den?», ville systemet stå fast og desorientert spørre hva *den* refererte til. En virkelig dialogkompetanse forutsetter et system som bl.a. kan ta vare på referensiell informasjon fra setning til setning: bare da kan anaforiske pronomen som «han» og «den», elliptiske konstruksjoner o.l. tolkes. Det blir utviklet slike «dialogkompetente» systemer også, og de utgjør da både en ramme man kan studere tekstlingvistiske problemstillinger innenfor, og et anvendelsesområde for tekstlingvistisk innsikt.

Nå er ikke alle enige om at det er hensiktsmessig å bruke tid og krefter på å utvikle lingvistisk sofistikerte spørsmål-svar-systemer. Høyst sannsynligvis vil det alltid være slik at det bare er begrensede og «regimenterte» undermengder av naturlige språk et dataprogram vil kunne analysere. (Enkelte hemningsløse optimister – eller blir det pessimister? – tror riktignok at det prinsipielt ikke er noen grenser for hvilke intellektuelle funksjoner en datamaskin kan programmeres til å utføre: men da snakker vi i hvert fall ikke om den overskuelige fremtid.) Dette betyr at en bruker forholdsvis raskt vil stange hodet i taket for systemets språkevner. Etter hvert som brukerne oppdager systemets grenser, vil de raskt tilpasse seg disse og holde seg til den kode systemet forstår. Men siden mennesker så raskt tilegner seg en slik kode, er det rimeligere å satse på enkle «front ends» og brukerinstruksjon enn på ambisiøse språksystemer, mener enkelte. Dette har utvilsomt noe for seg. Den vesentlige fordel ved naturlig-språklige spørsmål-svar-systemer er antagelig at de utgjør en lav barriere for nye og uøvede brukere, som kan tenkes å vike tilbake for å måtte lære et spesialisert dataspråk.

Naturlig-språklige «front ends» brukes også i andre forbindelser, f.eks. i forbindelse med *ekspertsystemer*. Et ekspertsystem er et system som i en eller annen representasjonsform har «ekspert-kunnskap» om et domene – det kan være om bank-transaksjoner, oljeboring, medisinsk diagnose, eller forsåvidt hva som helst – og en evne til å «resonnere» på grunnlag av den kunnskapen, som simulerer den menneskelige ekspertens evne. Naturlig-språklige grensesnitt brukes ikke bare for å «konsultere» slike ekspertsystemer, men også for å bygge dem opp, altså slik at ny kunnskap også kan formidles til systemet ved hjelp av naturlig-språklig input. Forskningen omkring kunnskaps-representasjon har forbindelseslinjer til lingvistisk semantikk.

Andre språk-kompetente systemer fungerer som *hjelpemidler for skribenter* ved å avsløre stavefeil, syntaksfeil, eller endog stilistisk uheldige uttrykksmåter. Slik programvare har åpenbart også et anven-

delsesområde innenfor språkpedagogikken.

Så har vi naturligvis de systemene som strever med å bli datalingvistisk respektable igjen: Systemene for automatisk oversettelse. I 50-årene var forventningene til maskinoversettelse ubegrensede og optimismen uhemmet: Fullt automatisert oversettelse av høy kvalitet ble antatt å ligge like om hjørnet. Når vi i dag ser tilbake på de systemene som ble utviklet i denne perioden, virker de håpløst primitive, både lingvistisk og datafaglig. Optimismen fikk et grunnskudd midt på 60-tallet, og maskinell oversettelse ble deretter stort sett tatt av den datalingvistiske dagsorden.

I de senere år har vi sett en fornyet interesse for maskinoversettelse, sammen med et mer nyansert og realistisk syn på hva som er mulig enn det som preget de tidlige forsøkene.

Både lingvistikk og databehandlingsteori er kommet adskillig lenger i dag enn de var da de første systemene ble utviklet. De oversettelsessystemene som utvikles i dag, benytter seg av datalingvistiske teknikker og lingvistisk teori fra de sene 70-årene og 80-årene. Blant annet inkorporerer de gjerne elementer fra nyere grammatiske teorier, utfører semantisk analyse, går til dels over setningsnivå, inneholder såkalte «fail-soft measures» som innebærer at også setninger systemet ikke kan analysere, får en slags oversettelse, og er modulært oppbygget, f.eks. slik at de grammatiske komponentene er klart adskilt fra analysekomponentene, osv. Det siste innebærer at systemene er lettere å modifisere. Et eksempel på et stort europeisk prosjekt av denne art, er EUROTRA, et oversettelsesprosjekt i regi av EF, med sikte på å muliggjøre automatisert oversettelse mellom EF-språkene.

Grunnforskning omkring maskinoversettelse er det lite av i Vesten. Men i forbindelse med det store japanske prosjektet for å utvikle 5. generasjons datamaskiner – eller «kunnskapsmaskiner», som det skal bli – er maskinell oversettelse et meget viktig delområde.

Datalingvistisk forskning har flere anvendelsesområder enn de nevnte, f.eks. i forbindelse med informasjonssøking i tekstmasser, som er et felt det forskes i her i landet, bl.a. ved Institutt for rettsinformatikk i Oslo. Jeg skal ikke bruke tid på å omtale flere slike områder, men heller stille spørsmålet hva slike systemer alment må kunne gjøre for å virke. Felles for dem er et større eller mindre element av simulert *språkforståelse*. Det vil si, systemene må ha representert en språklig kunnskap som setter dem i stand til å analysere et språklig input på ulike nivåer, velkjente i lingvistisk sammenheng. Systemer for talegjenkjennelse må kunne gjennomføre en fonologisk analyse som gir en fonemisk representasjon som output. Skriftlig input overflødiggjør dette nivået; men dernest må det gjennomføres en morfologisk og en leksikalsk analyse som gir en representasjon av input som en streng av formativer, en syntaktisk analyse som gir en syntaktisk struktur, en semantisk analyse som gir en logisk form, og kanskje en pragmatisk analyse som supplerer denne på grunnlag av kontekst og gir som

NO. 1: E82060001_2_1 (08/13/84, 08/13/84, 08/20/84)
"１９８１年度の船舶用電気技術。"

The electrical technology for marine vessel in 1981.

NO. 4: E82060004_2_1 (09/06/84, 09/06/84, 09/06/84)
"マイクロコンピュータ制御による電力節減。"

The energy conservation by the microcomputer control.

NO. 26: E82060043_5_1 (06/13/84, 06/13/84, 06/14/84)
"拡散近似も吟味される。"

The examination is also made the diffusion approximation.

Eksempler på automatisk oversettelse fra japansk til engelsk (ansvarlig: Toyoaki Nishida, Kyoto University).

produkt logiske representasjoner som kan danne input for en resonnerende komponent. For å kunne gjennomføre slike analyser trenger et system bl.a. en bakgrunnskunnskap i form av et rikt strukturert leksikon eller «ordforråd», og videre sett av fonologiske, morfologiske, syntaktiske, semantiske, pragmatiske og deduktive regler. I tillegg til regelsettene, som utgjør systemets språklige «kompetanse», må det spesifiseres analyseprosedyrer, altså algoritmer som trinn for trinn angir hvordan et gitt input skal analyseres på bakgrunn av regelsettene. Det vil si, systemet må ha en «performance»-komponent i tillegg til en «kompetanse»-komponent. På syntaksnivå kalles en slik analyseprosedyre for en *parser*. Parsing inngår i databehandlingsteori alment, men der dreier det seg om syntaksanalyse av programmeringsspråk, altså enkle, konstruerte språk. Parsing av naturlige språk er et empirisk prosjekt og reiser ganske andre problemer.

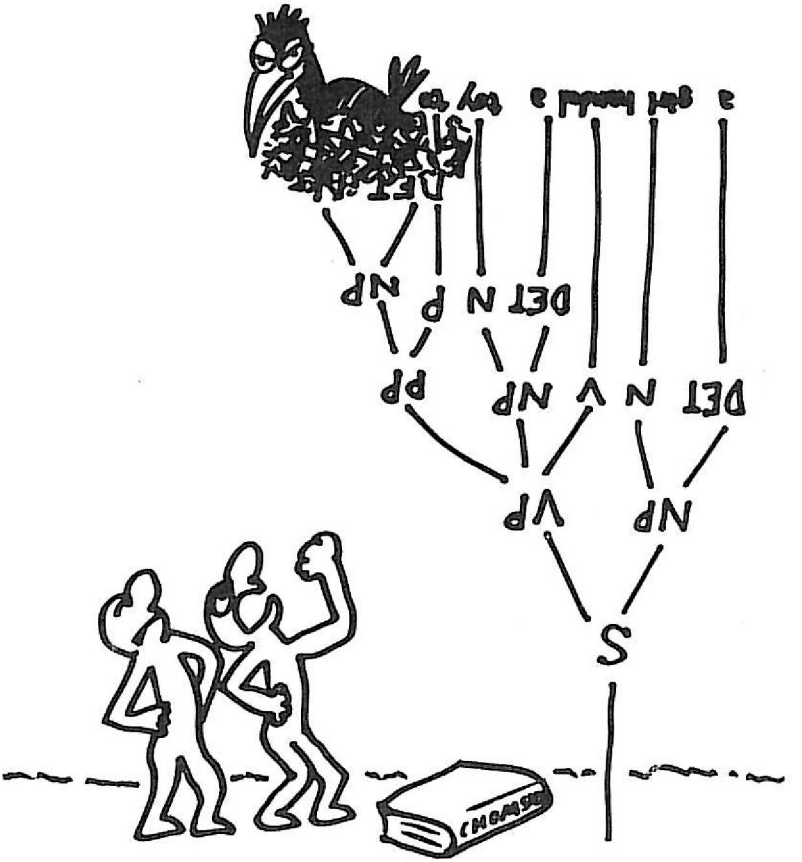
Arbeidet med regelkomponentene kan i stor utstrekning bygge på moderne lingvistisk forskning, i første rekke på arbeidet med formell syntaks innenfor generativ grammatikk, og på nyere formelle semantiske teorier. En aktuell semantisk teori er Montague-semantikken, etter logikeren *Richard Montague*, som utviklet en teori om hvordan setninger i et naturlig språk trinn for trinn kunne oversettes til et logisk formelspråk, som i sin tur var mengdeteoretisk fortolket. I de aller seneste år er det utviklet formelle semantiske teorier – såkalt «situasjonssemantikk» – som ivaretar kompleksiteten i naturlige språk langt bedre enn Montague-semantikken gjør. Situasjonssemantikken vil utvilsomt få betydning for datalingvistikken, og sikkert også for systemer for kunnskapsrepresentasjon. Formelle *syntaktiske* teorier har vært sentrale innenfor teoretisk lingvistikk i tredve år snart, og de har funnet et

utfordrende anvendelsesområde i datalingvistikken, der man nettopp trenger formaliserbare beskrivelser av strukturen i naturlige språk. De tidligste versjonene av generativ grammatikk, Chomskys transformasjonsgrammatikk fra 50- og 60-årene, vakte forventninger om datalingvistisk anvendelighet som ikke helt ble innfridd. Problemet var først og fremst at transformasjonsmekanismen var for kraftig fra et analysesynspunkt: Det var uhyre vanskelig å gjennomføre transformasjoner «baklengs» under en analyseprosess. Isteden ble det utviklet grammatikkformalismer som var skreddersydde for parsing, f.eks. såkalte «Augmented Transition Networks». Nyere former for generativ grammatikk er også velegnet i datalingvistisk sammenheng, og langt mer velegnet enn «Augmented Transition Networks» til å uttrykke lingvistiske generaliseringer. Særlig aktuelle i systemer som utvikles i dag, er Leksikalsk-funksjonell grammatikk, som er utviklet av lingvister og informatikere ved Stanford University og Xerox PARC, og «Generalized Phrase Structure Grammar», som er utviklet av lingvister i England. Ingen av disse grammatikkmodellene bruker transformasjoner. De er ikke bare datalingvistiske instrumenter, men også høyst interessante som lingvistiske teorier. Den utviklingen Chomskyansk transformasjonsgrammatikk har gjennomgått i 70-årene, har også gjort *den* mer interessant i datalingvistisk sammenheng enn den var før, bl.a. fordi transformasjoner har fått en langt mer begrenset rolle å spille også der.

Den datalingvistiske forskningen har på sin side påvirket utviklingen innenfor teoretisk lingvistikk. Jeg nevnte at et språkanalysesystem i tillegg til representasjoner av grammatisk kunnskap i form av generative regler også må inneholde en analysealgoritme som benytter denne «kunnskapen». Forholdet mellom grammatisk kunnskapsrepresentasjon og analysealgoritme er etter hvert også blitt et viktig grunnforskningsfelt innenfor teoretisk lingvistikk. Mens lingvistikken tidligere var mest opptatt av språkstruktur, eller «competence», er feltet utvidet til også å omfatte de prosessene strukturene inngår i, eller språklig «performance», om man vil. Datalingvistikken gjør det mulig å underkaste sider av «performance» den samme form for formalt studium som «competence» hittil har vært gjenstand for. På denne måten blir datamaskinen en modell, et paradigme for utforskningen av menneskelige språkprosesser. Dermed er vi inne på domenene «kunstig intelligens» og «cognitive science», der utforskning av språkevnen alltid har stått sentralt. Når grammatikkmodeller og prosessmodeller slik utforskes i sammenheng, kan de virke gjensidig motiverende på hverandre. Berwick og Weinberg formulerer det slik i en bok fra i år (Berwick og Weinberg: *The Grammatical Basis of Linguistic Performance,* The MIT Press 1984, s. xiv.): «... the joint work of linguistics and computer science is like the partnership between data structures and the algorithms that use them. Linguistics is that subdiscipline of cognitive science dealing with certain structures of knowledge; computer science

tells us which algorithms work best with certain data structures. Research on either side is mutually constraining.» Selv om man ikke skulle dele forfatternes syn på lingvistikken som en underdisiplin under «cognitive science», kan man akseptere dette som et meningsfylt og spennende forskningsprogram. Og også for lingvister som ikke er opptatt av prosessmodeller kan det være verdifullt å kunne implementere komplekse grammatiske teorier på en datamaskin for slik å kunne teste deres konsekvenser.

Jeg har nå omtalt en type lingvistisk grunnforskning der datamaskinen så å si anvendes som en modell av studieobjektet, og det kan være naturlig å relatere det til et skille som stundom trekkes mellom «datalingvistikk» og «datastøttet lingvistikk». Er det datastøttet lingvistikk jeg har beskrevet her? I den grad vi vil trekke et slikt skille, er det



Tegning: Øystein Reigem.

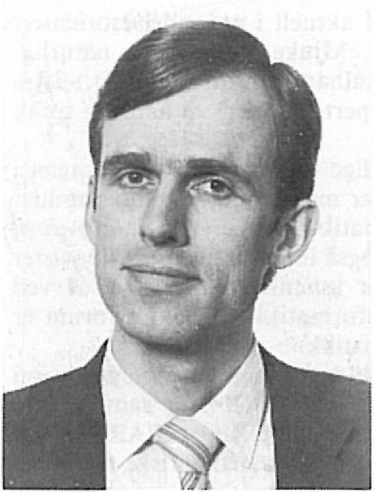
etter min mening ikke fruktbart å trekke det her. Også på det sistnevnte området er det tale om systemer som *simulerer språklig kompetanse*, altså systemer som bare kan utvikles ved hjelp av lingvistisk ekspertise: De forutsetter en lingvist som *leverandør* av innsikt. Så lenge det er tilfellet, er det tale om datalingvistikk. Datastøttet lingvistikk, derimot, er løsning av tradisjonelle lingvistiske forskningsoppgaver ved hjelp av ordinære databehandlingsteknikker som f.eks. sorteringsprogrammer, statistikkutregning osv. Eksempler er utarbeidelse av konkordanser og indekser, statistisk prosessering av ordforekomster i tekster eller av sosiolingvistisk materiale osv., oppbygging av tekstarkiver, med mere. Ved denne typen virksomhet er lingvisten hovedsakelig *mottager* av ekspertise. Noe forenklet kunne man si det slik: Datastøttet lingvistikk handler om praktiske teknikker som er nyttige for teoretikere; datalingvistikk handler om teori som er nyttig for praktikere.

Likevel er det i praksis nær sammenheng mellom de to områdene, og konkrete prosjekter vil ofte benytte seg av teknikker fra begge. Skillet er av begrepsmessig art, og poenget med å trekke det må være at datalingvistikk og datastøttet lingvistikk representerer ulike vektlegginger av problemstillinger, ulike siktemål, og at det i dag kan være viktig å understreke at vi bør satse sterkere på spesifikt datalingvistisk forskning. Datastøttet lingvistikk har vi allerede, akkurat som vi har datastøttet litteraturforskning, datastøttet arkeologi osv.

Når vi bør satse sterkere på datalingvistikk, er det fordi datalingvistikk er blitt et viktig grunnforskningsfelt internasjonalt, med innflytelse på vår forståelse av språk og språkbruk, og fordi det er nødvendig å bygge opp en kompetanse i Norge for at det skal bli mulig å utvikle og tilpasse systemer som skal kunne analysere norsk språk. Men ikke bare det. Jeg er så dristig å mene at det også vil være et verdifullt tilskudd til forståelsen av databehandlingens teori. Som Terry Winograd uttrykker det i en artikkel i Scientific American for september i år: «In the popular mythology the computer is a mathematics machine: it is designed to do numerical calculations. Yet it is really a language machine: its fundamental power lies in its ability to manipulate linguistic tokens - symbols to which meaning has been assigned.» Den samme innsikten ligger bak opprettelsen av «Center for the Study of Language and Information» ved Stanford University i fjor - et senter der forskere fra Stanford, SRI International og Xerox PARC samarbeider. Senterets program bygger bl.a. på følgende forutsetninger: «(1) Language use is fundamentally computational in that it is used by finite agents with limited resources to process, store, and communicate information; (2) computational practice is fundamentally linguistic in that computers are used by humans under the assumption that the symbols and processes of computers are about entities in the world».

Hvis man mener at man er opptatt av «humaniora i informasjonssamfunnet», bør man følgelig være opptatt av datalingvistikk. «Humaniora i informasjonssamfunnet» bør ikke bare være et spørsmål om å

stå på sidelinjen og være bekymret for at folk leser færre bøker enn før. Som humanister bør vi tilkjenne oss selv en mer positiv og offensiv rolle som forvaltere og utviklere av en relevant og meget nyttig form for ekspertise: Ekspertise i menneskelig symbolbehandling. Og vi må være villige til å utforske dette problemområdet innenfor forståelsesrammer som muliggjør tilstrekkelig presise teorier til at de kan danne grunnlag for simulering - vi må med andre ord kvitte oss med en del vanlige filolog-fordommer. Men det er kanskje en annen artikkel.



Helge J. Jakhelln Dyvik er professor ved Institutt for fonetikk og lingvistikk, Universitetet i Bergen. I samarbeid med Senteret deltar Dyvik i et datalingvistisk forskningsprosjekt hvor det særlig arbeides med problemer knyttet til forholdet mellom syntaktisk og semantisk analyse. Han arbeider også med utviklingen av et grunnfag i datalingvistikk ved UiB.

Datalingvistikk i Norge

Helge Lødrup

I USA og enkelte andre land er datalingvistikk et veletablert forskningsområde med en historie som strekker seg over et par dekader. I Norge er det et nyere område, men det er klart på frammarsj.

I tekniske og naturvitenskapelige miljøer har det en tid vært interesse for datalingvistikk.

Ved NTH/SINTEF er naturlig språk aktuelt i anvendelsesorientert forskning. Det har vært arbeidet med «Mjuke system» for naturlig-språklig dialog med edb-systemer. (Se Stålhane og Amble (1982).) Også i arbeidet med kunnskapsteknologi (ekspertsystemer) vil naturlig språk spille en viktig rolle.

Ved Det matematisk-naturvitenskapelige fakultet ved Universitetet i Oslo er det et aktivt miljø som arbeider med emner i grenseområdet mellom logikk, lingvistikk og informatikk. Lederen er *Jens Erik Fenstad* ved Matematisk institutt, men også informatikere og lingvister er med. Det er skrevet hovedoppgaver innenfor området både ved Matematisk institutt og Institutt for informatikk. Et viktig forum er Matematisk institutts «Seminar i lingvistikk».

Et forskningsprosjekt som pågår, gjelder semantisk interpretasjon innenfor leksikalsk-funksjonell grammatikk, LFG. Her samarbeider man med forskere ved Stanford University og Xerox PARC om å formalisere sider ved Barwise og Perrys situasjonssemantiske teori. (Se Fenstad et al (1984).)

I de språkvitenskapelige miljøene var *Svein Lie* ved Universitetet i Oslo først ute. Han laget en parser for norsk, basert på klassisk funksjonsanalyse, i samarbeid med *Knut Hofland* ved NAVFs EDB-senter for humanistisk forskning. (Se Lie (1980).)

I det siste har interessen for datalingvistikk vært sterkt økende blant språkforskere. Det er rimelig å se dette i sammenheng med utviklingen innenfor formell syntaks og semantikk i internasjonal språkvitenskap de senere årene, og med den økende interessen for dette i Norge. De sentrale modellene i internasjonal språkvitenskap, Government-Binding, leksikalsk-funksjonell grammatikk og generalisert frasestrukturgrammatikk, er alle interessante fra et datalingvistisk synspunkt. I språkvitenskapelige miljøer vil arbeidet med datalingvistikk bli en naturlig utvidelse av arbeidet med formell syntaks og semantikk.

Ved universitetene i Tromsø, Trondheim og Bergen har lingvistene tatt initiativer i datalingvistisk retning.

I Tromsø er det søkt om en stipendiatstilling i datalingvistikk, og muligheten for å tilby et mellomfag er under utredning.

I Trondheim er det tilløp til et samarbeid mellom lingvister og

forskere på NTH/SINTEF, i første omgang i form av «Kunnskapsteknologisk forum». Det er søkt om et professorat i datalingvistikk.

I Bergen er det vedtatt å bygge ut datalingvistikk som nytt fagfelt ved Det historisk-filosofiske fakultet. Det er ansatt en universitetsstipendiat, og fakultetet prioriterer et professorat i emnet. Virksomheten er tilknyttet Institutt for fonetikk og lingvistikk, som her samarbeider med andre språkmiljøer, bl.a. Prosjekt for datamaskinell språkbehandling ved Nordisk institutt, og med NAVFs EDB-senter for humanistisk forskning. Ved Institutt for fonetikk og lingvistikk kan man ta datalingvistikk som mellomfagstillegg i lingvistikk, og det vil snart bli tilbudt som grunnfag.

Som i Oslo arbeider man med datalingvistikk innenfor rammen av leksikalsk-funksjonell grammatikk. Grunnlaget ble lagt av *Per-Kristian Halvorsen*, som begynte arbeidet med en LFG-parser for norsk da han var forsker ved NAVFs EDB-senter. (Se Dyvik og Hofland (1983).) Forskningen er i første omgang knyttet til automatisk syntaktisk og semantisk analyse av norsk innenfor denne modellen. Og også i Bergen er det interesse for å legge situasjonssemantikk til grunn for den semantiske analysen.

Henvisninger

Dyvik, H. og K. Hofland (1983) Parsing basert på LFG: Et MIT/Xerox-system applisert på norsk. Foredrag ved De nordiske datalingvistikkdagene 1983, Uppsala. (Under utgivelse)

Fenstad, J.E. et al (1984) Equations, Schemata and Situations. Upublisert manuskript.

Lie, S. (1980) *Automatisk syntaktisk analyse.* Bergen. NAVFs EDB-senter for humanistisk forskning.

Stålhane, T. og T. Amble (1982) *Soft Systems.* Trondheim. RUNIT.

Cand. philol. Helge Lødrup er universitetsstipendiat ved Institutt for fonetikk og lingvistikk, Universitetet i Bergen.

Finländsk datalingvistik

Fred Karlsson

Datalingvistikens anglocentrism

Den datalingvistiska teorin har, speciellt inom området parsing, en klar anglocentrisk slagsida. De teorier och modeller för automatisk ordigenkänning och satslösning som hittills utarbetats är i allmänhet inte anpassade för att analysera strukturer av den art som finns i (mer eller mindre) syntetiska och agglutinerande språk såsom finskan och ungerskan.

Sådana drag är t ex den rikliga förekomsten av ordformer (varianter av lexemen), en intrikat böjnings- och avledningsmorfologi och en relativt fri ordföljd. T o m Winograd (1983) ger uttryck för uppfattningen att morfologiska fenomen i allmänhet är så språkspecifika och idiosynkratiska att det inte finns goda möjligheter att utveckla en allmän datalingvistisk morfologisk teori.

Också i praktiken, dvs inom data- og den grafiska industrin, har det flera gånger visat sig att program och systemlösningar som ursprungligen gjorts för engelskan och liknande indoeuropeiska språk i allmänhet inte går att konvertera för tillämpning på finska utan grundläggande modifikationer. Detta gäller bl a informationssökningssystem och rutiner i ordbehandlingsprogram såsom avstavning och (partiell halvautomatisk) korrekturläsning.

Forskningsbehovet

Det finns således brådskande både teoretiska och praktiska motiv för datalingvistisk utforskning av utomindoeuropeiska språk. Det centrala är självfallet de teoretiska bidrag en sådan forskning kan komma med.

Vid Institutionen för allmän språkvetenskap vid Helsingfors universitet har under perioden 1981-1984 pågått ett av Finlands Akademi bekostat projekt med titeln «Automatisk analys av finska». Det primära målet har varit att för morfologins och syntaxens del utreda vilka teoretiska krav en allmängiltig morfologisk respektive syntaktisk (eller snarare: integrerad morfosyntaktisk) parser bör uppfylla för att kunna analysera finska lika väl som engelska eller svenska. En bakgrund för detta projekt var den morfologiska pilotundersökning som Brodda & Karlsson (1980) utförde med Broddas BETA-program.

Koskenniemis morfologiska tvånivåmodell

Projektets centrala resultat hittills är Kimmo Koskenniemis (1983)

språkoberoende tvånivåmodell för morfologisk analys och syntes.

Tvånivåmodellen tillhandahåller dels en allmän formalism för beskrivning av morfologiska fenomen såsom affigering och morfofonologiska växlingar, dels en oberoende algoritm och ett datorprogram som implementerar tvånivåbeskrivningen av ett givet språk. Modellen är i princip obegränsad till sin täckning och är avsedd för tillämpning dels på hela språksystemet, dels på löpande text.

Tvånivåmodellen arbetar med parallella regler, inte sekventiellt ordnade regler som i den generativa (morfo)fonologin. Detta har den interessanta konsekvensen att de ontologiskt onaturliga mellanstadierna i en ordinär generativ derivationshistoria försvinner.

Tvånivåmodellen uttrycker, enkelt sagt, en tillåten korrespondens mellan den lexikala nivån och ytnivån. Den lexikala nivån består av lexem i kombination med sin potentiella uppsättning av morfotaktiska mönster. Alla stammar och ändelser är alltså försedda med uppgifter om vilka morfotaktiska mönster som kan tänkas komma härnäst. De morfotaktiska mönstren är implementerade som minilexikon bundna till varandra med sekventiella referenser. Den lexikala nivån tillåter också användning av morfofonem och morfologiska drag. Ytnivån består helt enkelt av graford.

De morfofonologiska reglerna uttrycks med den nämnda speciellt konstruerade lingvistiska formalismen, men implementeras var och en som en självständig finite state -automat. En tillåten korrespondens har upptäckts om vid parvis genomgång av tecknen i grafordet och lexikonträdet alla automater vid strängarnas slut är i tillåtna sluttillstånd.

Detta innebär bl a, att generering och igenkänning av former kan utföras med samma regeluppsättning och av samma program. Programmet kan, på et konkret sätt, köras i båda riktningarna. Vid igenkänning återfinns alla tolkningar om ytsträngen är flertydig. Sammansatta ord igenkänns produktivt: det enklaste fallet är att rotlexikonet pekar på sig självt som en morfotaktiskt möjlig fortsättning.

Koskenniemi (1983) har implementerat en fullständig beskrivning av finskans böjningsmorfologi. Karlsson (1983; i tryck) har utformat en beskrivning av finskans avledningsmorfologi som också implementerats i tvånivåmodellen. Därmed igenkänns också alla produktivt avledda ord.

Lexikonet i den finska implementationen omfattar f n cirka 10.000 lexem, de vanligaste orden i finsk normalprosa tagna från toppen av en frekvensordbok (Saukkonen & al. 1979). Sammantagna och kopplade till de nämnda morfologiska beskrivningarna innebär detta att c 90% av orden i vilken som helst text på normalprosa kan tolkas morfologiskt.

Tvånivåmodellen har tillämpats också på flera andra språk. En fullständig implementering av svenskans böjningsmorfologi har gjorts av Blåberg (1983) och en implementering av fornkyrkoslaviskans morfologi av Lindstedt (i tryck). Blåbergs svenska lexikon omfattar i

hakkaile (a)
hakkailutta
hakkautta
hakkaise
hakkautu
hakkaantu

hakkaaminen
hakkaaja (b)
hakkaajuus
hakkaajatar
hakkaajattaruus
hakkaajamainen
hakkaajamaisuus
hakkaajatarmainen
hakkaajatarmaisuus
hakkaajamaisempi
hakkaajamaisemmuus
hakkaajamaisin
hakkaajamaisimmuus
hakkaava
hakkaavuus
hakkaavampi
hakkaavammuus
hakkaavin
hakkaavimmuus
hakannut
hakanneisuus
hakanneempi
hakanneemmuus
hakannein
hakanneimmuus
hakattava
hakattavuus
hakattavampi
hakattavammuus
hakattavin
hakattavimmuus
hakattu
hakattuus
hakatumpi
hakatummuus
hakatuin
hakatuimmuus
hakkaamaton
hakkaamattomuus
hakkaamattomampi
hakkaamattomammuus
hakkaamattomin
hakkaamattomimmuus
hakkaavainen
hakkaavaisuus
hakkaavaisempi
hakkaavaisemmuus
hakkaavaisin
hakkaavaisimmuus
hakkaus
hakkauksellinen
hakkauksellisuus
hakkauksellisempi
hakkauksellisemmuus
hakkauksellisin
hakkauksellisimmuus

hakkaileminen
hakkailija (c)
hakkailijuus
hakkailijamainen
hakkailijamaisuus
hakkaileva
hakkailevuus
hakkaillut
hakkailleisuus
hakkailtava
hakkailtavuus
hakkailtu
hakkailematon
hakkailemattomuus
hakkailevainen
hakkailevaisuus
hakkailu

hakkailuttaminen
hakkailuttaja (d)
hakkailuttajuus
hakkailuttajamainen
hakkailuttajamaisuus
hakkailuttava
hakkailuttavuus
hakkailuttanut
hakkailuttaneisuus
hakkailutettava
hakkailutettavuus
hakkailutettu
hakkailuttamaton
hakkailuttamattomuus
hakkailuttavainen
hakkailuttavaisuus

hakkauttaminen
hakkauttaja (e)
hakkauttajuus
hakkauttajamainen
hakkauttajamaisuus
hakkauttava
hakkauttavuus
hakkauttanut
hakkauttaneisuus
hakkautettava
hakkautettavuus
hakkautettu
hakkauttamaton
hakkauttamattomuus
hakkauttavainen
hakkauttavaisuus

hakkaiseminen
hakkaisija (f)
hakkaisijuus
hakkaisijamainen
hakkaisijamaisuus
hakkaiseva
hakkaisevuus
hakkaissut
hakkaisseisuus
hakkaistava
hakkaistavuus
hakkaistu
hakkaisematon
hakkaisemattomuus
hakkaisevainen
hakkaisevaisuus
hakkaisu

hakkautuminen
hakkautuja (g)
hakkautujuus
hakkautujamainen
hakkautujamaisuus
hakkautuva
hakkautuvuus
hakkautunut
hakkautuneisuus
hakkauduttava
hakkauduttavuus
hakkauduttu
hakkautumaton
hakkautumattomuus
hakkautuvainen
hakkautuvaisuus

hakkaantuminen
hakkaantuja (h)
hakkaantujuus
hakkaantujamainen
hakkaantujamaisuus
hakkaantuva
hakkaantuvuus
hakkaantunut
hakkaantuneisuus
hakkaannuttava
hakkaannuttavuus
hakkaannuttu
hakkaantumaton
hakkaantumattomuus
hakkaantuvainen
hakkaantuvaisuus

(a) = root → der. V
(b) = root → der. N, A
(c) = hakka/ile → der. N, A
(d) = hakka/il/utta → der. N, A
(e) = hakka/utta → der. N, A
(f) = hakka/ise → der. N, A
(g) = hakka/utu → der. N, A
(h) = hakka/antu → der. N, A

Alle avledninger av det finske verbet «hakkaa» (hakke).

skrivande stund cirka 3000 lexem. Implementationer för engelska, franska, japanska och rumänska finns beskrivna i sammelvolymen *Texas Linguistic Forum 22.*

Halvautomatisk tagging

Den morfologiska fasen av projektet är till sina väsentliga delar avslutad, den syntaktiska håller på att begynna. En förutsättning för avancerad parsing är tillgång till taggade korpora av nödig storlek med hjälp av vilka man snabbt och tillförlitligt kan testa hypoteser och optimera regler.

För detta ändamål har jag med hjälp av *Jouko Lindstedt* konstruerat ett morfologisk orienterat halvautomatiskt taggande program, FINTAG, omfattande tretton moduler av BETA-regler som tillämpas i sekvens på inputtexten. Output är inputtexten så analyserad att (a) alla graford försetts med en ordklasstagg, och (b) alla ändelser i graforden är segmenterade. Ordklasstaggarna är «intaggade» (jfr Brodda 1982), närmare bestämt prefigerade, så att resultatet ser ut som följer: PR:TÄMÄ=N N:VUODE=N N:ALKU VF:ON A:KYLMÄ=Ä N:AIKA=A «detta års början är en kall tid».

De tretton BETA-modulerna (som omfattar cirka 7000 rader substitutionsregler) har ordnats i sekvens både på lingvistiska och strategiska grunder. Taggningsstrategin kan kort karakteriseras så här:

- prefigera tecknet + till ett ord så snart det erhållit sin ordklasstagg och alla ändelser segmenterats; följande regelmoduler stiger inte in i ord prefigerade med +
- de 200 vanligaste ordformerna taggas och segmenteras som helheter
- de 600 vanligaste adverben taggas och segmenteras som helheter (förutsatt att de är strukturellt homonyma med nominala böjningsformer)
- alla böjningsändelser och många avledningsändelser segmenteras utspridda enligt noggranna överväganden över de tretton modulerna
- närhelst det är möjligt prediceras ordklasstillhörigheten utgående från igenkända grammatiska former (i samma graford)
- frekvensbaserade stamlexikon används för att förutsäga slutna ordklasser samt adjektiv och verb
- former utan ordklasstagg vid tillämpning av den sista regelmodulen klassificeras som substantiv.

De första versionerna av FINTAG gav en helt korrekt analys åt cirka 85% av graforden i en text på 66.000 ord. Senare korrigeringar har höjt träffsäkerheten till något över 90%. De återstående 10% av graforden blir felaktigt analyserade och måste givetvis korrigeras för hand.

Det sålunda taggade materialet på 66.000 ord har senare för hand

sönderdelats i sina enkla satser (till antalet drygt 10.000) som kan utgöra poster vid direkta syntaktiska sökningar. Av denna korpus har framställts olika varianter, t ex en där varje sats representeras enbart av sina taggar medan det lexikala materialet filtrerats bort.

Denna taggade korpus utgör grunden med vars hjälp de första ansatserna till en (morfosyntaktisk) parser för finska gjorts.

En parser

En parser för ett språk av finskans typ möter andra slags problem och kan utnyttja annan sorts information än en parser för den indoeuropeiska språktypen. Klart är t ex, att ytstrukturen är mera informationsmättad pga den rikliga förekomsten av overta böjningsmorfem som mer eller mindre direkt signalerar syntaktiska funktioner som subjekt, objekt och attribut. Å andra sidan är ordföljden mycket friare, speciellt på sats- och meningsnivåerna, vilket medför komplikationer.

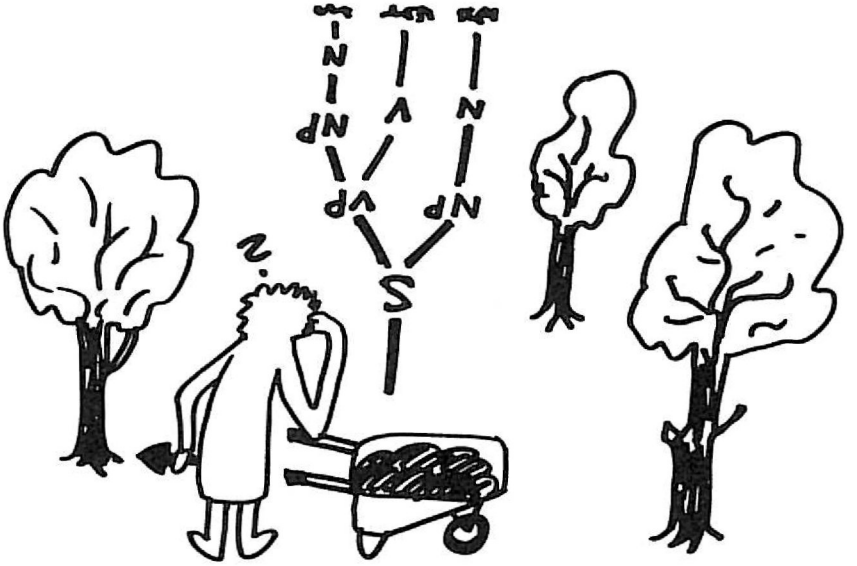
Den parser vi avser att konstruera skall i sin slutliga form stå på en tvånivåmorfologi av ovan beskriven form. Den morfologiska igenkänningen ger en fullständig analys som input till syntaxen.

Det är som bekant en öppen fråga vilken grundläggande formalism en syntaktisk parser bör uppbyggas kring. Som Koskenniemi påvisat, kan morfofonologin beskrivas i termer av finite state -automater. För syntaxens del är det klart att så enkla medel inte räcker till att lösa mera komplicerade konstruktioner.

För att tentativt nalkas detta problem har jag konstruerat och i Lisp implementerat en «realistisk» parser för finskans grundläggande satsstrukturer. Parsern gör någotsånär klassisk satslösning, dvs igenkänner statsdelar som subjekt, objekt osv samt kopplar ihop alla attribut med sina huvudord. Detta kan betraktas som ett minimum av vad en parser skall klara av.

«Realistisk» innebär bl a, att satsanalysen görs under ett enda pass från vänster til höger, att det inte används längre look-ahead än ett ord (dvs ordet omedelbart efter det som vid ett givet ögenblick analyseras), att besluten fattas deterministiskt så snart säkerhet föreligger, att back-tracking inte tillåts, och att den syntaktiska tolkningen skall byggas upp inkrementalt så att beslut vid givna ställen i strängen skall beakta endast sådana alternativ som är möjliga *givet* den analys som dittills alstrats. Parserns verksamhet bör således bl a återge det faktum, att det i början av satsens analys normalt står flera alternativ till buds än mot satsens slut.

Ifall ett givet beslut inte kan fattas givet lokal morfologisk information om ordet ifråga plus den dittills alstrade tolkningen, placeras ordet (med eventuellt identifierade attribut) i en väntebuffert. Tolkningen av den syntaktiska funktionen för innehållet i en väntebuffert avgörs genast då den fortgående analysen inkrementalt har möjliggjort en (säker) tolkning.



Tegning: Øystein Reigem.

Ett intressant resultat av denna modell är att det för finskans del vid analys av enkla satser inte tycks behövas mer än två buffertar samtidigt. Det finns mao inte samtidigt mer än två huvudord vars syntaktiska tolkning skulle vara oklar. Oftast är det i sådana fall fråga om satskonstruktioner där det finita verbet föregås av flera nominalfraser och strukturen upplöses först då finitet passerats.

Denna parser är inte generell i den bemärkelsen att programkod och språkbeskrivning vore separerade. Däremot har jag strävat till att de enskilda Lispfunktionerna skulle vara så isomorfa som möjligt med de regler som lingvisten postulerar i sin beskrivning. Det är alltså fråga om en procedurell grammatik. Ett intressant projekt vore att se i vilken mån den inkrementala tolkningsstrategin, determinismen, principen om begränsad look-ahead samt systemet med två väntebuffertar vore tillämpligt t ex på svenskan eller engelskan med sina fattigare morfologier.

Detta problem kommer att utredas i den fortsättning på projektet som planerats för åren 1985-1990. Tanken är då framförallt att i detalj försöka undersöka principerna för en generell parser.

Praktiska tillämpningar

Vårt projekt har koncentrerat sig på datalingvistisk grundforskning med inriktning på täckande modeller: hela subsystem av grammatiken, stora lexikon, löpande text. Dessa ambitioner är självfallet också av praktiskt intresse.

Flera aspekter av vårt arbete visade sig på ett tidigt stadium vara praktiskt tillämpbara. Bl a har vi utgående från de teoretiska modellerna konstruerat en så gott som felfri avstavningsalgoritm för finska samt moduler som behövs i informationssökningssystem för finska. Flera sådana är redan i praktisk produktion. Härvidlag har vi haft utmärkt nytta av Broddas BETA-system. Tvånivåmodellen bjuder i sig många möjligheter till praktiska tillämpningar, t ex vid informationssökning och automatisk korrekturläsning eller ortografisk granskning.

Övriga projekt

Datalingvistiken existerar inte som disciplin vid de finländska universiteten och det finns inte annan lingvistiskt inspirerad datalingvistisk forskning än den ovan berörda.

SITRA (Fonden till åminnelse av Finlands självständighets 50-årsjubileum) bekostar ett projekt inriktat på att konstruera ett interface på finska som er planerat att användas som en modul i expertsystem. Detta arbete görs av bl a *Harri Jäppinen* och *Esa Nelimarkka* och har hittils främst varit inriktat på konstruktion av en språkspecifik morfologisk analysator för finska.

Referenser

Blåberg, Olli 1983. Svensk böjningsmorfologi. En tvånivåbeskrivning. Trebetygsavhandling vid Institutionen för allmän språkvetenskap, Helsingfors universitet.

Brodda, Benny 1982. Problems with tagging - and a solution. Nordic Journal of Linguistics 5:2, 93-116.

Brodda, Benny & Fred Karlsson 1980. An experiment with automatic morphological analysis of Finnish. PILUS 40.

Karlsson, Fred 1983. Suomen kielen äänne- ja muotorakenne. WSOY, Porvoo.

Karlsson, Fred (i tryck). «Tagging and parsing Finnish». Utkommer i sammelvolymen från Fjärde nordiska datalingvistikdagarna, Uppsala 1983.

Koskenniemi, Kimmo 1983. Two-level morphology. A general computational model for word-form recognition and production. University of Helsinki, Department of General Linguistics, Publication No. 11, Helsinki.

Lindstedt, Jouko (i tryck). «A two-level description of Old Church Slavonic Morphology». Scando-Slavica.

Saukkonen, P. & al. 1979. Suomen kielen taajuussanasto. WSOY, Porvoo.

Texas Linguistic Forum 22, eds. Mary Dalrymple & al. Department of Linguistics, University of Texas at Austin.

Winograd, Terry 1983. Language as a cognitive process. Vol. 1: Syntax. Addison-Wesley, Reading, Mass.

Fred Karlsson er professor ved Institut för allmän språkvetenskap, Helsingfors universitet.

Datalingvistlinjen i Göteborg

Lars Ahrenberg

Det är väl ingen överdrift att påstå att intresset för datalingvistik är i stigande, både bland språkforskare, språkstuderande och, vågar man kanske påstå, också i samhället i övrigt. Detta ökande intresse beror naturligtvis delvis på att språkvetenskapen i allt större utsträckning använder datorn som forskningsredskap, men i än högre grad på att utvecklingen av datorsystem för kommersiellt bruk nu nått ett läge då automatiseringen av språkliga processer ter sig både möjlig och önskvärd, t ex inom sådana områden som informationssökning, processtyrning och användarens kommunikation med systemet. I den utvecklingen måste naturligtvis människor med kunskap om språket och människan som språkvarelse - dvs. lingvister, och i synnerhet datalingvister - spela en stor roll.

Vad är då en datalingvist för en slags person? Ja, i dagens läge är det väl oftast en språkvetare som för den egna forskningen börjat använda datorer och så småningom skaffat sig kunskaper om informationsbehandling, eller kanske omvänt, en datalog med lingvistiska intressen. Någon särskild utbildning i datalingvistik, med undantag för enstaka specialkurser då tillfälle erbjudits, har dock han eller hon i regel inte bakom sig.

Framtiden lär förmodligen bli annorlunda härvidlag. Fler och fler universitet världen över ger kurser och utbildningar i datalingvistik och relaterade specialiteter som artificiell intelligens, datorstödd textbehandling eller «Natural Language Processing». American Journal of Computational Linguistics' specialbilaga (december 1983) över sådana utbildningar upptar 85 olika universitet, men är, som man påpekar, inaktuell redan då den distribueras. Vanligast är att man ger sådana kurser inom ramen för utbildningsprogram i datavetenskap, informatik eller lingvistik. I Göteborg har vi dock valt att utforma en hel utbildningslinje speciellt för ändamålet, som alltså leder till en datalingvistexamen.

Att en utbildningslinje i datalingvistik kommit till stånd i Göteborg har flera naturliga skäl. Här finns Språkdata (institutionen för språkvetenskaplig databehandling) som i många år bedrivit forskning och forskarutbildning i ämnet. Här har också länge funnits ett tvärvetenskapligt intresse för området artificiell intelligens bland lingvister, psykologer och ADB-folk, kanske tydligast manifesterat genom arrangemanget av en Nordisk forskarkurs i artificiell intelligens i Mullsjö 1982.

Linjen är således ett samarbetsprojekt som innebär samarbete inte bara mellan institutioner utan också mellan olika fakulteter. De institutioner som är engagerade är på den humanistiska sidan, lingvistik,

språkvetenskaplig databehandling och filosofi och på den matematisk-naturvetenskapliga sidan institutionen för informationsbehandling. Utbildningen räknas dock som en humanistisk utbildning och utgör bl a en av många välkomnad humanistisk inbrytning på ett område som hittills varit dominerat av tekniker och naturvetare.

Linjen omfattar fyra år och rymmer både teori och praktik. Kurserna på linjen är utformade speciellt för linjens krav men fördelar sig med ungefär 1/3 på de lingvistiska ämnena allmän språkvetenskap och fonetik, med 1/3 på datalogi/ADB och med den sista tredjedelen på språkvetenskaplig databehandling. Den sista terminen är helt inriktad mot ett examensarbete som kan innebära projekt- eller praktikarbete inom eller utom universitetet.

Med datalingvistexamen som grund är det meningen att de 28 studerande med jämn könsfördelning som nu, höstterminen 1984, påbörjat utbildningen, ska kunna finna sina platser på arbetsmarknaden om fyra år. En enkät som gjordes innan linjen startades visade på ett relativt stort intresse. Datalingvistexamen kan också utgöra grund till forskarutbildning i språkvetenskaplig databehandling eller allmän språkvetenskap, eller, med viss påbyggnad, i andra ämnen som ingår i linjen.

Linjens uppläggning är, i detalj, följande:

* Termin 1

  Programmering, 10p
  Språk och språklig kommunikation, 5p
  Datalingvistiska problem, 5p

* Termin 2

  Strukturell översikt över ett naturligt språk, 5p
  Fonetik, fonologi, grafonomi, morfologi, 5p
  Automatisk analys och syntes av tal och skrift, 5p
  Automatisk morfologisk analys och syntes, 5p

* Termin 3

  Formaliserade syntaktiska beskrivningar av naturligt språk, 10p
  Syntaktisk parsing, 5p
  Semantik, 5p

* Termin 4

  Logik och matematisk lingvistik, 10p
  Algoritmer och datatyper, 5p
  Användning av informationssystem, 5p

* Termin 5

  Formaliserade semantiska beskrivningar av naturliga språk, 5p
  Datamaskinell lexikologi, 5p
  Semantisk parsing, 5p
  Databaser och informationssökning, 5p

* Termin 6

  Pragmatik, 5p
  Programspråk och kompilatorteknik, 5p
  Utveckling av informationssystem I, 5p
  Artificiell intelligens I, 5p

* Termin 7

  Valfri språklig kurs, 5p
  Artificiell intelligens II, 5p
  Utveckling av informationssystem II, 5p
  Valda datalingvistiska uppgifter, 5p

* Termin 8

  Valfri specialisering inför examensarbetet, 10p
  Examensarbete (uppsats, projekt- eller praktikarbete), 10p

Lars Ahrenberg er lektor/linjeledare ved Institutionen för språkvetenskaplig databehandling, Göteborgs universitet. Han forsker i bl.a. datalingvistikk med hovedvekt på parsing.

Sprogbeskrivelse til flersproglig maskinoversættelse

Hanne Ruus

Et af de områder inden for datalingvistikken, der får særlig opmærksomhed for tiden, er maskinel oversættelse. Der kan allerede købes maskinoversættelsessystemer, der oversætter mellem bestemte sprogpar, hvor det ene sprog som oftest er engelsk, og mange maskinoversættelsessystemer er under udvikling (jf. Jostein H. Hauges rapport fra Tutorial on Machine Translation i Humanistiske Data 2-84).

I konstruktionen af et maskinoversættelsessystem må man først bestemme, hvilke dele af sproget der er viktige i oversættelsesprocessen. Dernæst må man beskrive disse dele, så beskrivelsen let kan bruges af det datamatiske system, der skal udføre oversættelsen.

Når man konstruerer et oversættelsessystem, der kun oversætter mellem to bestemte sprog, kan man tillade sig at bruge viden, der kontrasterer de to sprog, i alle oversættelsens faser. Når man vil lave et maskinoversættelsessystem, der ikke bare er en sammenbygning af en række systemer for to sprog, må man søge at bygge på den viden, der er fælles for alle de sprog, der skal oversættes mellem. I et ægte flersproget maskinoversættelsessystem må man sørge for at begrænse den viden, der kun gælder for to bestemte sprog, og for, at den har sin veldefinerede plads, så den ikke bliver blandet med generelle oplysninger, der bruges af alle sprogene. Tilsvarende må man bestemme, hvor i systemet man har gavn af oplysninger, der kun gælder for et sprog.

Når man oversætter en tekst på et sprog – kildesproget – til et andet sprog – målsproget –, er det et basalt krav, at de to tekster betyder det samme. Systemet må altså sørge for, at tekstens betydning ikke ændres under oversættelsesprocessen. Beskrivelsen af naturlige sprogs betydning er imidlertid en meget vanskelig opgave. Det vil enhver have erfaret, der har prøvet at diskutere betydningen af bestemte ord og vendinger med andre sprogbrugere.

For at begrænse vanskelighederne med betydningsbeskrivelsen mest muligt kan man forsøge at bestemme, præcis hvor i oversættelsesprocessen man skal have oplysninger om betydning. Man kan endvidere bygge på andre mere veludviklede dele af sprogbeskrivelsen som ordenes opbygning og bøjning (morfologi) og ordenes kombinationsmuligheder i større enheder som sætningsdele og sætninger (syntaks).

I mange nutidige maskinoversættelsessystemer deler man oversættelsesprocessen op i tre faser: den centrale fase, hvor gloserne i kildesprogsteksten udskiftes med gloser og udtryk på målsproget. Denne fase kalder man overførselsfasen (eng. transfer). Før den centrale fase har man en analysefase, hvor man analyserer den tekst, man skal oversætte

for at finde tekstens og dens sætningers sproglige struktur. Efter den centrale fase har man en syntesefase, hvor man sørger for at sætte ordene i den rigtige rækkefølge og for at bøje dem rigtigt, så de følger de regler, der gælder for målsproget.

I Eurotra-projektet, som blev introduceret i Humanistiske Data 2-82, har man valgt at udarbejde et analysemodul og et syntesemodul for hvert sprog. For hvert sprogpar har man to overførselsmoduler. For tiden er der syv officielle sprog i EF: engelsk, tysk, fransk, hollandsk, dansk, italiensk og græsk. For sproget dansk skal der da udvikles følgende moduler: et analysemodul, et syntesemodul og tolv overførselsmoduler. Med disse moduler vil det blive muligt at oversætte mellem dansk og de øvrige EF-sprog, når der samtidig bliver udviklet analyse- og syntesemoduler for de øvrige sprog.

I Eurotraprojektet er arbeidet organiseret sådan, at der findes en forskningsgruppe for hvert sprog og en central gruppe, der koordinerer arbejdet. Forskningsgrupperne for de forskellige sprog skal udarbejde et analysemodul og et syntesemodul for deres eget sprog, desuden skal de samarbejde med de andre sproggrupper om at udvikle overførselsmoduler og om at specificere, hvilken sproglig viden der skal være til stede i de strukturer, der er inddata til de forskellige overførselsmoduler og uddata fra disse.

Den overvejende del af det lingvistiske arbejde i Eurotraprojektet har hidtil drejet sig om at bestemme, hvilken lingvistisk viden det er ønskeligt at have adgang til i overførselsfasen, og om at undersøge, hvordan denne viden kan specificeres, så man dels kan sikre, at alle, der arbejder i projektet, har den samme forståelse af denne viden, dels kan beskrive denne viden på en sådan måde, at det er muligt at beregne, hvad af den der er repræsenteret i en vilkårlig tekst.

For tiden er Eurotra i en forberedende fase. Ved afslutningen af denne fase skal der foreligge en præcis beskrivelse af det programmelsystem, der skal udvikles til projektet. Programmelbeskrivelsen indeholder blandt andet en prototype, der kan udføre de vigtigste operationer og en omhyggelig beskrivelse af den formalisme, man skal bruge til at udtrykke lingvistisk viden for at få denne behandlet af programmelsystemet. Der skal endvidere foreligge en beskrivelse af forskellige slags lingvistiske strukturer, som skal beregnes, og af den lingvistiske viden, der skal udtrykkes i strukturerne.

Den vigtigste af de lingvistiske strukturer er den struktur, der er grænseflade mellem analysemodul og overførselsmoduler og mellem overførselsmoduler og syntesemodul. Som output fra analysen skal strukturen indeholde nok information om kildesprogsteksten til, at det er let at udskifte kildesprogets gloser og vendinger med målsprogets. Som output fra et overførselsmodul skal strukturen indeholde tilstrækkelig information til, at syntesemodulet kan generere målsprogsteksten.

Da man skal udarbejde mange overførselsmoduler, nemlig to for

hvert sprogpar, for tiden i alt 42, er det væsentligt, at de opgaver, der udføres i overførselsfasen, er så få og enkle som muligt. I det ideelle tilfælde indeholder et overførselsmodul bare regler for, hvordan man for hvert kildesprogsord vælger dets målsprogsoversættelse. Reglerne er formuleret, så de kræver tilstedeværelsen af bestemte typer lingvistisk viden, morfologisk, syntaktisk og semantisk, f.eks. oversættes det engelske verbum *know* til det danske verbum *vide*, hvis dets objekt har form af en sætning eller en infinitivkonstruktion, mens *know* oversættes til *kende*, hvis dets objekt har form af et nominalsyntagme jf. *hun kender deres opholdssted* vs. *hun ved, hvor de opholder sig.*

Der findes imidlertid klasser af ord, som man på forhånd ved vil være vanskelige at behandle med enkle overførselsregler. De drejer sig om artikler, hjælpeverber og præpositioner. Disse ordklasser skal somme tider oversættes til bøjningsendelser, somme tider til ord. Tilsvarende skal bøjningsendelser somme tider oversættes til hele ord. F.eks. svarer en engelsk konstruktion med hjælpeverbet *will* ofte til en konstruktion uden hjælpeverbum på dansk jf. eng. *they will arrive tomorrow* da. *de kommer i morgen.*

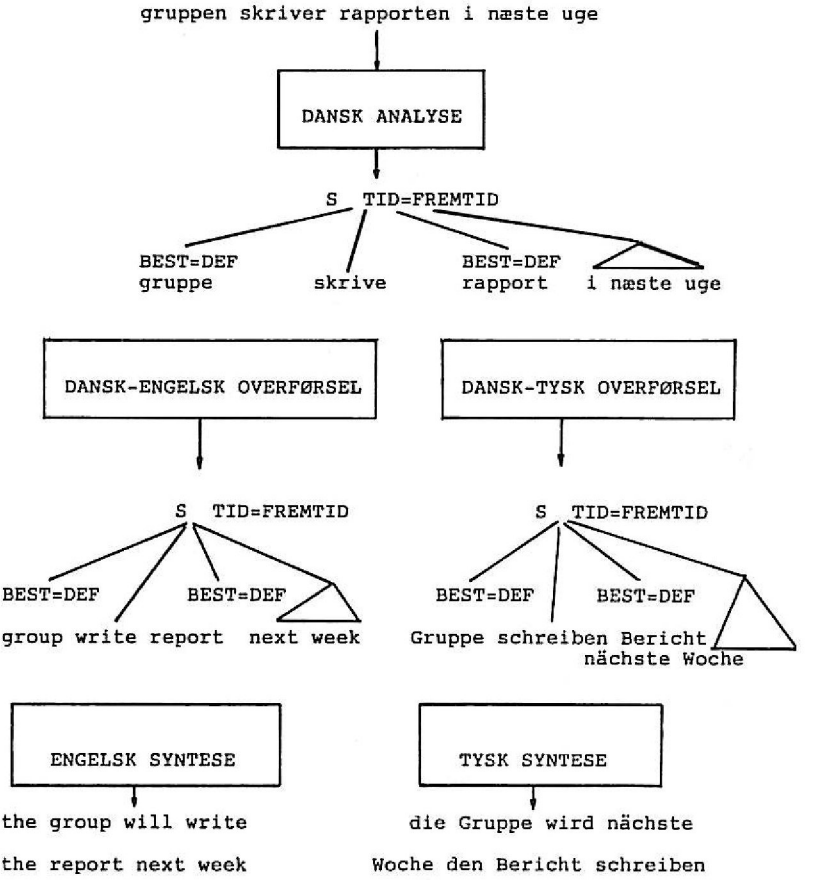
Artikler bruges blandt andet til at udtrykke bekendthed, hjælpeverber til at udtrykke tid og måde. For at undgå indviklede overførselsregler søger man da at finde fælles værdier for tid, måde og bekendthed. Disse værdier kan så beregnes i analysemodulerne og bruges til at generere det rigtige målsprogsudtryk i syntesemodulerne. Hvis man f.eks. har et verbum bøjet i le future i den franske kildetekst, skal der beregnes en værdi FREMTID af det franske analysemodul. Denne værdi skal så bruges af det danske syntesemodul til at beregne, om der skal stå hjælpeverbet *vil* i den danske tekst, eller om den danske tekst skal have verbet i præsens. Tilsvarende skal det danske analysemodul kunne bestemme, om en præsensform skal give anledning til værdien FREMTID. Det skal også for forekomster af verbet *ville* kunne bestemme, om verbet betegner fremtid eller vilje eller begge dele.

Ideerne til områder, der bør behandles med fællessprogligе værdier, får man ved at sammenligne sprogenes grammatikker og undersøge, om de forskelligheder, man finder, giver anledning til indviklede overførselsregler. På denne måde kan man f.eks. ud fra en sammenligning af danske passivkonstruktioner med deres oversættelser til engelsk, tysk og fransk se, at passiv bør behandles fællessprogligt: dansk kan udtrykke passiv med en bøjningsendelse eller et hjelpeverbum, mens engelsk og tysk kræver et hjælpeverbum og fransk ofte bruger en sætning i aktiv f.eks. da. *morgenmaden serveres fra kl. 7* vs. eng. *breakfast is served from 7 o'clock.*

Når man vil udarbejde forslag til, hvilke værdier der skal beregnes, kan man bruge forskellige fremgangsmåder. Man kan undersøge det fænomen, man behandler, på alle sprogene og søge at finde det minimale antal distinktioner, der gør det muligt at regne sig fra en konstruktion på et sprog til den rigtige, tilsvarende konstruktion på de

andre sprog. Herved bygger man udelukkende på de sprog, man behandler. Man får værdier, der er fælles for netop dette sæt af sprog. Hvis man vil inkludere flere sprog, risikerer man, at der må tilføjes flere distinktioner og nye værdier. Til gengæld har man gode chancer for, at de foreslåede værdier faktisk lader sig beregne, da man har taget udgangspunkt i forskellige konstruktioner på de sprog, man skal beregne værdier for.

Hvis man gerne vil sikre sig, at nye sprog kan tilføjes, uden at de forskellige sæt af fællessproglige værdier skal ændres, må man stile efter et sæt af universelle, semantiske værdier. Sådanne værdier kan berede vanskeligheder på to måder: Da disse værdier ikke på forhånd



Ovenstående er en forenklet skematisk fremstilling af oversættelsen af en dansk sætning til engelsk og tysk. Man ser, hvordan de fællessproglige værdier FREMTID og DEF beregnes af dansk analyse og bruges til at beregne hjælpeverber og artikler på engelsk og tysk.

hører sammen med bestemte konstruktioner i de forskellige sprog, kan de både være vanskelige at beregne og vanskelige at få en fælles forståelse af.

I Eurotra-projektet er der foreløpig foreslået universelle semantiske værdier for semantiske træk ved ordbetydninger (f.eks. HUMAN, ANIMATE) og for de semantiske relationer mellem verbet i en sætning og de andre sætningsled (f.eks. AGENT, PATIENT, EXPERIENCER), mens der findes forslag for mere pragmatisk bestemte, semantiske værdier for områderne tid, måde og bekendthed.

Arbejdet med at definere de fælles værdier for tid, måde osv. kan ikke gøres færdigt en gang for alle. Man må først lave et gennemarbejdet forslag og afprøve dette. Under afprøvningen kan der komme forskellige slags vanskeligheder: dels kan det vise sig, at en specificeret værdi er for vanskelig at beregne, dels kan det vise sig, at de værdier, man havde defineret ikke er tilstrækkelige, idet man ikke får de rigtige konstruktioner på målsprogene, selv om man bruger de definerede værdier. I begge tilfælde må man utarbejde et revideret forslag til værdier og afprøve det.

Efterhånden som man finder frem til sæt af værdier, der virker for de forskellige betydningsområder, bliver disse sæt inkluderet i den mængde oplysninger, man skal specificere i overførselsstrukturen. Man får således forøget den mængde af fællessproglig viden, der skal og kan beregnes.

I den næste fase af projektet skal der udvikles en første prototype, der kan oversætte mellem sprogene inden for et ordforråd på ca. 2500. Under konstruktionen af denne prototype vil de foreslåede fællessproglige værdier blive afprøvet. Det kan da ikke udelukkes, at pragmatisk bestemte, semantiske værdier vil erstatte universelle og omvendt. Det kan også vise sig, at det på enkelte områder ikke er muligt at finde et sæt fælles værdier, der er operationelt. Hvor det er tilfældet, må problemerne klares i de to-sprogede overførselsmoduler. Det vil altså først i løbet af udviklingsfasen vise sig, hvor mange fællessproglige værdier man kan finde og bruge. Jo flere fællessproglige områder, man kan behandle med fælles værdier, desto større udbytte vil man have af at udvikle et flersproget system.

Hanne Ruus er lektor ved Institut for nordisk filologi, Københavns Universitet. Hun arbeider for tida med EFs maskinoversettelsesprosjekt og edb-baserte undersøkelser i eldre nordiske tekster. I tillegg underviser hun i datalingvistikk.

Edb og talemålsforsking

Erfaringar frå prosjektet «Talemål hos ungdom i Bergen»

Helge Sandøy

1. Prosjektet TUB

Arbeidet i prosjektet «Talemål hos ungdom i Bergen» (TUB) kom i gang i 1977 og hadde som mål å undersøke utviklinga i bergensmålet og variasjonen i talemålet hos ungdommen i Bergen kommune (d.v.s. «storkommunen» etter 1972). NAVF finansierte prosjektet til ut 1983.

Etter at innsamlingsfasen var over, hadde prosjektet eit svært stort lydbandmateriale. Det var opptak av ymse typar samtalar med 104 informantar, i alt på 68 timar og 15 minutt. (Ei kort orientering om prosjektet står i *Talemål i Bergen 1/83*.) Om lag 75% av materialet blei transkribert fullt ut for å gjøre syntaktisk og tekstlig analyse lettare, for seinare å kunne gi ut dialekttekstar og for å halde materialet meir ope for nye problemstillingar.

Sjølv om ein avgrensa systematisk det materialet som skulle transkriberas, blei det svært store tekstmengder til saman. Det omfatta i alt 220.000 løpande ord. Spørsmålet blei derfor tatt opp om korleis ein kunne få ei mest mulig rasjonell behandling av materialet. Etter å ha hatt samrådingsmøte med datafolk vedtok prosjektet i 1980 å dataføre transkripsjonen for å kunne utnytte betre datateknikken i analysen av materialet. Ein vurderte vinningane i analysearbeidet til å vere så store at ein kunne ta kostnadene med dataføringa. Vinningane ville ligge i raskare analyse og større nøyaktigheit i oppteljinga.

2. Krav til materiale og til bruk

Transkripsjonen var fonemisk, unntatt på eit par punkt der den fonetiske variasjonen var av spesiell interesse. Ein slik transkripsjon krev ein del ekstrateikn utover skriftspråksortografien. Derfor måtte eit maskintilgjengelig teiknsett settas opp. Det blei konstruert i korrespondanse med internasjonal lydskrift (IPA) (med noen tilleggsteikn for pause, utydelig tale, svake element, og namn/ord sitert med fremmend/utalandsk uttale).

Samtalane i lydbandopptaka var dels mellom to informantar + opptaksleiar, dels mellom fire informantar + opptaksleiar. I samtaleutskriftene var det ønskelig ikkje bare å få fram replikkvekslingane, men også å ta vare på kva som var samtidig tale, d.v.s. kva tid samtaledeltakarane snakka i munnen på kvarandre. Vi valde derfor ei oppstilling med

parallelle og synkroniserte linjer, der kvar deltakar hadde si linje i linjebundelen:

```
2931  K1:
2932  L :
2933  M1:  < de: ska vå de: ska +vå:r skol6 > /              < di:    så +ve: > /       < n6 > /
2934  K2:                                                vi  < v6 > /
2935  M2:  < jo vi fik sån fik sån "tilbud > /   +ja: /      < asså   e:C6 +møt >  asså  < fik >


2936  K1:
2937  L :
2938  M1:                ja di: så +ve: /
2939  K2:
2940  M2:  "tilbud6 ja /                    +e:g ska nå +ne: å "hø:r6


2941  K1:                       når^ ska "de:r6 da /
2942  L :
2943  M1:
2944  K2:
2945: M2:  ka d6 +e: få nåk6 /                          nei kå ti +va:
```

Eksempel frå gruppesamtale nr. 60.
(Tala til venstre er linjenummer. L = leiar, K1 osv. er kvinnelige og mannlige informantar. < og > markerer samtidig tale. / er teikn for kort pause, for at siste segmentet er utydelig. Lydskrifta elles er det ingen grunn til å forklare nærare her.)

Alle samtaletekstane er utstyrt med samtalenummer og informantnummer. Desse numra saman med linjenummer er lagt inn som referansar i KWIC-konkordansen som er laga på grunnlag av alle tekstane. I referansedelen av konkordanslinja er det dessutan òg lagt inn informantopplysningar som vi bruker i analysen. Kvart eksempel vi finn i KWIC-konkordansen, kan derfor gi oss mange opplysningar: 1) lyden/ordet/uttrykket i språklig kontekst, 2) om det har vore samtidig tale eller avbrott i den teksten linja gjengir, 3) referanse tilbake til teksten slik at ein kan finne ein større samtalekontekst, 4) opplysningar om sosialgruppe, kjønn, alder og bydel for den informanten som har brukt ordet/uttrykket, 5) opptakssituasjon belegget er henta ifrå. I konkordansen er belegga sortert alfabetisk. Men p.g.a. lydskrifta måtte vi sette spesielle krav til alfabetiseringsprinsippa.

```
kås2b306 65G015A1424  (-) +å: / (-) (-) har "d^åkar "lanste:d nåk6 ste:d^a / (-) (-) (-) nei +de: ad e: nå
kås2b306 65G015A0939     +lei av / (-) < at iC6^ d6 +e: > (-) nåk6^ sån^ sæ:li d^æ:^ ne:r6^ // (-) (-) så>
---------

mås2b106 60G012C2935  (-) (-) (-) < de: ska vå de: ska +vå:r skol6 > / (-) (-) (-) < di: (-) så +ve: > /
kås2b306 65G015A0424     kåm6r +op // (-) (-) vå +va: "dåkar skole^ op6^ i "muntli^ i fjor^ / (-) va "dåkar
```

Eksempel på to konkordanslinjer med nøkkelordet (key-word) *nåk6* (= 'noe') og to med *skol6* (= 'skole'). (Kolonnane i referansefeltet til venstre viser: 1: kjønn, 2-3: bydel, 4-5: aldersgruppe, 6: sosialgruppe, 7-8: språkholdning, 10-12: opptaksnummer, 13-15: informantnummer, 16: opptakssituasjon, 17-20: linjenummer i teksten som belegget er henta ifrå. I teksten står // for lengre pause, (-) tyder mellomliggande replikk frå annen samtaledeltakar.)

Bjørn Eide har gjort det nødvendige programmeringsarbeidet til samtaleoppstillingane og tilrettelegginga for konkordansutvikling.

3. Dataføringa

Dei store kostnadene med EDB-bruken låg i den omstendelige dataføringa. Det var tale om svært vanskelige tekstar p.g.a. lydskrifta, dei hadde ei innfløkt oppstilling, og analysen var avhengig av stor nøyaktigheit.

Den handskrivne transkripsjonen blei maskinskriven og seinare overført til datamaskina ved optisk lesing (OCR). Denne framgangsmåten blei valt – i staden for pønsjing direkte på terminal – fordi det på denne tida var for få tilgjengelige terminalar, og dessutan for å kunne utnytte habil skrivehjelp utafor byen. På det tidspunktet var det heller ikkje aktuelt å bruke småmaskiner (tekstbehandlingsmaskiner) å skrive inn tekstmaterialet på. Kostnadene ved vår framgangsmåte var neppe større enn ved direkte pønsjing, men vi måtte plagas med store tekniske problem med den optiske lesaren, slik at det blei mye hefting i framdrifta.

Det gikk òg mye tid til korrekturen. Jamt over blei det tale om tre-fire korrekturar før alt var rett (i forhold til den handskrivne transkripsjonen).

Både transkripsjon (med kontrollar), maskinskriving, optisk lesing og korrektur i fleire omgangar foregikk samtidig (i nesten 1 1/2 år) og kravde mye organisering for at alt materialet skulle komme rett gjennom alle fasane. Til mye av arbeidet hadde vi timelønte assistentar, men prosjektmedarbeidarane måtte instruere, samordne og kontrollere arbeidet.

Sia kostnadssida er eit viktig punkt i vurderinga av EDB-bruken, kan det vere nyttig å ha ei viss konkretisering av arbeidstida som er gått med. Arbeidstimane for den timelønte hjelpa vi brukte til dataføringa, fordeler seg slik:

Maskinskriving:	440 t
Optisk lesing, programmering, korrekturinnføring	1200 t
Korrekturlesing:	700 t

I tillegg kjem den tida dei tilsette i prosjektet brukte til å organisere og kontrollere dette arbeidet. Seier vi at sjølve dataføringa tok nærare to årsverk, er det i alle fall ein viss peikepinn på omfanget.

Inn i reknestykket må ein sjølvsagt ta med at medarbeidarane var ukjente med EDB og terminalbruk i utgangspunktet, slik at denne tida har òg vore læretid.

4. Morfologisk merking

Ved automatisk analyse eller leiting i språktekstar volda alle homonyma problem. Ein transkripsjon i lydskrift gir i tillegg det problemet at eitt og same ord kan få svært mange former ettersom lydar lett fell bort i naturlig rask tale. Variantane eit ord kan forekomme i, er så mange at det i praksis er uråd å predikere dei. Lemmatisering (d.v.s. å gi alle orda ei «oppslagsform», f.eks. i normalortografi) kan vere ein måte å løyse denne siste vansken på. Homonymiproblemet kjem ein langt på veg unna med å merke alle orda morfologisk.

For å kunne prøve ut og forbetre EDB-opplegget lét vi den første tjuandedelen av det dataførte materialet utgjøre eit prøveprosjekt. Der prøvde vi òg ut eit nøyaktig merkesystem med både lemma og opplysning om ordklasse og bøyingsform. Etter erfaringane frå dette prosjektet kom vi til at tidsforbruket til merkinga ikkje stod i forhold til nytta vi ville ha i analysen av dei språkvariablane vi hadde planlagt å arbeide med.

I den endelige merkinga sløyfa vi lemmatiseringa og gikk over til bare å merke for ordklasse ved verb, substantiv (her også merke for felleskjønn eller nøytrum), adjektiv og pronomen. Dessutan fikk ein del ord vi allereie hadde definert som interessante i prosjektet, ein talkode som viste kva variabel og variant dei var belegg på. Dei fleste homonymiproblema blei løyst med denne merkinga. Og til dei konkrete analysane vi hadde planlagt, viste dette seg å vere nok. Om ein ønskjer det, kan nøgnare merking føyas til seinare. (Det er alt gjort for verb og substantiv frå tre bydelar.)

```
kÅs2b306 65G015A1424  (-) +å: / (-) (-) har "dˆåkar "lansteid nåk6 steːdˆa / (-) (-) (-) nei +deː ad eː nå    p  hl
kÅs2b306 65G015A0939    +lei av / (-) < at iC6ˆ d6 +eː > (-) nåk6ˆ sånˆ sæli dˆærˆ neːrèˆ // (-) (-) <d>    p  hl
---------
mÅs2b106 60G012C2935  (-) (-) (-) < deː ska vå deː ska +vårː nkn16 > / (-) (-) (-) < diː (-) så +veː > /    f
kÅs2b306 65G015A0424    kÅn6r +op // (-) (-) vå +vaː "dåkar skoleˆ op6ˆ i "muntlˆ i fjorˆ / (-) va "dåkar  f
```

Eksempel på merkte konkordanslinjer
(Til høgre for teksten står kode for ordklasse (ev. genus) og ev. variabel- og variantkode).

Den forenkla merkinga blei gjort i KWIC-konkordansen og ført rett inn på skjermterminal. Vi brukte der eit program Harald Solevåg alt hadde laga for merking av språklig materiale (Solevåg 1977). I ein alfabetisert konkordans vil forekomstar av same ord (og form) stå samla etter kvarandre, og merkeprogrammet utnyttar dette til raskare innsetting av kodane. Bjørn Eide skreiv programma som la materialfilane til rette for merkeprogrammet.

Vi reknar med at merkearbeidet tok om lag 550 timar.

5. Analysearbeidet

Den morfologisk merkte KWIC-konkordansen har vore utgangspunktet for det meste av analysearbeidet. Ymse seleksjonsprogram har plukka ut materiale frå konkordansen på grunnlag av den transkriberte forma av ordet (nøkkelordet på linja) og den morfologiske koden ytst til høgre på linja. Ved seleksjonen er heile konkordanslinjene som inneheld potensielle belegg på ein språklig variabel, ført over på eigen fil. I mange tilfelle har ein måtta bearbeide denne filen vidare, fordi ein har vilja legge inn tilleggsopplysningar, d.v.s. merke meir nøyaktig. Ved f.eks. variablane *då* og *no* viste det seg nødvendig å ta med opplysningar om syntaktisk funksjon (Myking 1983). Somtid har ein òg komme til at ein del av materialet i filen er uaktuelt å bruke som belegg på den meir nøyaktig definerte variabelen. Da har ein stroke dei linjene. Med meir nøyaktig merking i utgangspunktet, kunne ein sjølvsagt ha sloppe unna ein del av denne bearbeidinga etterpå. Men merkingsfasen hadde dermed tatt mye lengre tid, og trulig meir enn det ein ville vinne inn att seinare. Dessutan – og det er det viktigaste – er det svært vanlig at ein under analysearbeidet oppdagar at ein treng nye eller finare kategoriinndelingar av materialet. Da blir det aktuelt å finsikte materialet att. Den forenkla merkinga vår gav oss altså i mange tilfelle denne arbeidsmåten: først grovseleksjon or KWIC-konkordansen, deretter fininndeling ved bearbeiing i eigne filar.

Når ein enten skal merke filen nøgnare, eller ein skal reinske ut materiale som eigentlig ikkje er belegg på det ein vil undersøke, kan det vere nyttig å ha sortert materialet etter visse prinsipp: alfabetisk, baklengs alfabetisk (nyttig ved fonologiske eller bøyingsmorfologiske studiar), etter bydel, kjønn o.s.v. Visse delar eller problemtypar i materialet kan da samlas for seg, materialet blir meir oversiktlig og lettare å gi seg i kast med. Slike sorteringar eller redigeringar av filane kan maskina gjøre raskt.

I dei ferdige filane kan ein telje opp belegg for kvar kategori og for alle krysskombinasjonar av kriterium. I staden for å gjøre dette for hand fikk vi i somme tilfelle laga program som både sorterte og talde. Men i mange tilfelle fann vi det nyttigast og raskast å kjøre statistisk analyse direkte på materialfilane med konkordanslinjer. Kvar linje i filen inneheld alle dei nødvendige opplysningane om språklige og

utomspråklige variablar, slik at maskina kan rekne direkte frå grunnlagsmaterialet.

For noen få språklige variablar er dette ei rask og god løysing. Men skal ein samanlikne variasjonen for fleire variablar (studere samvariasjonen), kan slike filar bli både store og uoversiktlige, ettersom linjene også inneheld så mye unødvendig materiale, og derfor er uhandterlige. For slike tilfelle har vi plotta inn tala på belegg på dei ymse variantane i ein eigen fil som er organisert som eit meir vanlig skjema.

Seleksjons- og sorteringsprogramma er skrivne av Eva Møller, Bjørn Eide og Helge Sandøy. Medarbeidarane i prosjektet har sjølve skrive og kjørt statistikkprogramma ved hjelp av SPSS-pakka på Sperry 1100. Dei har måtta rekne seg ein del tid til å sette seg inn i både dei teoretiske og praktiske sidene ved dette arbeidet. Ein del prøving og feiling har det sjølvsagt blitt ettersom språkvitarar tradisjonelt får lite utdanning på dette feltet.

6. Andre bruksmåtar

Slik materialet nå ligg på filar - både oppstilt i tekstar og sortert i KWIC-konkordans - kan det brukas til fleire formål enn det som var hovudmålet med prosjektet TUB. Vi planlegg ei utgåve av tekstutdrag frå samtalane, først og fremst til undervisningsformål. Med tilgang til fotosettar som er utstyrt med lydskrift, er det enkelt å få feilfri sats, utan ny korrektur på desse innfløkte tekstane.

Vi har òg levert tekstar til ei hovudoppgåve om replikk- og samtaleanalyse. Dessutan har vi levert materiale til ymse grammatiske analysar. Belegg på grammatiske fenomen hentar ein lettast frå KWIC-konkordansen. Det er òg planlagt ein spesiell fonotaktisk undersøkelse av TUB-materialet som skal gi data til eit prosjekt om talesyntese.

7. Vurdering av EDB-bruken

7.1 Manuell mot elektronisk databehandling

Ei vurdering av EDB-bruken må gjøras i forhold til konkrete alternative arbeidsmåtar. Ettersom utgangspunktet her var at tekstane skulle transkriberas, var alternativet å bruke dei handskrivne transkripsjonsarka når ein skulle rekne belegg til analysane. Avgrensar ein seg til dei statistiske analysane i hovudprosjektet, ville kanskje slik manuell teljing koste mindre enn heile prosessen med å dataføre materialet. Men ein kan tenke seg visse ulemper med denne arbeidsmåten: For det første det banale problemet med å halde orden på tusenvis med lause ark som ein stadig måtte bla fram og tilbake i. For det andre er ikkje eit menneske like nøyaktig som maskina til å leite. Og for det tredje - og det er det viktigaste -: I det ideelle analysearbeidet har ein klåre hypotesar i utgangspunktet slik at kategoriane er veldefinerte, og teljinga kan gå beint fram med å skrive f.eks. strekar i eit skjema. I praktisk analysear-

beid er det ofte slik at ein etter ei stund kjem i tvil om kategoriane er rette. Det kan komme på tale å omdefinere både variablane og variantkategoriane. Denne revurderinga av hypotesar og variablar skjer når ein sit med materialet framfor seg. I slike tilfelle er det ein umistelig fordel å ha belegga samla i velredigerte filar (eller datautskrifter) – i staden for å ha dei spreidd på dei fleire tusen transkripsjonsarka. (I verste fall har ein heller ikkje notert referansar gode nok til å finne tilbake til belegget og konteksten.)

Her ser ein altså tydelig at om EDB-bruken er kostbar, gir han svært mange praktiske fordelar.

7.2. Transkripsjon mot ekserpering direkte frå band

Kan ein tenke seg alternative løysingar enda lenger tilbake i arbeidsprosessen i prosjektet, er ein aktuell utveg å ekserpere belegg rett frå lydbandopptaka i staden for å transkribere tekstane. I dei delane av TUB-materialet som ikkje var transkribert, ekserperte vi belegg på ein del variablar (der det transkriberte materialet gav for få belegg). Ekserperinga skjedde ved at ein lytta igjennom bandopptaka og skreiv ned alle belegg på ein variabel. Dessutan noterte ein òg ein kontekst på eitt ord både føre og etter ordet og ein referanse til teljeverket på avspelaren. Belegg og kontekst blei notert (i lydskrift) både for å lette kontrollen og for å ha eit visst grunnlagsmateriale om ein trong å revurdere variabel- og variantdefinisjonane. Desse to muligheitene ville ein miste om ein bare talde belegg – noe som sjølvsagt ville gå raskare.

Vi har prøvd å undersøke kor nøyaktig slik ekserpering er kvantitativt, og har komme til at med éi førstelytting og éi kontrollytting (av ein annen person) greier ein å få med seg mellom 95 og 100% av faktisk eksisterande belegg (Lødrup 1982). Det er i praksis fullt tilfredsstillande, så lenge ein kan rekne dei ikkje oppskrivne belegga som tilfeldige.

EDB-bruken ved denne ekserperingsmetoden gjeld bare den statistiske behandlinga. Ein må legge tala på belegg inn i ein fil, som ein bruker til statistikkprogramma. I visse tilfelle kan det sjølvsagt òg vere mest praktisk å skrive inn sjølve orda som er belegg.

Fordelane med denne arbeidsmåten kan vere fleire: Først og fremst sparer ein mye tid (d.v.s. kostnader). For det andre kan beleggmaterialet vere kvalitativt meir nøyaktig enn eit transkripsjonsmateriale av di ekserperaren konsentrerer seg bare om det som er relevant for variabelen han arbeider med. Ein transkriptør kan ikkje konsentrere seg om alle dei mulige problemstillingane materialet skal settas inn i.

Men vi har òg registrert ulemper med denne arbeidsmåten: I det praktiske analysearbeidet blei den noterte konteksten for liten når variabelen måtte revurderas. Men dersom ein skulle skrive ned mye kontekst og ekserpere belegg til svært mange variablar, ville tida ein sparte i forhold til full transkripsjon, krype fort saman. Dessutan er det

ein viktig forskjell at det ekserperte materialet stort sett bare kan brukas til den eine variabelen det er tiltenkt, mens ein fullstendig transkripsjon kan ha eit stort bruksområde. Han kan brukas både til problemstillingar ein planlegg alt i utgangspunktet, og slike ein kjem på seinare. (Største avgrensinga gjeld fonologiske språkvariablar, som blir låst til det transkripsjonssystemet ein valde i utgangspunktet.)

7.3. Analysen må vente lenge

Det rasjonelle med den måten TUB brukte EDB på, ligg i rask og nøyaktig utnytting av materialet når det først er dataført. Den fordelen veks med størrelsen på materialet. Men skal ein utnytte denne fordelen, forutset det at arbeidsfasane er skilt: først innsamling, deretter transkripsjon og dataføring, og til slutt analysearbeid. At analysen må vente heilt til den mødesame transkripsjons- og dataføringsfasen er avslutta, er den største ulempa med opplegget TUB valde. Det førte til ein lang og slitande arbeidsbolk der hypotesane bare «låg på hylla og venta». I denne tida fikk ein ikkje oppleve noe fruktbart møte mellom hypotesar og data/resultat. Arbeidet gav ein få utfordringar ein kunne utvikle seg faglig på. Og dessutan var det lite rom for revurderingar av prosjektet ettersom så mye var låst frå starten av, p.g.a. den store materialmengda.

Om TUB-opplegget bør brukas om att i nye prosjekt, må sjølvsagt vere avhengig av formålet med prosjektet. Eg ser dei mange fordelane i det store materialet vi nå har tilgjengelig om bergensmålet. Det er alt brukt til mange analysar, og mange nye formål kjem det til å bli brukt til vidare framover òg. Legg ein stor vekt på å ha tilgjengelig ein slik «database», kan ein rettferdiggjøre å dataføre så mye materiale. Det kan ein om ein reknar kartleggingsarbeidet som viktigare enn arbeidet med teorien.

Men reknar ein dei teoretiske spørsmåla for å vere hovudoppgåva i talemålsforskinga i dag, løyser ein ikkje dei først og fremst med eit stort materiale. I den greina som kallas sosiolingvistikk, er det viktigare f.eks. å klekke ut gode eksperimentelle språkbrukssituasjonar, som kan auke innsikta i og forbetre teorien om korleis og korfor menneska utnyttar språkvariasjonen. I arbeidet både med slike og andre problemstillingar, der ein føler teorien i dag er ufullstendig eller lite tilfredsstillande, er det viktig å gjøre avstanden kortare mellom hypotesedanning og ev. avkrefting/styrking. Det vil seie at analysen må komme fortare, slik at hypotesane kan presiseras og forskaren kan modnas. For å få til det må lydbandmaterialet vere mindre, men heller meir relevant for spesielle problemstillingar, og arbeidsmåten må vere raskare. Den skisserte måten med ekserpering direkte frå lydband og med EDB-bruk bare til statistisk behandling av beleggtala er den mest rasjonelle arbeidsmåten til det formålet.

Litteraturliste

Kort orientering om prosjektet «Talemål hos ungdom i Bergen». I: Talemål i Bergen 1/83. Bergen.

Lødrup, Helge 1982. Om ekserpering i sosiolingvistiske undersøkelser. (Prosjektnotat.)

Myking, Johan 1983. Adverba *då* og *no* i bergensmålet. Lingvistisk og sosiolingvistisk variasjon. I: Talemål i Bergen 2/83. Bergen.

Solevåg, Harald 1977. Dokumentasjon av TAGGING-systemet. (Nordisk institutt.) Bergen.

Helge Sandøy – amanuensis i norsk talemål ved Nordisk institutt, Universitetet i Bergen – er dagleg leiar for prosjektet «Talemål hos ungdom i Bergen.»

Personregisterloven og behovet for datavern

Thore Gaard Olaussen

Lovens utforming

Lov om personregistre m.m. trådte i kraft 1. januar 1980. Bortimot et tiår før ble det første utvalget med oppgave å vurdere reguleringer på bruk av persondata, nedsatt. Utvalget med professor *Tore Sandvik* som formann la fram sin utredning i april 1974 med forslag om lovreguleringer for privat sektor innenfor områder som kredittopplysning, databehandling, (databehandlingsforetak), adresserings- og distribusjonsvirksomhet og markeds- og opinionsundersøkelse[1]. Ettersom mandatet i første rekke rettet søkelyset mot bruk av persondata i kredittopplysningsvirksomhet, ble utvalget i ettertid ofte kalt «kredittopplysningsutvalget».

Det andre utvalget med *Helge Seip* som formann ble oppnevnt i 1972 med oppgave å vurdere bruk – og mulig misbruk – av persondata og personregistre i offentlig sektor. Utvalget la fram sin utredning i februar 1975 med lovforslag som bl.a. innebar generell konsesjon på personregistre i offentlig forvaltning[2]. Imidlertid ble det fra utvalgets side foreslått at man burde sammenføye de to lovutkastene fra Sandvik-utvalget og Seip-utvalget. Vi fikk således en lov som i tråd med den Sverige innførte allerede i 1973, omfattet konsesjonsplikt på alle personregistersystem som medfører bruk av edb. I tillegg fikk vi i Norge også konsesjonsplikt på manuelle registersystem som inneholder særlig sensitiv informasjon[3].

Loven som ble behandlet og vedtatt av Stortinget 1978, kan vi grovt oppsummere i følgende punkter:

1. Et datatilsyn opprettes til å behandle konsesjonssøknader ved etablering av personregister og å utøve kontroll med slike registre. Datatilsynet skal også behandle konsesjonssøknader fra kredittopplysningsinstitusjoner, databehandlingsforetak, adresserings- og distribusjonstjenestefirma og opinionsinstitutt og utøve kontroll med disse.
2. Lovens personbegrep gjelder «juridiske personer» – dvs. at den foruten fysiske personer også omfatter «sammenslutninger» og «stiftelser»[4].
3. Konsesjonsplikten i loven dekker som nevnt både registersystem som medfører bruk av edb og manuelle registre som inneholder
   - informasjon om rase, eller politisk eller religiøs oppfatning
   - informasjon om at en person har vært mistenkt, tiltalt eller dømt i straffesak

- informasjon om helseforhold, eller misbruk av rusmidler
- informasjon om seksuelle forhold
- annen informasjon om familieforhold enn slik som gjelder slektskap eller familiestatus, formuesordning mellom ektefeller og forsørgelsesbyrde[5].

4. I konsesjonssøknaden må det spesifiseres hvilken informasjon som skal brukes foruten begrunnelse for slik registrering. Datatilsynet på sin side spesifiserer bruken av registeret og eventuelt hvilke informasjonselementer som det ikke kan gis tillatelse til å bruke.
5. Loven har innført innsynsrett som gir alle som ønsker det, rett til å få vite hvilken informasjon som er lagret om seg selv. Denne retten gjelder imidlertid ikke «registre som bare brukes til statistikk, forskning eller generelle planleggingsformål»[6].
6. Data innsamlet av private meningsmålingsinstitutter skal ikke lagres i mer enn et halvt år i en form som kan identifisere enkeltpersoner[7].
7. Loven regulerer også overføringer av data til utlandet. Et konsesjonspliktig register skal ikke sendes ut av landet uten tillatelse fra Datatilsynet[8].

Det går tydelig fram at loven ikke bare gjelder edb og bruken av denne teknologien i samfunnet. Det er jo også lovens intensjon å dekke både den nye teknologi og de mer tradisjonelle hjelpemidler for informasjonsregistrering. I mange henseender har imidlertid både loven og utøvelsen av denne vanskeligheter med å nå fram overfor nye teknologiske utfordringer. For det første nytter loven en terminologi som «datafolk» vanskelig finner seg «hjemme i». Lovens betegnelse er «Personregisterloven» – og ikke «Dataloven». Man bruker uttrykk som «innføre opplysninger i et register» eller «opprettelse av personregister»[9]. For det andre risikerer man at teknologien «løper fra» loven, idet lovbestemmelsene står i fare for ikke å dekke utviklingen på det teknologiske området. Dette bekymrer også Datatilsynet[10]. Et tredje moment som bør nevnes, er at Datatilsynets saksbehandlere har juridisk bakgrunn og lite erfaring med edb-teknologi og bruk av data.

Konsekvenser for forskningsvirksomhet

Vi skal spesielt se på hvilke følger innføringen av Personregisterloven har eller eventuelt vil få på forskningens bruk av data:

1. Søknad om konsesjon kan forsinke (og således fordyre) forskningsprosessen. På grunn av ressursknapphet og bemanningssituasjon i Datatilsynet tok det uforholdsmessig lang tid før man greidde å absorbere søknadene fra allerede *eksisterende* dataregistre. Resultatet har således vært sen saksbehandling, og lang venting for mindre «viktige» saker. Datatilsynet har selv påpekt i flere sammenhenger at bemanningen ikke står i forhold til saksmengden[11]. Systemet er slik at man må vente på Datatilsynets behandling av konsesjonssøknaden *før* man gjør bruk av de aktuelle data.

2. En god del av de data som brukes innen visse forskningsprogram (eksempelvis innen fagområdene medisin, psykiatri, psykologi og sosiologi), inneholder informasjon av sensitiv karakter. For gitte forskningsoppgaver er det ikke til å unngå at forskeren må ha adgang til å bruke denne type informasjon. Hvordan skal en helseundersøkelse gjennomføres uten adgang til å bruke data om folks helseforhold, eller hvordan skal valgforskeren gå fram uten å få bruke opplysninger om folks stemmegivning? Lovens §6 sier helt klart at registrering av sensitive data ikke må finne sted uten at det er nødvendig, og det heter at «registrering av personopplysninger skal være saklig begrunnet ut fra hensynet til administrasjon og virksomhet i det organ eller foretak som foretar registreringen». Spørsmålet som med berettigelse kan stilles fra forskningshold, er om Datatilsynet er ekstra varsomme med å gi konsesjon til forskningsprosjekt som vil bruke sensitive data, eller om vilkårene for konsesjon vil være så restriktive at innsamling eller bruk av dataene vil vanskeliggjøres.
3. Bortsett fra regelen om innsynsrett skjelner ikke loven mellom *administrativ* bruk av data og data brukt til *forskningsformål*. Datatilsynet er således ikke forpliktet til å ta «særlige» hensyn til forskningen i konsesjonsbehandlingen. Fra forskningshold er det lagt vekt på å understreke hvordan forskning og statistikkproduksjon avviker fra ren administrativ bruk av data om enkeltpersoner[12]:
   i) Man er opptatt av *tendensene* i materialet, og analyserer datasettet med dette for øyet ved hjelp av krysstabeller, korrelasjoner og annen aggregert statistikk.
   ii) Tiden som går med til bruk av data til analyse er forholdsvis kort og prosessen av mer sporadisk karakter sammenlignet med etablerte dataregistersystem til administrativ bruk.
   iii) Antallet på de som vil ha tilgang til identifiserbare persondata vil være meget begrenset, fordi selve identifikasjonen på den enkelte vanligvis vil være av mindre interesse i forskningssammenheng.
4. Konsesjonen vil gjelde så lenge forskningsprosjektet pågår. Datatilsynet kan etter prosjektavslutning pålegge prosjektledelsen å slette data. Dette er også i samsvar med det syn at et konsesjonspliktig register ikke bør eksistere når det ikke lenger er saklig behov for det[13]. Det er åpenbart at for den forskning som er avhengig av denne type data, vil slike krav by på spesielle problemer. Pålegg om sletting vil bl.a. gjøre det umulig å gjennomføre såkalte panelundersøkelser eller etterundersøkelser av noe slag. Et annet viktig moment er at sletting av datagrunnlaget for et forskningsprosjekt vil vanskeliggjøre kontroll av forskningsresultatenes holdbarhet i ettertid.

Finnes løsninger?

Allerede i lovforberedelsene var det flere fra forskningssiden som tok

Tegning: Øystein Reigem.

opp spørsmålet om Datalovens innvirkning på en del viktige forskningsfelt. Enkelte fryktet at innføring av en datalovgivning ville innskrenke mulighetene til å bruke persondata i en slik grad at det ville vanskeliggjøre effektiv forskning på felter som er avhengig av tilgang til denne type informasjon. En del rapporter fra Sverige beskrev svenske forskeres erfaringer med innføring av tilsvarende datalov. Inntrykket var ikke særlig oppmuntrende. Kritikken gikk på den svenske datainspeksjonens til dels unødvendige innblanding i forskningsvirksomheten og den forsinkelse dette medførte[14].

Stein Rokkan var en av dem som på norsk side gjorde oppmerksom på forskningens vanskeligheter i forhold til den datalov som Seip-utvalget hadde foreslått[15]. Han påpekte at man fra forskningshold måtte klart tilkjennegi sine interesser når det gjaldt adgang til å bruke data, og han rådet bl.a. NAVF til å gjøre de utspill som var nødvendige[16]. Etter forslag fra bl.a. NSD vedtok styret i NAVF å etablere et eget sekretariat for datavern-spørsmål (NAVFs datafaglige sekretariat). Sekretariatet ble etablert i tilknytning til NSD i Bergen, og innledet straks forhandlinger med Datatilsynet.

Fra forskningssiden var det to hovedspørsmål som måtte avklares. Det første var å få en forsikring om at dataloven ikke skulle hindre eller forsinke i noen grad forskningsprosjekt som er avhengig av tilgang til og bruk av persondata. Forskningsrådet aksepterte her en prosedyre der

meldinger sendes Datatilsynet om hvilke NAVF-prosjekt som omfattes av loven. Med hver melding følger en innstilling fra NSD. Denne ordningen gjelder alle fagområder innen NAVF-systemet.

Det andre hovedspørsmålet var å få i stand arkivordninger for forskningsdata som sikret mulighetene for oppbevaring av data også etter at forskeren har avsluttet sitt arbeid med datamaterialet. Det man ønsket å unngå var at viktige forskningsdata ble tapt for alltid p.g.a. restriktive konsesjonsvilkår fra Datatilsynets side. Datatilsynet var heller ikke rette instans til å vurdere hvilke forskningsdatasett som burde sikres oppbevaring for fremtidig bruk, ble det fremholdt fra forskningssiden. Datatilsynets oppfatning av lovbestemmelsene var at hovedregelen var sletting av data etter avslutning av forskningsprosjekt. I rammekonsesjonsavtalen møtte imidlertid Datatilsynet forskningsinteressene på halvveien. Her heter det at «personopplysninger som er registrert på edb eller andre media, og som det er særlig grunn til å anta at det vil bli behov for i fremtidig forskning, kan overføres til Norsk samfunnsvitenskapelig datatjeneste (NSD), ...»[17]. Avtaleteksten bærer preg av en rekke forbehold fra Datatilsynets side: I avtalen er det nevnt at et eget utvalg, oppnevnt riktignok av NAVF, skal vurdere hvilke data som kan overføres til den arkivinstans som har Datatilsynets godkjennelse. Endelig helgarderte Datatilsynet seg ved å presisere at overføring bare kan finne sted dersom det ikke er i strid med forskerens taushetsplikt. Om dataoverføring virkelig ville være brudd på taushetspliktsregler, tok ikke Datatilsynet standpunkt til.

Dette siste spørsmålet har således vært uavklart helt fram til NAVF i år fikk fremlagt en innstilling som ikke fant at overføring av forskningsdatasett til arkivinstans var i strid med forskeres taushetsplikt[18]. Datatilsynet som var representert i det utvalget som fremla innstillingen, aksepterte denne fortolkningen.

I samsvar med rammekonsesjonen har NAVF nå oppnevnt et råd for dataarkivering med professor *Jan Fridthjof Bernt* som formann. Rådet er videre sammensatt med representanter fra humanistisk, samfunnsvitenskapelig og medisinsk fagråd. Forskere som har bygd opp datamateriale i henhold til bestemmelsene i NAVFs rammekonsesjon og som ønsker dette materialet deponert i dataarkiv, må få dette avklart gjennom dette nyoppnevnte rådet.

I utgangspunktet kan opplegget virke byråkratisk og tungrodd. Det er imidlertid viktig å være klar over at man her har fått til en ordning som skal tilfredsstille to til dels kryssende hensyn: 1) personvernet der lovbestemmelsene ønsker å begrense bruken av persondata dersom slik bruk ikke er helt nødvendig (relevanskravet), og 2) forskningsinteressene som på dette området ønsker friere tilgang til og bruk av data.

Personverninteressene og forskningsinteressene har likevel sammenfallende syn på et avgjørende område, og det gjelder datasikkerhet. Dette går både på sikring mot uautorisert tilgang på data og sikring mot ødelegging av data. I Datatilsynet har man hatt et arbeid i gang med

utforming av forskrifter for datasikkerhet, men hittil har man av ressursmessige hensyn ikke kunnet gå inn i problematikken med full tyngde. Et sett forskrifter som innebærer forpliktelser når det gjelder sikring av data, vil selvsagt ha ressursmessige konsekvenser for den ansvarlige for datamateriale. For forskningen er det derfor viktig at man får i stand ordninger som ivaretar arkiveringsbehov for forskningsdata, og at slike oppgaver legges til instanser som både har ressurser og kompetanse til å løse dem[19].

For å få en konstruktiv debatt om disse spørsmålene mellom instanser som i utgangspunktet har ulike interesser å ivareta, bør vi etter min mening utvide «verneområdet» noe. Debatten om datasikring kan ikke begrenses kun til personvernet - vi må også kunne snakke om et informasjonsvern, og således få en videre horisont på disse problemene. For å diskutere spørsmål om «verneverdige» data er vel heller ikke Personregisterloven det «beste» utgangspunkt - kanskje snarere tvert imot. Men nettopp derfor burde det være av almen interesse at Datatilsynet ble engasjert i denne problematikken og var med i diskusjonen om løsningene på disse spørsmålene[20].

Noter

1. NOU 1974: 22 «Persondata og personvern».
2. NOU 1975: 10 «Offentlige persondatasystem og personvern».
3. For flere detaljer vises til Jon BING og Knut S. SELMER *A Decade of Computers and Law*, Universitetsforlaget, Oslo 1980, og til Ot.prp. nr. 2 (1977-78) «Om lov om personregistre m.m.»
4. Jfr. lovens pgf. 1, annet ledd.
5. Jfr. lovens pgf. 9, første ledd.
6. Jfr. lovens pgf. 7, annet punktum.
7. Jfr. lovens pgf. 33, annet ledd.
8. Jfr. lovens pgf. 36, første og annet ledd.
9. Det kan opplyses at Seip-utvalget i sin innstilling brukte betegnelsen «Datalov» og «persondatasystem» istedenfor personregister.
10. Det vises til presseoppslag, bl.a. Bergens Arbeiderblad 7.8.84.
11. Det kan eksempelvis vises til årsmeldingene fra Datatilsynet og til Arve FØYEN «Utredning om endringer i personregisterloven», Complex nr. 1/83, Universitetsforlaget, Oslo 1983.
12. Jarle BROSVEET og Ørjar ØYEN «Norwegian Data Law and the Social Sciences», foredrag til CESSDA/IFDO konferansen i Køln 9.-11. august 1978.
13. Dette syn blir gjenspeilt i det såkalte relevanskravet som er bygget inn i lovens pgf. 6, første ledd.
14. Det vises spesielt til Tore DALENIUS og Anders KLEVMARKEN, red. *Personal Integrity and the Need for Data in the Social Sciences*, Samhällsvetenskapliga forskningsrådet, Stockholm 1976.
15. Stein ROKKAN «The Production, Linkup and Communication of Social Science Data: A Survey of Developments in Norway», notat, Bergen 1977.
16. Stein ROKKAN «Norsk samfunnsvitenskapelig datatjeneste og den kommende datalov» i *Personvern og samfunnsforskning. Seminarrapport*, NSD-rapport nr. 17, Bergen 1977.
17. NAVFs rammekonsesjon punkt 4.9, første ledd.

18. NAVF-innstilling «Oppbevaring og gjenbruk av personidentifiserbare forskningsdata underlagt lovbestemt taushetsplikt». Utredning fra utvalg nedsatt av NAVFs styre, Oslo 1984.
19. Jeg vil her vise til en interessant artikkel av Lars GULDBRANDSEN «Varig vern av våre gallupdata» i Forskningsnytt nr. 4/84. Konklusjonen i artikkelen er at vern av denne type data kulturelt sett er like viktig som vern av trykte dokumenter, og bør ivaretas på en tilsvarende betryggende måte.
20. Det vises til NOU 1984: 3 «Frå informasjon til kulturarv», innstillingen fra Arkiveringsutvalget med Berge FURRE som formann. Her er det lagt fram forslag om hvordan sikre informasjon fra ulike medier for fremtidig bruk.

Cand. polit. Thore Gaard Olaussen er konsulent ved Norsk samfunnsvitenskapelig datatjeneste.

Standardising Transcriptions of L. Wittgenstein's Nachlass

Michael Kulemann

Ludwig Wittgenstein left behind a lot of unpublished written material. The greater part of this material is now available in machine-readable form at the Norwegian Computing Centre for the Humanities. Part of it has been transcribed during the last years by a group of Norwegian scholars interested in Wittgenstein's Nachlass. Another part of the transcriptions are a product of the «Tübinger Wittgenstein-Archiv» and were made during the years 1977 to 1980. Nearly the whole Nachlass is available as a microfilm produced by Cornell University (USA).

The first question one could ask is: Why should you make a text machine readable at all, that is available on microfilm (or on paper copies of that film)? If you are interested in the text, why not read the film (or its copy)? Isn't a transcription a waste of time? I think that the Nachlass in machine-readable form has three major advantages:

(1) The Nachlass consists of hand- and typewritten texts that were corrected by Wittgenstein again and again (for an example see the copy below). They are therefore very hard to read. So to *save* time you have to make a transcription anyway.
(2) You can use the computer to do boring work for you: make word lists, find all paragraphs where a given word is used and other things like that (especially in a huge corpus like Wittgenstein's Nachlass such tools are useful and maybe there are even lines of research that could not possibly be followed without them).
(3) Computers can be used to prepare texts for phototypesetting. So making a text machine readable is a first step towards a printed edition of this text.

Looking at the figure you will immediately see the difficulties that arise when trying to transcribe the texts: a complicated two-dimensional structure has to be transcribed in a one-dimensional, sign-by-sign way. Systems that do this cannot be designed at one stroke: while transcribing a text you will from time to time find phenomena that cannot be represented in the code system. Then you will have to change them. So the code system evolves and once in a while even undergoes a radical change.

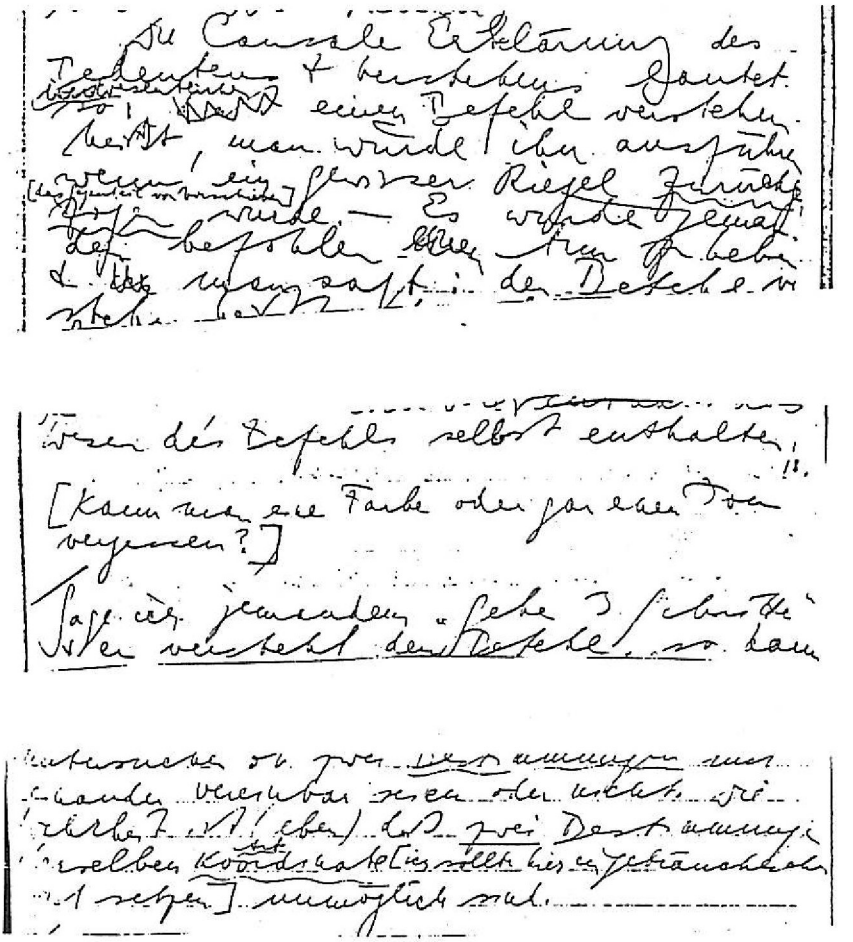
This process yielded the following situation: Between 1977 and 1980 texts were transcribed according to three major code systems and in each text there could be minor differences between the code system actually used and the main system «officially» used at that time. Those «official» systems were labelled I, II and III. The code system used in

Norway (COSY TRAWMA) was developed on the basis of code system III by *Asbjørn Brændeland*. So just now there are texts transcribed according to four code systems and an unknown number of minor variants of them. There were other things that contributed to the coding differences between the texts that presently constitute the machine readable part of L. Wittgenstein's Nachlass: The Norwegian transcribing conventions do not allow word divisions (and I think that is a good idea), while the Tübingen conventions allowed them and even enforced them: If Wittgenstein himself had to divide the last word of a page, then the code for «newpage» would be found in the middle of a (divided) word. In the Tübingen transcriptions there are signs for the German Umlaute (ä, ö, etc.) you could not represent on normal input and output devices (terminal, printer) here in Norway. On the other hand Umlaute were transcribed (as usual internationally) as: ae, oe, etc. But that is not all. Wittgenstein himself sometimes had to use a typewriter without Umlaute. So there are some texts transcribed in Tübingen where the Umlaute are represented as: ae, oe, etc.

At this point (if not earlier) anybody interested in Wittgenstein's *philosophy* and not in his *word divisions* and subtleties about how to represent *Umlaute* might lose his (or her) temper and argue: Whoever is interested enough in Wittgenstein's philosophy to look into his Nachlass and is able to understand it will have no difficulty in interpreting the different representations of the Umlaute correctly. The word divisions will not be a problem at all. The different coding systems might need a bit of learning, but that cannot be so difficult. So: care about the real problem: *understanding* Wittgenstein. All those things about a standardisation (so easy that anybody can do it without the help of a computer) is just a waste of time.

This argumentation is sound as far as it goes: one can do a lot of things with the transcriptions, even if they are not standardised. But there are some things one cannot do. I want to give some examples of this. «Tätigkeit» is a central concept in Wittgenstein's Spätphilosophie. For one reason or another it might be interesting to have a look at all the places where Wittgenstein mentions it. Normally this would be easy: there are standard programs that can be used to find those places. But because there is an Umlaut in «Tätigkeit» there will be difficulties in using those programs. A similar problem results from the word division: you will not be able to find the places where the word you are looking for is divided. There are reasons for standardising the code system too. In Oslo programs have been developed which presuppose the Norwegian code system (e.g. a program to increase readability by printing the transcriptions without codes and without the texts Wittgenstein deleted). So only standardisation makes it possible to use all the advantages of machine readable transcriptions.

Part of the standardisation must be done manually. In transcriptions



Part of one of Wittgenstein's handwritten texts.

according to code systems I and II information is missing that should be there according to the Norwegian code system. But another part of the standardisation can be done automatically: It is possible to bring all the information that is there in the form required by the Norwegian code system and writing conventions.

A program that does this should be general. Otherwise you have to have a tailor-made program for nearly every text. So the code changing program that has been developed during the last two months takes as input not only a transcription but also a description of the code system that is used in that transcription. In order to use the program on different texts the only thing you have to change is the description of the

code system.

Using the code changing program in connection with the Norwegian code checking program there is an effective, semi-automatic way of recoding the Tübingen transcriptions: One runs the code changing program and takes the result as input for the code checking program. The error messages which are the output of the code checking program can be used to correct the description of the code system used by the code changing program.

Even so, changing the Tübingen transcriptions will be a long and complicated process needing a lot of work. I think two things can be learned from this example:

(1) It is worthwhile doing some work before beginning to make transcriptions of difficult texts. If the system for the transcription is good at the beginning not so many changes have to be made later.

(2) One should not change the transcription system unless one is forced to do so. But if one has to make changes they must be well documented.

M.A. Michael Kulemann has worked for the Tübinger Wittgenstein-archiv and recently spent two months at the Norwegian Computing Centre for the Humanities in connection with the data processing of Wittgenstein's Nachlass.

RAPPORTER

Spørjeundersøking om bruken av statistiske metodar i språk- og litteraturforskinga

Ole Lauvskar

NAVFs EDB-senter for humanistisk forskning sende 12.3 dette året ut eit spørjeskjema om bruken av statistiske metodar blant språk- og litteraturforskarane. Eit av måla var å få eit oversyn over kor stor bruken av slike metodar er hos denne gruppa. Eit anna mål var å kartleggja behovet for å bruka slike metodar. Den nye datateknologien har gjort det lettare å arbeida med kvantitative problemstillingar. Bruken av statistiske metodar vil difor vera meir aktuelt å tenkja på – òg blant språk- og litteraturvitarar.

Her vil eg berre gje eit kort resyme over kva som kom fram av undersøkinga. Interesserte kan henvenda seg til NAVFs EDB-senter for å få ein meir fyldig rapport. Rapporten er på om lag 20 sider.

Det vart send skjema ut til 470 språk- og litteraturvitarar. I skrivet som følgde med spørjeskjemaet, vart ein oppmoda om å gje skjemaet vidare til hovudfagsstudentar. Dette vart i liten grad gjort. I alt var det 203 som svara. Det gjev ein svarprosent på om lag 43.

Det var visse skilnader på svarprosentane ifrå det eine universitetet (eller høgskulen) til det andre. Skilnadane var noko større viss ein ser på kva stillingar dei som svara hadde. Svarprosenten vart jamnt over høgare di høgare ein kom i det akademiske hierarkiet.

Deler ein den totale svarprosenten i to, ein for språk- og ein for litteraturvitarar, får ein 44 og 43. Det må kunna seiast å vera overraskande likt. Det hadde vore naturleg at svarprosenten hadde vore ein del høgare hos språkvitarane enn hos litteraturvitarane, sidan språkvitarane jamnt over brukar meir statistiske metodar enn kva litteraturvitarane gjer.

Av språkvitarane svara 40% at dei brukte eller hadde brukt statistiske metodar. Den tilsvarande svarprosenten for litteraturvitarane er 16. Alt i alt svara 32% at dei hadde brukt slike metodar.

Dei som skreiv at dei hadde brukt statistiske metodar vart bedne om å svara på seks ekstra spørsmål. Av svara ser ein at det var like mange

som sa dei hadde brukt meir avanserte metodar som enkle metodar (utrekning av gjennomsnitt, oppteljingar osv.). Men det var færre som hadde funne fram til *eigna* metodar. Halvparten kunne fortelja at dei hadde brukt edb i arbeidet, og dei fleste av desse var fornøgd med programma. Om lag to tredjepartar av dei som svara hadde tileigna seg sine statistiske kunnskaper ved eige initiativ, medan halvparten skreiv at dei hadde motteke hjelp hos andre når dei hadde utført analysane. Det vart òg spurd om kva del av arbeidet som var mest tid- og arbeidskrevande. Av dei som svara, svara dei fleste at det var tilretteleggjing av dataane og meir metodiske spørsmål som var dei største problema. Jamfører ein svara som språk- og litteraturvitarane gav, finn ein liten skilnad. Men ein bør vita at det berre var 11 litteraturvitarar som sa at dei hadde brukt statistiske metodar. Det gjer at prosentar som vert utrekna er svært usikre.

Eit spørsmål var firedelt. Her skulle ein gje uttrykk for om det var tilstrekkeleg tilgong til konsulenthjelp, læremiddel, opplæringstiltak og edb-program der kor ein forska. På dette spørsmålet skulle alle svara, anten dei hadde skrive at dei hadde brukt statistiske metodar eller ikkje. Nedanfor er ein tabell som syner i prosentar kva som vart svara på desse spørsmåla. Dei som ikkje svara, er ikkje medrekna.

	Brukt stat. met.			Ikkje brukt stat. met.		
	God nok tilgong	Ikkje god nok	Usikker	God nok tilgong	Ikkje god nok	Usikker
Konsulenthjelp	16	54	30	19	28	53
Læremiddel	19	45	36	13	25	63
Opplæringstiltak	5	67	28	12	32	55
Edb-program	15	45	40	17	31	51

Ein ser at dei som hadde brukt statistiske metodar, er mindre fornøgd med den aktuelle tilgongen enn dei som ikkje hadde brukt slike metodar. Men det er òg færre som har skrive usikker i den første gruppa enn i den siste. Ei rimeleg forklaring er at dei som ikkje hadde brukt statistiske metodar, heller ikkje hadde undersøkt kor stor tilgongen var. Dei har jamnt over òg mindre behov for ein auka tilgong.

Dei som svara at dei ikkje hadde brukt statistiske metodar, vart bedne om å svara på eitt fleirdelt ekstraspørsmål. Her gjekk det fram av svara at dei fleste av språkvitarane meinte at statistikk var relevant innafor deira fagfelt. Ein fjerdepart av dei som meinte det, meinte òg at statistikk var relevant for deira personlege forskingsoppgåver. Blant litteraturvitarane meinte dei fleste at statistiske metodar ikkje var relevante innafor deira fagfelt.

I kommentarane til svara var det enkelte som hevda at viss ein brukte statistiske metodar, kunne dei kvantitative metodane få ei for stor vekt i høve til dei kvalitative. Dette synet fann ein først og fremst hos litteraturvitarane. Som det blant anna framgår av tabellen over, er det mange som kjenner eit behov for større mulegheiter til å nytta statistiske metodar enn kva ein har i dag. Særleg ser behovet for opplæring ut til å vera udekka.

Nordiske arkivdager i Ebeltoft 2.-5. august 1984

Anne Hals

På de 14. nordiske arkivdager som ble holdt i Ebeltoft, Danmark fra 2.-5. august 1984 ble det forsøkt med en ny arbeidsform i forhold til tidligere arkivdager. Bortsett fra under *Erik Lønroths* avslutningsforedrag var deltakerne hele tida inndelt i ulike arbeidsgrupper. Man kunne velge mellom følgende grupper:

- arkivenes arbeidsprogrammer
- edb og arkivene
- registerregler: administrasjons- og forskningsregistre
- etisk kassasjon
- tilgjengelighetsregler
- tverregistrering av privatarkiv
- statens eiendomsrett til arkivalia
- administrasjonshistorie

Jeg vil i det følgende bare referere fra diskusjonen i arbeidsgruppa «Edb og arkivene».

Uten nærmere spesifisering av emnet hadde alle deltakerne levert inn skriftlige innlegg før møtet. Innleggene spente fra spørsmål om edb-baserte journalsystemer i offentlig forvaltning til Arkivverkets egen bruk av edb. På møtet valgte man imidlertid å konsentrere seg om følgende problemområder:

1. Mottak av edb-arkivalia

Mens Arkivverket i Danmark og Sverige har mottatt edb-arkivalia i form av magnetbånd siden midten av 70-tallet, er det først i år at man har gjort forberedelser til å motta magnetbånd her til lands. Den erfaring man hittil har gjort i de andre nordiske land viser at det ikke er problemfritt å motta edb-arkivalia. For det første er ofte edb-materialet i statsforvaltningen lite standardisert. Man bruker forskjellig maskinutstyr, båndene blir produsert med forskjellige tegntetthet, forskjellig

tegnkode osv. Likeledes har mange institusjoner ulik praksis når det gjelder vedlikehold av eldre magnetbånd. Det kan derfor ofte være vanskelig å konvertere eldre bånd over til formater som vi kan lese i dag. Det viste seg å være en viss uenighet blant deltakerne om hvilke krav Arkivverket kunne stille til de avleverende institusjonene når det gjaldt standardisering av magnetbånd. Mens man på norsk hold har utarbeidet helt klare kravspesifikasjoner til utforming av magnetbåndformat, var man på svensk hold mer reservert overfor muligheten av å få institusjonene til å følge slike krav. I Danmark er standardiseringsproblemet noe mindre, fordi de aller fleste av forvaltningens edbregistre blir kjørt på den statlige regnecentralen. For det andre krever det etter hvert ganske mye arbeid og store kostnader for Arkivverket å vedlikeholde magnetbåndene. I Sverige hvor Riksarkivet til nå har tatt imot ca. 3.000 magnetbånd, regner de med å måtte konvertere alle båndene minst hvert 10. år, dels på grunn av den teknologiske utviklingen og dels fordi jo eldre magnetbåndene blir, jo større fare er det for at båndene bli avmagnetisert og verdifull informasjon går tapt. I tillegg til konvertering av magnetbånd, må man også med jevne mellomrom (ca. hvert 2. år) kjøre tester på båndene.

Det var enighet blant alle deltakerne om at de magnetbånd som skal avleveres til Arkivverket, skal være helt nye og at de skal avleveres i 2 eksemplarer. Det ble også sterkt understreket av flere at man må sørge for å få avlevert tilfredsstillende dokumentasjon til de ulike edbsystemene. Dokumentasjonen må både være av teknisk og systembeskrivende art. Enkelte var imidlertid usikre på om mye av den dokumentasjonen som blir utarbeidet i forbindelse med driften av et edb-system, alltid er tilstrekkelig for arkivformål. Det ble reist spørsmål om hvor

Arkivdagene ble holdt i idylliske Ebeltoft - hovedstaden i Molboland.

langt man kunne gå i å kreve dokumentasjon som primært blir laget av hensyn til Arkivverket.

2. Edb-baserte journalsystemer

Fra norsk side presenterte *Ivar Fonnes* og *Anne Hals* retningslinjer for innføring av edb-baserte journalsystemer i statsforvaltningen. Retningslinjene er laget fordi en rekke statsinstitusjoner har begynt å ta i bruk eller planlegger å ta i bruk edb i journalføringen. Formålet med retningslinjene er å sikre at edb-journalsystemene blir utformet på en slik måte at de spesielle forsvarlighets- og sikkerhetskrav som Arkivverket stiller til journalføring, blir ivaretatt.

Jan Frisk ved Göteborgs statsarkiv viste til et edb-basert journalsystem som flere av de kommunale myndigheter i Göteborg har innført. Dette systemet er basert på fritekstsøking innenfor faste felt. Man har derfor valgt å sløyfe klassifisering av dokumentene dvs. henvisning til arkivnøkkel. I dette systemet blir det heller ikke tatt papirutskrifter av journalen. All søking foregår direkte fra skjerm og backup tas utelukkende i form av magnetbåndkopier av databasen. Både fra dansk og norsk hold ble det uttrykt skepsis til å sløyfe arkivnøkkel. Ikke minst av hensyn til behandlingen av det fysiske arkiv, er det en fordel å holde de enkelte dokumenter samlet i emneordnete saker. Det ble også uttrykt tvil om sikkerheten ved systemet var så god at man kunne sløyfe enhver papirutskrift.

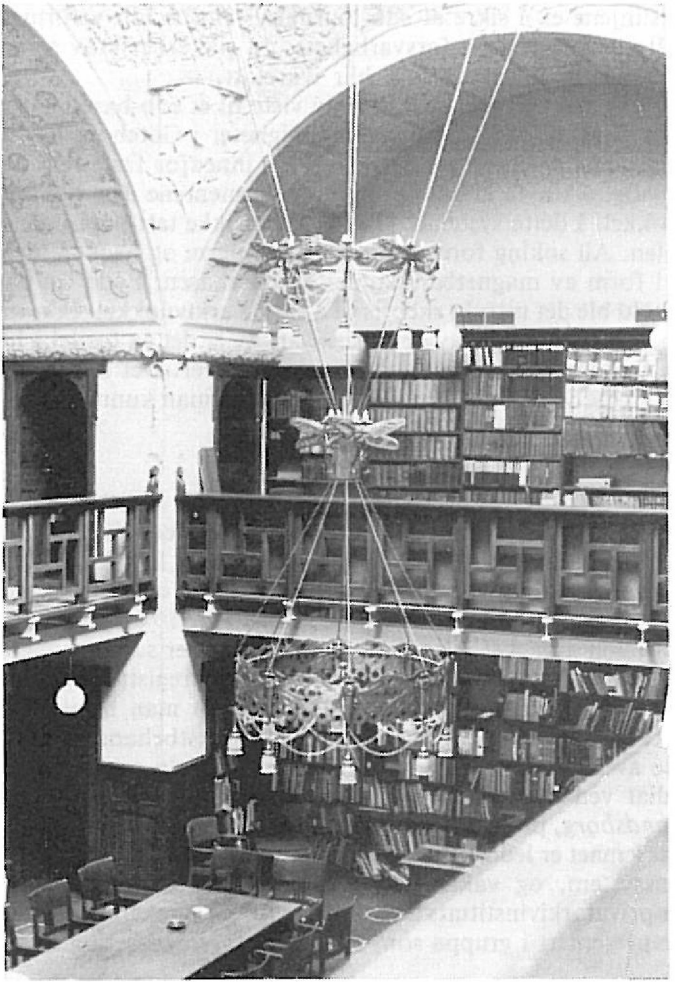
3. Edb-bruk i Arkivverket

Det er ingen av de nordiske arkiver som har mye erfaring i bruk av edb som internt hjelpemiddel. Det danske riksarkivet har i noen år hatt en terminal tilknyttet universitetet i København. Ved det svenske riksarkivet er det først i sommer anskaffet en IBM PC. I Norge har man nå i tillegg til den 4-bruker Altos som Riksarkivet eier sammen med Norsk privatarkivinstitutt og Sekretariatet for fotoregistrering, anskaffet mikromaskiner ved de fleste statsarkiv. Mens man hittil i de andre nordiske land hovedsaklig har arbeidet med tekstbehandling og tilrettelegging av edb-lagret tekst for trykking ved hjelp av fotosats, kunne stipendiat ved NAVFs EDB-senter for humanistisk forskning, *Hege Brit Randsborg*, presentere sitt registreringsskjema for katalogskriving. Dette skjemaet er ledd i arbeidet med å utvikle et edb-basert arkivinformasjonssystem, og vakte stor interesse blant de andre deltakerne. (Norsk privatarkivinstitutts edb-opplegg for en samkatalog for privatarkiv ble presentert i gruppa som diskuterte tverregistrering av privatarkiv.)

Under diskusjonen ble det nevnt en del andre edb-oppgaver som er på planleggingsstadiet, bl.a. kartlegging av institusjoner som har store edb-registre, oversikter over kassasjonsbeslutninger, registrering av

mikrofilmbestand, registrering av avleverte magnetbånd og deres tilstand m.m.

Når det gjelder edb-registrering av primærkilder blir dette arbeidet nå stort sett ledet av institusjoner utenfor Arkivverket som Demografiska databasen i Haparanda og Registreringssentralen for historiske data i Troms. Overføring av arkivmateriale til edb har vist seg å være så



Innlagt i programmet var et besøk til Ervervsarkivet i Århus. Arkivbygningen, som ble oppført i 1902, har en praktfull mottakelseshall og bibliotek i jugendstil.

ressurskrevende at de fleste var enige om at dette ikke i vesentlig grad kan prioriteres av Arkivverket, men må utføres i samarbeid med andre.

4. Edb-utdanning for ansatte i Arkivverket

I alle de nordiske land har man i de seinere år forsøkt å gi arkivarene en del kunnskaper om edb. Spørsmål som ble tatt opp i gruppa var hva slags edb-kunnskaper trenger den enkelte arkivar og hvordan skal man få gitt arkivarene de nødvendige kunnskapene. I tillegg ble det reist spørsmål om man ikke måtte basere seg på at arkivarene befinner seg på forskjellig nivå når det gjelder kunnskap om edb. Man trenger noen få med direkte edb-utdannelse, men i tillegg er det også et spørsmål om man ikke innenfor hvert arkiv må utdanne spesielle ressurspersoner (f.eks. minst I arkivar ved hvert statsarkiv/landsarkiv) som kan lede arbeidet med inspeksjon av edb-arkivalia m.m. innenfor vedkommende arkivdistrikt, og som til en viss grad kan veilede og gi opplæring til resten av institusjonens ansatte.

Gruppa kom fram til at det kunne være ønskelig med et fellesnordisk opplegg for dataundervisning. Man ble derfor enige om å be de nordiske riksarkivarene å søke om midler fra Nordisk kulturfond til et nordisk forskerkurs med sikte på å gi videreutdanning i edb for ansatte i Arkivverket.

Anne Hals er førstearkivar ved Riksarkivet. Hun arbeider bl.a. med Arkivverkets bruk av edb og mottak av edb-materiale fra statsforvaltningen.

On methods for using population registers in historical research

Rapport fra en konferanse ved Universitetet i Umeå 13.-17. august 1984

Eirik Lien

Denne konferansen var en fortsettelse av de to nordiske konferansene med tittel «Historiske databaser i Norden» som ble holdt i Norge i 1980 og i Danmark 1982. Denne hadde imidlertid en videre deltakerramme idet invitasjonen hadde gått ut over det nordiske området. Det deltok

omkring 40 stykker, omtrent halvparten fra Sverige og resten fordelt på Europa (Danmark, Finland, Norge, Nederland, Belgia, England, Italia), USA, Canada og Japan. Arrangør var Demografiska databasen i Umeå, med *Jan Sundin* som hovedansvarlig.

Denne gangen var konferansen strukket ut til ei hel uke (de er blitt lengre for hver gang), en endring som jeg synes var positiv. Det ble dermed god tid til samtale og diskusjon etter hvert innlegg.

Den første delen av konferansen var viet prosjektrapporter, der omtrent hver eneste deltaker fikk anledning til å orientere om sitt eget arbeid og få reaksjoner på det. Det var interessant å få en oversikt over det som skjer også andre steder enn i Norden. Spesielt med fjernsynets sommerserie «Shogun» i friskt minne, kunne en få perspektivet på *Akira Hayamis* foredrag «The shumon aratame cho: Japanese population registers».

En tredjedel av konferansen var satt av til disse prosjektorienteringene. Tema for øvrig var «Completeness of Data», «Record Linkage», «Hazard Models», «Sampling» og «Other Sources». Jeg skal kommentere en del av dem.

Under «Completeness of Data» tok bl.a. *Andrea Schiaffino* fra universitetet i Bologna opp i hvor stor grad en kan stole på offentlige registre som en sikker kilde for demografisk forskning. Med italienerens sans for dramatikk framstilte han levende hvordan han unnlot å kontakte myndighetene på det stedet han flyttet til. Med slike tilstander blir det lett skjeve utvalg og huller i registrene! Det er i hvert fall tydelig at forskjellige land strir med ulike typer problemer.

Under temaet «Record Linkage» luftet *Hans Chr. Johansen* fra Odense universitet spørsmålet om hvą som vil være den beste enheten å lenke informasjonen til: personen, familien, bostedet. Ut fra de særegne forholdene i Danmark, der navnevariasjonen er adskillig mindre enn i de andre nordiske land og der en dermed har færre holdepunkt i *navnet* som identifikasjon, har han arbeidet med å bruke adressen som lenkingsenhet. Ved å ta utgangspunkt i entydig husnummerering i skattelistene har en referanser å knytte personinformasjonene til.

Temaet «Hazard Models» ga en – for meg – ny dreining av hvordan en kan studere demografiske data. Innfallsvinkelen er å ta i bruk sannsynlighetsteori med sannsynlighetsmål basert på virkelige tilfeller, f.eks. fødselsrater. Ved å bruke disse ratene under andre forhold eller i et annet tidsrom kan en få fram forskjeller som kan gi grunnlag for analyser, med andre ord simulering. Metoden virker på meg – og en del andre – noe teoretisk og fjern fra det å studere levende menneskers atferd. Det er mange utenomliggende forhold, som er tids- og stedsavhengige, som ikke kan komme med i en slik betraktningsmåte. Det levende menneske med egen vilje og forstand som handler rasjonelt under ulike vilkår, har lett for å forsvinne. Det ble gitt to foredrag under dette temaet, symptomatisk(?) nok av to amerikanere, *Myron Gutman* fra Houston og *George Alter* fra Bloomington.

Arrangementet fant sted i Humanisthuset på Umeå universitet som i sommerstillheten gav mer enn tiltrekkelig plass for oss. Arrangørene hadde også lagt opp sosiale innslag med mange muligheter til forbrødring og forsøstring, alt fra sightseeing i Umeå og omegn til (nesten) kollektiv avnytelse av surstrømning. I et forsøk på å viske ut sosiale og nasjonale forskjeller fikk alle tildelt databasens T-skjorte. Ved avslutningen av konferansen inviterte universitetet ved prorektor til bankett på en av byens restauranter.

En av kveldene var satt av til omvisning på Demografiska databasen med orienteringer og demonstrasjoner. Skoleprosjektet MIS (Människan i samhällsomvandlingen) med *Egil Johanson* som leder hadde også en avdeling, både med utstillinger under selve konferansen og med demonstrasjoner. Formidlingsaspektet er en viktig del av det arbeidet som drives ved databasen, og vil stadig bli viktigere. Denne konferanseserien vil etter planen bli ført videre neste høst i Stockholm, og da nettopp med formidling som et av hovedtemaene. Stockholms historiska databas vil stå som arrangør.

Denne konferanseserien har så langt vært et inspirerende tiltak og virkelig vært med på å fremme samarbeid i Norden. At konferansen denne gangen fikk et mer internasjonalt preg enn de forrige, var ikke noen ulempe, men det gikk muligens en del på bekostning av den nordiske deltakelsen.

Eirik Lien er konsulent ved Edb-tjenesten for humanistiske fag, Universitetet i Trondheim.

ECAI 84 - 6th European conference on artificial intelligence

Øystein Reigem

ECAI 84 ble holdt i Pisa 5.-7. september, med tilhørende tutorials den 3. og 4. I tillegg til selve konferansen ble det arrangert en industri-sesjon og en sesjon om ESPRIT og kunstig intelligens den 4. (ESPRIT er det europeiske svaret på det japanske femtegenerasjonsprosjektet.) En utstilling av utstyr pågikk hele uken.

Kursene 3. og 4. september tok for seg fire emner, nemlig programmeringsspråk for kunstig intelligens, ekspertsystemer, analyse av naturlig språk og roboter og robotsyn. Selve konferansen tok i tillegg opp emnene kognitiv modellering, planlegging og søking, systemstøtte, kunnskapsrepresentasjon, automatisk programmering, teorembevising,

logikkprogrammering, anvendelser i industrien, læring (dvs. systemer som lærer), samt filosofiske implikasjoner. Det var dessuten diverse innlegg som ikke falt inn under noen av de nevnte emnene. I alt ble det holdt ca. 160 foredrag, og man kunne velge mellom 5-6 parallelle sesjoner. Totalt deltok 7-800 personer på hele eller i deler av arrangementet.

Jeg har valgt å begrense referatet til emnet ekspertsystemer, og for å være enda snevrere kun referere *Bob Wielingas* tutorial-innlegg. Bob Wielinga er førstelektor ved universitetet i Amsterdam, avdelingene for psykologi og samfunnsvitenskapelig informatikk.

Et ekspertsystem er et datamaskinprogram som skal fungere som en ekspert på et bestemt område. Systemet må besitte kunnskap om området, ikke bare fakta, men også regler for hvordan slutninger kan trekkes.

Ekspertsystemer kan ha flere roller: De kan være konsulenter, eller de kan være satt til å styre en prosess.

Eksempler på dagens anvendelsesområder for ekspertsystemer er

- medisinsk diagnose og valg av behandlingsmetoder
- feilfinning i komplekse apparater
- tolking av komplekse data og signaler (f.eks. massespektrogram)
- konfigurering av datamaskiner
- forutsigelse av potensielle mineralforekomster
- undervisning i problemløsing i avanserte fagområder som fysikk, medisinsk diagnose, osv.

Wielinga innledet med en utgreiing om første generasjons ekspertsystemer (dvs. fram til i dag).

De tidligste systemene trakk slutninger ved «svake» metoder:

- Søking, f.eks. gjennom systemets samling av fakta for å verifisere et utsagn. Dette er en enkel, men svært tidkrevende metode.
- Generelle problemløsingsmetoder som generering og testing, dvs. at systemet på hvert trinn i løsningsprosessen genererer en mengde forslag til (del)løsninger som så alle blir testet. Uten en strategi for å begrense det som blir generert, vil den genererte mengden lett bli stor.

Tidlige systemer var gjerne

- uavhengige av problemområdet. Dette høres i utgangspunktet bra ut, men i praksis er kunnskap så mangfoldig at en generell struktur ikke kan romme alle muligheter
- «ufokuserte», dvs. at istedenfor mer målrettede strategier brukes f.eks. søking og generering som tidligere forklart. I stor grad på grunn av dette siste, hadde systemene en tendens til å
- eksplodere kombinatorisk.

(For å illustrere hva kombinatorisk eksplosjon betyr, kan man tenke på

et sjakkspillende program. For hvert trekk programmet tenker framover, mangedobles antall kombinasjoner av trekk, dvs. situasjoner, som programmet må vurdere, men for å spille best mulig, må programmet tenke flest mulig trekk framover. Dersom programmet ikke har noen strategi for å skjære ned på antallet situasjoner, vil dette føre til en kombinatorisk eksplosjon.)

Som sagt var de tidlige systemene uavhengige av problemområdet, men for 10-15 år siden begynte man å se på integrasjon av metoder og kunnskapen om problemområdet. Systemene fikk da en arkitektur avhengig av problemområdet. Man la inn heuristiske regler. Man bygget inn (enkle) «slutnings-maskiner» (inference machines). Slutningsmaskinen er den delen av systemet som på grunnlag av den lagrede kunnskapen og data fra brukeren trekker slutninger, dvs. konstruerer

ny kunnskap.

De såkalte *kunnskapsbaserte systemene* man kom fram til, var karakterisert ved

- at man holdt kunnskap og slutningsmaskin separat
- kunnskapen var representert eksplisitt og symbolsk
- bruken av kunnskapen kunne gjøres «transparent» ved at systemet ga forklaringer på sine resonnementer eller ga eksemplifiseringer
- kunnskapen i systemet kunne endres. Adskilt kunnskap og slutningsmaskin gjorde dette lettere.

For å illustrere, ga Wielinga følgende analogi mellom programmering og kunnskapsbaserte systemer:

- Programmering: Algoritme + data gir et program som styrer en (virtuell) maskin. (Virtuell vil i dette tilfellet si at en kan tenke seg at programmet gjør den fysiske maskinen til en maskin som kan løse problemet som programmet er konstruert for.)
- Kunnskapsbaserte systemer: Kunnskap + data gir en kunnskapsbase som styrer en slutningsmaskin.

Førstegenerasjons ekspertsystemer representerer ofte kunnskap som slutningsregler. Et eksempel fra MYCIN, som er et system for medisinsk diagnose, illustrerer dette:

if the gram stain of the organism is negative
and the morphology is rod
and the organism grows anaerobically
then the identity of the organism is bacteroides (0.6)

(Tallet i parentes angir med hvilken sannsynlighet slutningen kan trekkes). Eksempler på *data* vil være konkrete fakta som f.eks. at en bestemt bakterie er anaerob.

Et ekspertsystem inneholder mye viten om anvendelsesområdet. Typisk er 100-5000 slutningsregler + tilhørende data. Data kan ofte være ufullstendige eller upresise, og kunnskapen unøyaktig. Ekspertsystemer er ofte i stand til å forklare resonnementet bak sine slutninger fordi kunnskapen er representert eksplisitt og symbolsk, og fordi slutningsmaskinen er enkel.

Wielinga gikk så videre med en omtale av et bestemt system, nemlig R1, som konfigurerer datamaskiner av typen VAX. R1 hjelper altså til med å velge og plassere komponenter fysisk i maskinkabinettet på grunnlag av spesifikasjoner fra kjøperen om hvilke egenskaper maskinen skal ha. R1, som siden er døpt om til XCON-XSEL, er et av de få virkelig vellykkede ekspertsystemer som er laget.

Tredje del av Wielingas foredrag var en påpeking av problemene med dagens systemer. Han satte opp følgende punkter:

- kunnskapsinnsamling. Mye kunnskap er heuristisk og ikke faktisk
- avgrensning av problemområdet

- valg av arkitektur
- «grunnhet». Kunnskapen i dagens systemer er for grunn. Det er ingen underliggende viten om *hvorfor*
- brukergrensesnitt
- konsistensen av kunnskapsbasen
- overføring av teknologien. Når systemet er konstruert, hvordan overfører en systemet fra forskningslaboratoriet til brukerne? Hvordan forklarer en virkemåten for brukerne?
- vedlikehold
- kostnader. Bak de fleste av dagens systemer ligger det minst 5 årsverk.

Vi trenger altså både metodologi og teknikker for kommende generasjoner av ekspertsystemer.

Wielinga avsluttet innlegget sitt med å snakke om metodologi. Han nevnte følgende strategier:

- strukturere oppgaven med å samle inn kunnskapen. F.eks. kan oppgaven splittes i deloppgaver
- «systemisk» angrepsmåte på prosessen med å «tappe» og tolke viten fra eksperten(e)
- utvikle dokumentasjonsmetoder
- utvikle redskaper.

Kunnskapsinnsamlingen tenkte Wielinga splittet i følgende sekvens av deloppgaver:

- en orienterende fase
- analyse av problemområdet
- problemdefinisjon
- funksjonell analyse
- oppgaveanalyse
- ekspertiseanalyse.

I tillegg kommer analyse av brukeren og problemområdets omgivelser. Disse vil influere på de tre forrige trinnene.

«Tapping» og tolking av opplysninger fra eksperten(e) er vanskelig fordi verbale data er

- vanskelige å tolke
- ufullstendige og kanskje upålitelige
- ikke-operasjonelle

Innsamling av verbale data kan gjøres ved

- intervju
- introspeksjon. (Hva gjør en i den og den situasjonen?)
- egenrapportering. (Eksperten snakker høyt mens hun løser problemet)
- dialog
- rapport ut fra tilbakeblikk på problemløsingsprosessen.

Disse metodene er listet i stigende orden når det gjelder pålitelighet og fullstendighet, men også vanskelighetsgrad når det gjelder tolking og grad av samarbeid fra ekspertens side. En må her velge riktig metode til riktig tidspunkt.

Tolkingen av de verbale data skal lede fram til en konseptuell modell for problemområdet/problemløsingsprosessen, og til en kunnskapsbase. For å kunne tolke, trenger en en tolkingsmodell, dvs. en global idé om hva slags konseptuell modell en sikter mot. Uten en slik modell, får en lite ut av de verbale data.

Tolkning kan foregå på flere nivåer, fra en intern, personlig form hos eksperten til en implementasjon på en datamaskin. De forskjellige nivåene representerer en stigende grad av abstraksjon:

- Identifisering. Denne er individuell, og det er ingen konsistens
- Konseptuell. Her arbeider man med konseptuelle relasjoner, primitive konsepter, konseptuelle modeller
- Epistemologisk. Her ser man på *typer* av konseptuelle relasjoner, *typer* av relasjoner, *typer* av kunnskapskilder
- Logisk. Formalismer
- Implementasjon. Mekanismer.

I dag spør en ofte ekspertene om å formulere sin kunnskap på de to nederste nivåene. Dette leder ofte til et misforhold mellom systemets og ekspertens måte å resonnere på. Wielinga var spesielt opptatt av å kunne starte på det epistemologiske nivået (noe han kom tilbake til i større detalj i et eget foredrag under selve konferansen).

Som et eksempel på bruk av epistemologisk analyse nevnte Wielinga et arbeid av William J. Clancey, som i 1982 utviklet systemet NEOMYCIN som en forbedring av MYCIN. Clancey tok regelbasen i MYCIN, supplerte denne med (epistemologisk) tilleggskunnskap fra eksperter, og tolket det hele på nytt. Resultatet ble et system med en totalt forskjellig slutnings- og konseptstruktur. Mens MYCIN har en «flat» regelbase, har NEOMYCIN forskjellige typer regler: Hierarki av kausale regler, data/hypotese-regler, fakta-regler. Mens MYCIN benytter omfattende søking under problemløsingen, er NEOMYCIN styrt av metaregler som innbefatter strategier for problemløsing. NEOMYCIN kan også gi «dypere» forklaringer på sine resonnementer siden reglene har begrunnelser. Begrunnede og bedre strukturerte regler gjør det også lettere å modifisere NEOMYCINs kunnskap. Alt i alt ble NEOMYCIN et mye bedre system og nærmere en ekspert enn det gamle systemet.

Toward a Computer Ethnology

Inntrykk fra et internasjonalt symposium 16.-23. september i The National Museum of Ethnology, Osaka, Japan

Jostein H. Hauge

Japan har lenge gått nye veier både når det gjelder innhenting og spredning av informasjon. Ikke minst gjelder dette den økende bruk av ulike elektroniske medier. Symposiet «Toward a Computer Ethnology» føyer seg - om enn av andre grunner - inn i dette bildet.

Takket være store pengebidrag fra en japansk fabrikkeier, *Toyosaburo Taniguchi,* har det vært mulig å opprette et fond for vitenskapelig kontakt og samarbeid med forskere fra Vesten. For fondsmidlene arrangeres det ca. 50 internasjonale symposier hvert år. Et betydelig antall utenlandske fagfolk fra de fleste vitenskapsdisipliner inviteres til ca. 1 ukes samvær med japanske kolleger, både vitenskapelig, kulturelt og sosialt. Som oftest er gruppene små, gjerne bare ti personer til sammen, hvorav fem kommer fra vestlige land.

National Museum of Ethnology i Osaka har siden 1977 fått æren å arrangere et årlig symposium i denne serien som inntil nå i hovedsak har vært viet etnologiske emner knyttet til kulturer i Asia, Afrika og Amerika. I 1984 ble denne temaserien brutt ved at datamaskinelle metoder i etnologisk forskning ble valgt som tema.

Om symposiet

De fem japanske deltakerne var med ett unntak fra nasjonalmuseet i Osaka (den femte var professor i databehandling ved universitetet i Kyoto). Gjestene fra Vesten var *professor Joseph Raben*, (USA), *Burghard B. Rieger, dr. Louis D. Burnard* (England) og undertegnede fra Norge. Professor *Antonio Zampolli* var forhindret fra å være til stede.

Computer Ethnology er et begrep som er tatt i bruk ved nasjonalmuseet i Osaka og ved starten av symposiet ble det forklart slik av lederen, *professor Shigeharu Sugita*:

1. Studium og utvikling av datamaskinelle systemer for etnologisk forskning, tilpasset både symbolsk og ikke-symbolsk informasjon.
2. Utvikling av nye etnologiske forskningsmetoder gjennom bruk av datamaskin, med vekt på modellering og simulering.
3. Sammenlignende studium av datamaskiner som kulturelt objekt. Studiet omfatter både den rent tekniske utvikling av datamaskinene og deres virkninger på det omgivende samfunn.

Gjennom foredrag og drøftinger gjennomgikk våre japanske verter emner som for eksempel den praktisk orienterte feltarbeiders nytte av databehandlingsmetoder og hvilke kunnskaper etnologer må/bør ha i edb. Stor vekt ble lagt på metoder for bildebehandling og datagrafikk i humanistisk forskning og behovet for datamaskinelle informasjonssystem som kan omfatte et bredt register av etnologiske kildetyper. En rekke av de emnene som ble tatt opp, hadde klar relasjon til utviklingsarbeid som i dag drives ved nasjonalmuseet i Osaka (jfr. nedenfor).

Gjestene fra Vesten ble slått av hvor høye mål man i Japan setter for edb-arbeidet i museer og hvor djervt man satser på å nå de mål man har satt seg. Edb-arbeidet tar utgangspunkt i prinsippet om at edb-teknologien skal være *menneskelig tilfredsstillende* dvs. at datamaskinene skal utvikles til å bli et arbeidsverktøy som passer for menneskene og som kan utfylle deres spesielle egenskaper på en naturlig måte. Det er underforstått at maskinene skal gjøre rutinearbeidet, men like viktig er det at de skal kunne samspille med fagmedarbeiderne i deres skapende, vitenskapelige virksomhet.

I dag er ikke datamaskinene på langt nær et velegnet hjelpemiddel i så henseende, ble det hevdet. De er vanskelige å programmere, er for langsomme, rigide og for lite intelligente ved søking etter og fremvising av informasjon. Særlig gjelder dette visuell informasjon. Innlegging av symboldata i datamaskiner både i form av alfabeter og i tegnform, f.eks. japanske og kinesiske tegnsystemer, er i dag for tungvint og tidkrevende.

Den strenge dommen over dagens datamaskiner kan synes overraskende når man ser på alt det datamaskinene tross alt kan utrette i dag. De resultater som våre japanske verter kunne fremlegge i form av eksperimentelle metoder og teknologier (jfr. nedenfor) gjør at jeg er overbevist om at vi kan vente store sprang fremover når det gjelder databehandlingsmetodene også i humanistisk forskning.

Gjestene fra Vesten hadde ulik faglig og profesjonell bakgrunn, og deres faglige bidrag hadde derfor forskjellige utgangspunkt. Foredragene tok opp bredden i det etnologiske grunnlagsmaterialet og implikasjonene for de etnologiske edb-metoder (Raben), kunnskapsbaserte databaser for museal forskning (Burnard), datalingvistiske semantiske metoder (Rieger), og tilbakeblikk på utviklingen av datamaskinelle metoder i de humanistiske fag med vekt på historie- og museumsfeltet (Hauge).

Som egne innslag i symposiet ble det arrangert studiebesøk på fabrikker som anvender avansert, robotisert fremstilling av elektronisk utstyr, bl.a. laserplateprodukter (Panasonic). Det ble også arrangert møter med representanter fra forskningssentra hvor humanistiske forskere arbeider med industriell design (Kyushu Institute of Design) eller med automatisk oversettelse japansk/engelsk og vice versa.

Vi fikk også innblikk i det arbeidet som pågår med å standardisere

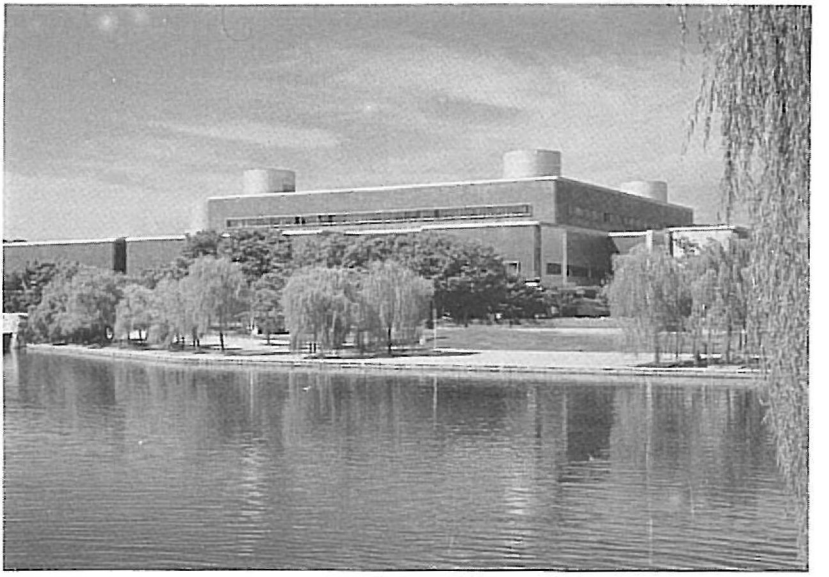
det japanske tegnsystemet kanji ved hjelp av datamaskinelle metoder og tilrettelegging av tekstbehandlings- og terminalutstyr for denne standard (ca. 3000 forskjellige kanjitegn). De utenlandske deltakerne var overrasket over hvor utbredt terminal- og tekstbehandlingsutstyr med japanske tegn er alt i dag. Gjennom kontakt med forskere ved Electrotechnical Laboratory i Ibarachi, som samarbeider med Institute for New Generation Computer Technology (ICOT) fikk vi innsyn i pågående utviklingsarbeid knyttet til femtegenerasjons datamaskiner. Vårt inntrykk var at det til i dag ikke har vært arbeidet så mye med de datalingvistiske sidene ved prosjektet (blant annet maskinell oversettelse) som den hjemlige presseomtale kan tyde på.

Bruken av datamaskinelle metoder i arkeologisk forskning ble drøftet i møter med ulike forskergrupper. Et sentralt mål er også her å skape informasjonssystemer som spenner over hele bredden i et arkeologisk grunnlagsmateriale (kart, skisser, kataloginformasjon, gjenstander, bilder, oversiktskart, satelittinformasjon).

De foredrag som ble holdt på symposiet, vil i 1985 bli utgitt i bokform. Interesserte kan få nærmere opplysninger ved henvendelse til Senteret.

National Museum of Ethnology

National Museum of Ethnology ligger vakkert til i Expo '70 Memorial Park i Osaka (jfr. bildet). Museet ble grunnlagt i 1974, åpnet i 1977 og



National Museum of Ethnology, Osaka.

er det ledende japanske museum i etnologi (også kalt kulturell antropologi). Det er integrert i det nasjonale universitetssystemet som et felles forskningsinstitutt for etnologiske studier samtidig som det er åpent for publikum (10 mill. besøkende til nå). Museet har i alt ca. 150.000 gjenstander som er utvalgt etter en nøye gjennomtenkt plan, og 9000 av dem er utstilt i rommelige utstillingshaller.

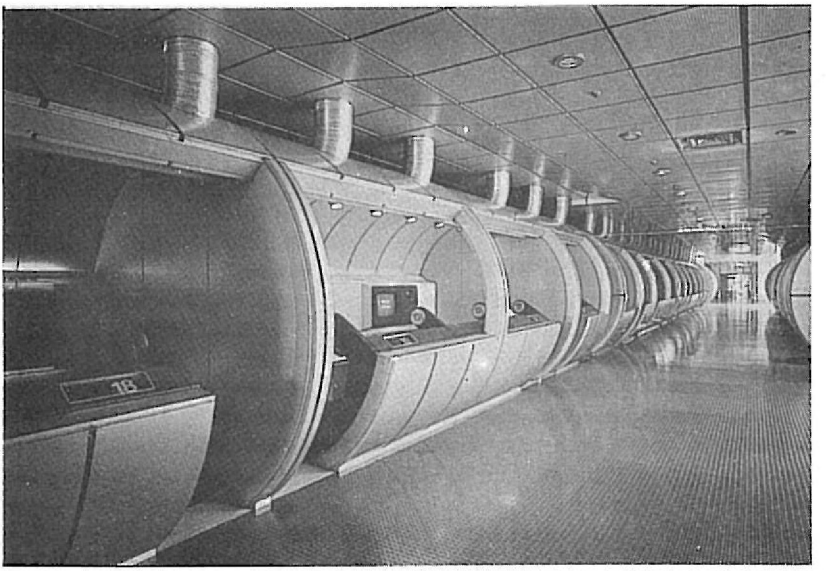
I alt er 250 personer knyttet til virksomheten, 62 av dem i forskerstillinger innenfor de 5 forskningsavdelinger som arbeidet er organisert i (studiet av asiatiske, afrikanske, europeiske og amerikanske kulturer og kulturer i Oceania). En av avdelingene har i oppgave å studere bruk av edb i etnologi.

Det finnes også et stort informasjons- og dokumentasjonssenter der referansedata til ulike typer primærmateriale er tilrettelagt for bruk i museets datasenter. Også biblioteksarbeidet er edb-basert.

Alt fra starten i 1974 har det vært satset sterkt på å ta i bruk datamaskinelle metoder i alt museumsarbeid. Museet er i dag trolig teknisk sett det mest velutviklete i verden.

Nedenfor vil jeg forsøke å gi et riss av det jeg fant mest interessant ved museets edb-arbeid, og starter med museets såkalte videotek.

Museet legger i all sin virksomhet vekt på best mulig formidling av den etnologiske viten gjennom sine utstillingsmetoder og presentasjonsstrategier. Et viktig ledd i dette arbeidet er videoteket (jfr. bildet). Videoteket er et audiovisuelt formidlingsopplegg som består av et



Videoteket har 40 visningsboder og over 700 programmer med etnologiske emner.

teknisk avansert system for lagring, behandling og visning av video- og audiokassetter. Målet er å kunne presentere på en autentisk måte ulike folkeslag, kulturer og kulturelle ytringsformer.

I dag finnes det over 700 videoprogrammer med japansk og engelsk tale hver på ca. 15 minutter og flere hundre lydkassetter som omfatter et mangfold av språk, musikk og sang. Instituttet har til disposisjon ca. 40 lytte- og visningsboder (jfr. bildet) hvor en ved hjelp av en programbok og en terminal kan velge ut bestemte programmer. Systemet er teknisk avansert ved at all lagring, behandling og visning skjer ved hjelp av roboter og overvåkes med edb.

Videoteket har til nå vært en stor suksess, og det forstår jeg godt etter å ha forsøkt det. Noe overraskende for meg var det å finne at listen over de 30 mest populære videoprogrammer for en bestemt måned viser at det mest populære programmet var «History of French Cooking».

Selv om robotsystemet har kapasitet til å håndtere opp til 2000 videokassetter, er det nå i gang arbeid med å overføre både audiokassetter og videokassetter til (noen få) laserplater. Laserbasert lagring vil forenkle driftsopplegget betydelig.

På edb-siden har museet en imponerende samling av store og mindre datamaskiner som til dels er spesielt tilpasset behandlingen av ulike typer kildemateriale. Sentralt i databehandlingen står to IBM 4341 som bl.a. utgjør kjernen i museets generelle informasjonssystem.

Edb-arbeidet startet for alvor i 1977 og har frem til 1983 i hovedsak vært rettet mot å tilrettelegge kataloger og referansedata i store databaser. Det ble opplyst at den største databasen inneholder over 1 million innførsler om bøker, dokumenter, gjenstander og annet etnologisk kildemateriale.

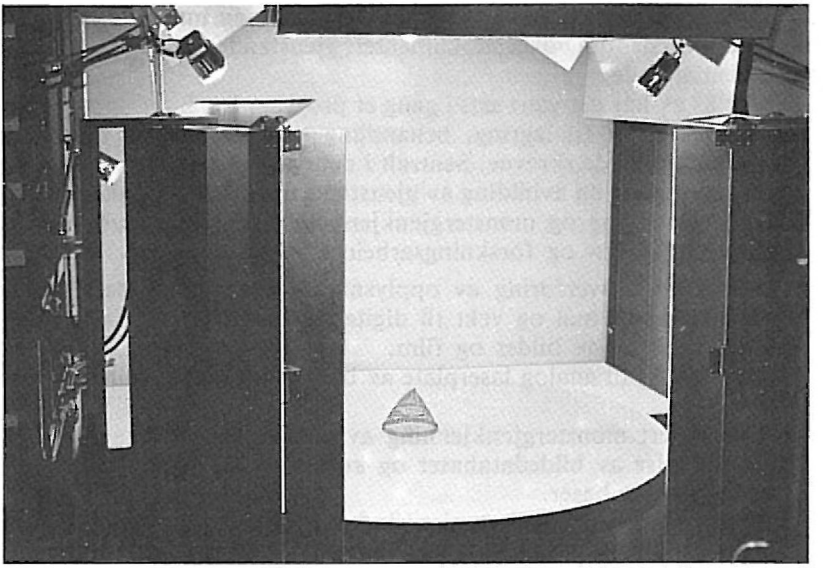
Fra 1983 av har det vært satt i gang et program for å angripe og løse problemer knyttet til lagring, behandling og presentasjon av andre primærkilder enn de skrevne. Sentralt i dette arbeidet er automatiserte rutiner for måling og avbilding av gjenstandsmateriale og ulike former for bildebehandling og mønstergjenkjenning. I dag har museet bl.a. følgende utviklings- og forskningsarbeid i gang:

1. Automatisk overføring av opplysninger om gjenstanders form, dekor, farge, mål og vekt til digitalisert form.
2. Digitalisering av bilder og film.
3. Overføring til analog laserplate av bl.a. film, bilder, musikk, tale osv.
4. Edb-basert mønstergjenkjenning av etnologiske gjenstander.
5. Opprettelse av bildedatabaser og koblingen av slike til generelle referansedatabaser.
6. Bruk av satelittdata i etnologisk forskning.
7. Bruk av avansert grafisk programvare (bl.a. shaded graphics).
8. Edb-metoder for bevegelsesanalyse (bl.a. danseformer).
9. Tale- og musikkanalyse ved hjelp av edb.

Det vil føre for langt å gi en omtale av alle disse aktivitetene. For undertegnede var ikke minst bildebehandlingsmetodene av stor interesse. Som vist på bildet har museet utviklet et gjenstandsregistreringssystem som automatiserer det grunnleggende museale arbeidet med ulike gjenstandstyper. Ved hjelp av et spesielt opptaksstudio blir gjenstandene plassert (og rotert) på en dreieskive, fotografert i sort/hvitt og farger fra ulike vinkler og høyder slik at en oppnår en detaljert digitalisert representasjonsform av gjenstanden i en datamaskin. Via laserskrivere og grafiske terminaler kan avbildinger av gjenstandene fremstilles på papir og skjerm samtidig som en kan påføre sekundær informasjon om gjenstandene (på japansk) og foreta utfyllende målinger interaktivt ved hjelp av datautstyret.

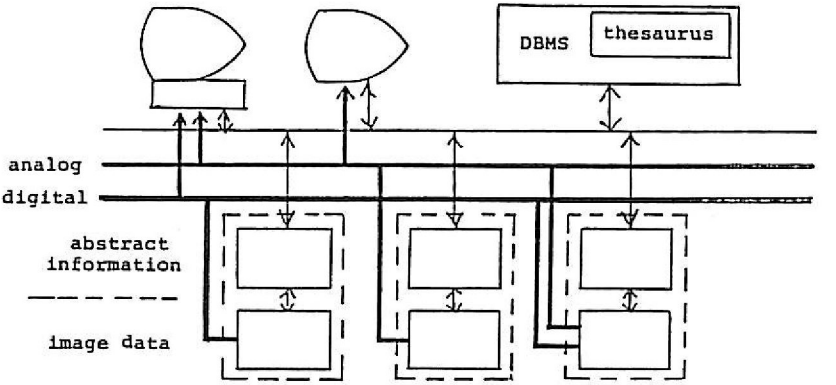
Ved å bruke de nye metodene hevder museet å ha redusert registreringstiden for vanlige gjenstander fra 1 time til 5-10 minutter, samtidig som en får en representasjon av gjenstanden i datamaskinell form.

For tiden utvikler man edb-opplegg for en effektiv fremsøkning og visning av gjenstands- og billedmateriale ved bruk av et relasjonsdatabasesystem (QBE). Dette arbeidet har for øvrig vist hvor viktig det er å ha effektive tesauri for søkeformål. Som resultat av utviklingsarbeidet håper museet å få et fullt integrert informasjonssystem hvor differensierte opplysninger om primærmaterialet integreres med databaser for lagring av avbildninger av primærkildene i digitalisert form. På sikt håper museet å kunne tilrettelegge for edb alle de primære, etnologiske



Studio for automatisk overføring av opplysninger om gjenstandenes form og mål til edb.

kildetyper. I dette arbeidet vil både teknikker for analog og digital overføring til laserplater bli brukt (jfr. diagram).



Plan for integrasjon av bildeinformasjon og referansedata.

Avslutningsvis vil jeg ønske at det kan bli finansielt og praktisk mulig for norske kulturforskere og museumsmedarbeidere å studere nærmere den faglige og edb-baserte virksomheten som pågår ved National Museum of Ethnology i Osaka. Som det kanskje fremgår av beskrivelsen ovenfor, kunne referenten gjerne tenke seg å være cicerone ved et slikt besøk.

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

Svein Engelstad, Britt Kroepelien og Espen Ore

Edb som verktøy i kunsthistorie med videre, er blitt diskutert i internasjonale kunsthistoriske fagmiljøer i flere år. Her hjemme er denne debatten kun i sin aller spedeste begynnelse.

Scuola Normale Superiore i Pisa i samarbeid med J. Paul Getty Trust arrangerte 24.-27. september 1984 *Second International Conference on Automatic Processing of Art History Data and Documents.* First International Conference ble avholdt i Pisa i 1978. Konferansen samlet ca. 300 deltakere, 3 av dem var fra Norge.

I forbindelse med konferansen er det utgitt 3 bøker, hver på ca. 500

sider. Disse beskriver diverse pågående prosjekter om anvendelse av edb på alle områder innen kunsthistorien. Bøkene skulle tjene som bakgrunn for foredrag og debatt, men ble først utdelt 1/2 time før konferansens begynnelse.

Følgende hovedtemaer ble behandlet på konferansen i denne rekkefølge:

- Leksikon
- Thesauri/Emneordlister
- Biografier
- Generelle kataloger
- Spesielle kataloger
- Bibliografier
- Dokumenter og kilder
- Integrering
- Ikonografi

I forbindelse med konferansen ble det hver dag også arrangert demonstrasjoner av edb-prosjekter innen kunsthistorie.

Diskusjonen foregikk hovedsakelig mellom på forhånd oppsatte paneldeltakere. For hvert tema var det én leder som holdt hovedforedraget, videre 2-4 paneldeltakere med korte innlegg. Der det var tid, ble det åpnet for noen spørsmål fra salen. Foredragene til lederne og innleggene til paneldeltakerne vil bli samlet i en 4. bok fra konferansen, som forhåpentligvis vil foreligge i november.

Konferanseopplegget virket ganske logisk og meningen var vel at alt skulle bygge opp til og opp under *Ikonografi*, som var tema siste dag. Dessverre fulgte ikke de forskjellige foredragene hverandre godt nok opp til at den store helheten ble skapt. Det var mange som snakket langt forbi hverandre, slik at det ikke ble skapt en reell kommunikasjon.

De fleste foredragsholderne drøftet kunsthistorisk metode på flere plan, men uten konkret anvendelse i forbindelse med edb. Nettopp kombinasjonen av kunsthistorisk metode og edb trodde man jo skulle være konferansens *hovedtema*. Når dette er sagt, må det også fremheves at en del sesjoner likevel hadde en generell interesse for dem som arbeider med lagring og gjenfinning av informasjon.

Referat fra enkelte sesjoner

Åpning

Under åpningssesjonen kunne en representant for J. Paul Getty Trust informere om at det planlegges et internasjonalt senter for arbeid innen kunsthistorie. Senteret skal plasseres i Los Angeles, men det vil ha som mål å nå frem til et større internasjonalt samarbeid m.h.p. bruk av standarder og thesauri. Det er planlagt en database med biografisk materiale fra renessansen.

Leksika

Denne sesjonen ble ledet av *G. Nencioni* fra Accademia della Crusca, Firenze. Han beskrev grundig utviklingen av leksika og ordbøker fra antikken til i dag. Deretter tok han for seg nye muligheter innen leksikonarbeid som bruk av edb kan by på. Et tradisjonelt leksikon (i betydningen ordbok) er statisk, mens språket utvikler seg. Mens en trykket bok ikke lar seg endre, kan et datamaskinlagret leksikon til enhver tid oppdateres, samtidig som det fortsatt vil kunne inneholde eldre betydninger av de enkelte ord. Et datamaskinelt leksikon gir også fritt valg av sorteringskriterier.

Til slutt beskrev Nencione en del pågående leksikonprosjekter i Italia. Bl.a. kunne han fortelle at man i Firenze arbeidet med et leksikon over tekniske termer fra det 16. århundre. Slike termer er til dels svært forskjellige fra det språket som kjennes fra renessansens litteratur.

De norske deltakerne på Second International Conference on Automatic Processing of Art History Data and Documents. F.v. Svein Engelstad, Britt Kroepelien og Espen Ore.

Thesauri

E. Svenonius fra UCLA tok utgangspunkt i det hun kalte gjenfinningsthesaurus. I et slikt thesaurus er ikke ordets betydning, men dets tilknytning til andre ord den viktigste informasjonen. Et første krav til et gjenfinningssystem som går utover fritekstsøk, er at man finner all den informasjon man søker, men ikke uvedkommende informasjon utover dette. Et thesaurus' rolle er å forminske problemene forbundet med gjenfinningsevne (recall) og presisjon.

Svenonius viste så til at thesauri først ble brukt innen realfag, der terminologien er temmelig konsis. Dette er ikke tilfelle innen et fag som kunsthistorie.

I en oppsummering ble enkelte mulige utviklinger omtalt. Svenonius mente at fritekstsøk vil bli brukt sammen med kontrollerte vokabularer og at thesauri vil være hjørnesteinen i integrerte kunnskapsbaserte systemer.

De andre paneldeltakerne kom inn på enkelte praktiske sider ved generering og bruk av thesauri. *T. Petersen* fra Art and Architecture Thesaurus, Bennington, Vermont understreket også at man i dag bruker teorier og metoder som er 20-30 år gamle.

Generelle kataloger

Panelets formann var *P. Homulos* fra National Museums of Canada. Da han ble bedt om å innlede denne sesjonen, hadde han satt i gang et prosjekt for å undersøke hvor problemene lå. Han hadde mottatt svar på et utsendt spørreskjema fra over 100 institusjoner verden rundt. Resultatet av spørreundersøkelsen var at Homulos kunne peke på noen momenter som ble oppfattet som vesentlige innen arbeidet med generelle kataloger:

- Standarder
- Prosjekter som måtte startes på nytt
- Viktigheten av informasjon
- Internasjonale interesser

Det viste seg at *mange* prosjekter arbeider med å utvikle standarder. Det er heller ikke uvanlig at prosjekter avbrytes for så å startes på nytt. Dette kan ha flere årsaker, så som at man har begynt med feil teknikk, eller det kan ha vært avbrudd i bevilgninger e.l. Homulos mente at det gikk klart frem fra undersøkelsen at det nå blir lagt mer vekt på informasjon enn for 5-10 år siden. Han ville ikke ta stilling til om det skyldes «informasjonsrevolusjonen». Selv om de fleste prosjekter er på institusjonsnivå, viser det seg at de fleste er interesserte i nasjonale og internasjonale kontakter.

Etter Homulos kunne *K. Baetjer* fra Metropolitan Museum of Art, New York, fortelle om et pågående samarbeidsprosjekt: *The Museum Prototype Project*, der 8 museer samarbeider med J.P. Getty Trust.

Målet var opprinnelig å lage et integrert system som skulle koble sammen alle informasjonsprosjektene som er sponset av Getty Trust. Dette viste seg altfor ambisiøst, og målet er nå nærmere avgrenset til å lage en katalog over de vestlige maleriene som finnes ved de 8 museene. I forbindelse med denne katalogen vil det naturligvis også måtte etableres et standard katalogformat.

Fra Ungarn kunne *László Szabó*, Magyar Nemzeti Galeria - Budapest, fortelle at man nå var i ferd med å kartlegge ungarske kunstverk. Arbeidet var nå såvidt startet, formatet for registreringsskjemaene er ennå ikke helt klarlagt. I begynnelsen skal man lagre opplysninger om 15-20 tusen kunstverk på en ungarsk mikromaskin der innmatingen foregår ved hjelp av hullkort.

Dokumenter og kilder

Denne sesjonen ble ledet av *P. Barocchi* fra Scuola Normale Superiore, Pisa. Hun nøyde seg stort sett med å foreta en oppsummering av det arbeidet med kilder som foregår i Italia. Hun poengterte bl.a. behovet for samarbeid mellom kunsthistorikere og folk fra arkivverket.

Resten av paneldeltakerne gikk inn for at det måtte arbeides med felles standarder, og enkelte kom også inn på de problemer som er forbundet med å bruke f.eks. utstillingskataloger som kilder siden de nettopp er preget av svært forskjellig terminologi.

De øvrige sesjoner

Svært meget av konferansens tid gikk med til å diskutere behovet for standarder og standardisering. Det er åpenbart at for å drive en effektiv edb-behandling av et hvilket som helst humanistisk materiale, er der visse standarder man bør enes om. Man bør rett og slett snakke samme språk. Diskusjoner om behovet for standarder hørte derfor, etter vår mening, ikke hjemme på denne konferansen.

Enkelte prosjekter som ble presentert på konferansen

ICONCLASS er et kunsthistorisk, ikonografisk klassifiseringssystem spesielt for vestlig kunst. Iconclass er utviklet ved Universitetet i Leiden i Nederland over en periode på ca. 35 år. Ansvarlig for Iconclass er for tiden *Leendert D. Couprie.*

Selve klassifiserings-systemet er utarbeidet «manuelt», det vil si etter tradisjonelle forskningsmessige metoder. Registrene over de tusenvis av koder er derimot ført ved hjelp av edb. Kodingssystemet gjør materialet som bearbeides med Iconclass spesielt velegnet for databehandling: Temaet *Hyrdenes tilbedelse* blir redusert til følgende 5 tegn *73 B 25*. Ved å kode data slik, får man en komprimert form som ikke legger beslag på for stor lagerplass. Videre kan man stille en rekke svært spesialiserte spørsmål ved hjelp av kommandoer som: «AND, OR, NOT» etc. og

kodene, som både er tall, bokstaver og parenteser. Kodene kan kombineres på mange forskjellige vis. De viktigste inndelingene er: 1 Religion og magi, 2 Natur, 3 Mennesker, 4 Samfunn og sivilisasjon, 5 Abstrakte idéer og begreper, 6 Historie, 7 Bibelen, 8 Litteratur, 9 Klassisk mytologi og antikkens historie.

Iconclass er publisert i 17 bind som omhandler klassifiseringen, anvendelsen av klassifiseringen og alfabetiske registre til klassifiseringen. Registerdelen er også tilgjengelig som en database som kan implementeres på andre systemer. Blant de som bruker Iconclass kan nevnes Harvard College, Marburg Index, Witt Library etc.

På området for internasjonale kunsthistorisk-bibliografiske registre finnes det to ledende prosjekter: *RAA* (Répertoire d'Art et d'Archéologie) i Paris og *RILA* (Répertoire International de Literature d'Art) i Massachusetts.

RAA har utkommet siden 1910. RAA tar for seg vestlig kunst og kristen arkeologi i Middelhavs-området mellom 2. og 20. århundre. En større edb-omlegging fant sted i 1973. RAA ligger nå på en database med muligheter for å søke på ord i titlene og i sammendragene av

Crispin de Passe d.e. (1565-1637): Neptun passerer fjellet der Deukalions ark ble liggende da floden trakk seg tilbake.
Dette bildet blir kodet slik i Iconclass-sytemet: 92 H 17 1 : 91 E 72
9 står for scener fra klassisk mytologi og antikk historie
91 viser til gjengivelser fra klassiske skapelsesmyter
92 viser til gjengivelser av olympiske guder
91 E 7 er Deukalions flod
92 H 17 er typiske fremstillinger av Neptun (Poseidon)

artiklene.

RILA ble startet i 1973 som et register for samtidig kunstlitteratur. Utover registrering lages også sammendrag av de aktuelle artiklene. Opplysningene blir lagret i en database og gjøres tilgjengelige både i publisert, trykt form og ved on-line kommunikasjon med databasen. RILA dekker Europa fra det fjerde århundre og Nord-Amerika fra det sekstende århundre, begge steder frem til dags dato.

Det er en viss forskjell i strukturen på disse to registre, eksakt hvilke skal vi ikke komme inn på her, men det burde være en mulighet for samarbeid. Nettopp spørsmålet om formalisert samarbeid ble for første gang tatt opp på denne konferansen. På visse områder utfyller disse to registre hverandre, og på andre overlapper de hverandre, så det gjøres en del dobbeltarbeid.

L. Heusinger fra Universitetet i Marburg gjorde rede for et annet og meget interessant prosjekt - Marburger Index. Det er en billedkatalog, publisert på mikrofilm og ikke i bokform. Den består i dag av 560.000 illustrasjoner - 30.000 katalogtekster - 6 forskjellige registre og en databank.

Som utgiver av Marburger Index står Bildarchiv Foto Marburg. Dette er Philippuniversitetet i Marburgs billedarkiv, og ble grunnlagt i 1913 for å utstyre universitetets kunsthistoriske avdeling med billedmateriale. I 1961 ble Bildarchiv Foto Marburg utpekt av det tyske vitenskapsråd til hovedsete for all dokumentasjon av tysk kunst. I og med et så utvidet ansvarsområde begynte ledelsen å orientere seg om mulighetene for et mer effektivt informasjonssystem.

For kunsthistorikeren er det som kjent selve kunstverket som er det viktigste. Men siden dette ikke alltid er tilgjengelig, vil en reproduksjon være det beste alternativ. Denne gir jo langt mer variert informasjon enn bare en verbal beskrivelse. Med dette som utgangspunkt fikk man i 1975 ideen til Marburger Index og tenkte seg denne katalogen som begynnelsen på et nytt og mer omfattende informasjonssystem.

Det er imidlertid velkjent at kunst- og kulturhistoriske samlinger ikke har solid nok økonomi til selv å kunne influere på utviklingen av grunnleggende teknologi. Man valgte derfor å utforske de allerede eksisterende metoder. Dette ble gjort, ved hjelp av prøve- og feilemetoden, og resultatet ble et mikrofilmarkiv som erstatter trykk, lysbilder og indeks. Overføring fra fotografi til mikrofilm ble gjort ved hjelp av datamaskin.

Men billedmaterialet var meget stort - og jo større det ble, desto vanskeligere ble det å finne frem til det enkelte verk. Og behovet for et effektivt søkesystem tvang seg hurtig frem. Lagring av millioner av data er med dagens teknologi ikke noe stort problem. Derimot er det en kjempeoppgave å gi dataene en logisk strukturert form, slik at en opplysning bare behøver å forekomme én eneste gang for å kunne trekkes inn i andre sammenhenger. *MIDAS* (A Method of Indexing and Documenting Art Systematically) er et system som ble utviklet i

tilknytning til Marburger Index. Hittil er bare de deler som er nødvendig for visuell kunst og kunstnere utviklet. Takket være tre millioner DM i støtte fra Stiftung Volkswagenwerk kunne 100.000 data fra Marburg Index bli samlet og systematisert. Siden 1983 er nye 30.000 arbeider blitt registrert. Teksten ved hvert stikkord slutter med referanser til de tilsvarende fotografier på mikrofilm. Alle tekstene blir lagret i en databank og kan kalles frem og bli brukt on-line ved terminal. Videre blir tekstene overført ved datamaskin til seks forskjellige kataloger som også er blitt publisert på mikrofilm. De fire viktigste er en primær ikonografisk katalog, en sekundær ikonografisk katalog, en portrettkatalog og en kunstnerkatalog. Hvert år blir ytterligere 20.000 data fra Marburger Index samlet og systematisert, og katalogene blir tilsvarende utvidet. Prosjektet har etter hvert fått økonomisk støtte fra mange andre institusjoner og samlinger.

Marburger Index er forskjellig fra de fleste andre edb-prosjekt ved at det i utgangspunktet ikke var ment å skulle løse bare én enkelt institusjons problem. Allerede fra begynnelsen ble prosjektet sett på som en samlet publikasjon for hele det kulturelt-akademiske marked. De første mikrofilmene ble allerede solgt tre måneder etter at prosjektet startet. Videre var hensikten å effektivisere produksjonen av reproduksjoner, ikke å samle data til en systemutviklingsprosess, og dermed gi resultater hurtigst mulig. MIDAS er dessuten resultatet av mange års møysommelig analyse av de humanistiske behovene, og er derfor fullstendig uavhengig av program- og maskinvare. Videre er det de samme personer som sto bak utviklingen av Marburger Index som var ansvarlige for MIDAS, og dermed unngikk prosjektet de ellers så vanlige friksjonene mellom faglig og teknisk nivå.

I dag er Marburger Index produsert i mer enn 200 eksemplarer, og solgt til flere land.

R. Holt fra On-Line Computer Systems, Maryland, var paneldeltaker i sesjonen for integrasjon og var den eneste i panelet som hadde datateknologisk utdannelse. Han ga inntrykk av å kjenne kunsthistorikerens problemer etter å ha deltatt i flere store kunsthistoriske prosjekt, bl. a. «The Census of antique works of art and architecture known to the Renaissance», som ble demonstrert på kongressen.

Han pekte på analogien mellom ordene og syntaksen i et vanlig språk med grunnreglene for et databasespråk. Ønsker man et integrert informasjonssystem, er det derfor helt nødvendig å finne frem til en felles språklig plattform. Esperanto var et eksempel på et slikt kunstig språk. Men, hevdet han, når dette forsøket ikke ble særlig vellykket, skyldes det at språket utelukkende ble brukt internasjonalt. Det var derfor en forutsetning for et felles dataspråk, mente han, at det også ble tatt i bruk i alle nasjonale prosjekt. Først da kunne man nærme seg informasjonsintegrering.

Alt i alt var det en meget interessant konferanse, innvendingene til tross. Det er tydelig at det ligger meget ugjort arbeid innen kunsthistorie

og edb.

For videre interesserte kan vi henvise til bøkene utgitt av Scuola Normale i Pisa, samt til en fyldig rapport vi planlegger i NAVFs EDB-senters rapportserie. Vi håper at rapporten vil være tilgjengelig ved årsskiftet.

Svein Engelstad og Britt Kroepelien er hovedfagsstudenter i kunsthistorie ved Universitetet i Bergen.
Espen Ore er edb-konsulent ved NAVFs EDB-senter for humanistisk forskning.

Besøk ved Istituto di Linguistica Computazionale - CNR, Pisa

Espen S. Ore

Undertegnede besøkte Istituto di Linguistica Computazionale 28. september 1984. Instituttet har et arbeidsområde som spenner fra bruk av edb innen klassisk filologi til kunnskapsbaserte databasesystemer. Blant de som tok seg tid til å informere om, og demonstrere forskjellige prosjekter, var *A. Bozzi, G. Ferrari* og *I. Prodanof.*

Klassisk filologi

Innen dette fagfeltet var flere prosjekter i gang. Et av dem, som for tiden blir viet mest oppmerksomhet ved instituttet, hadde som mål automatisk lemmatisering av latin. Lemmatiseringen foregår både gjennom morfologiske og grafiske analyser. Den morfologiske analysen gjør bruk av lister over pre-, in- og suffikser som gir uttrykk for ordklassetilhørighet og grammatisk form. Slike lister er sannsynligvis mer effektive overfor et språk som latin enn overfor språk som f.eks. norsk og engelsk der reglene for ordkonstruksjon og fleksjon ikke gir den samme entydighet som man finner på latin.

Ved siden av disse listene brukes det også lister som inneholder varianter av samme morfem. En slik liste kan f.eks. inneholde morfene a-, ab-, af-. Slike lister viser ikke bare et morfems realisasjon i forskjellige fonetiske kontekster, men også variasjon over tid. På denne måten skal det samme programmet kunne brukes til lemmatisering av både klassisk latin og middelalderlatin.

Programmet benytter lister med leksemer. Alle disse tre typer lister inneholder også koder for oppslagselementene som angir hva slags

leksikalske eller morfologiske ledd de kan koples sammen med.

Lemmatiseringsprogrammet skal brukes på tekster som leses inn ved hjelp av OCR. Bozzi kunne fortelle at han har prøvelest med en av de nyeste Kurzweil OCR-maskinene. Han var fornøyd med resultatene selv for bøker som var utgitt så lang tid tilbake som 1780-årene.

Annen virksomhet

Ved instituttet er det også én gruppe som arbeider med parsing av italienske tekster, og én som arbeider med kunnskapsrepresentasjon med tanke på databasesystemer.

Arbeidet med parsing er nå i det man kaller tredje fase. Arbeidet har hele tiden bygget på ATN-basert grammatikk. (ATN: Augmented Transition Network(s). ATN er utviklet fra matematikkens endelige tilstandsautomater. ATN-grammatikker er for tiden en av de mest brukte metodene for analyse av naturlig språk.)

Den første parseren som ble konstruert hadde en meget stor grammatikkdel. Den ble prøvd på vitenskapelige tekster og på science fiction litteratur. Hensikten med denne testen var bl.a. å undersøke om det var behov for forskjellige parsing-strategier for forskjellige typer tekst. Resultatene syntes å vise at det var et slikt behov. Alle de tre parserne er skrevet i Magna-Lisp.

Versjon nr. 2 bygget på en meget stor ordbok med ca. 100.000 forskjellige ord eller 1.500.000 forskjellige former. I tillegg gjorde man bruk av en frekvensordliste. Hovedordboken var for stor til å være direktekoplet, så parseren måtte brukes i satsvise kjøringer. Senere ble det laget en forenklet ordbok med ca. 80.000 vanlige former. Denne lot seg bruke interaktivt, og hvis parseren under en kjøring kom over en form som ikke fantes i ordboken, kunne den legges inn med én gang. Denne parseren hadde problemer med sideordning (konjunksjon), men den taklet underordning (subjunksjon). Parseren ble betraktet som lite effektiv idet den lett ga store mengder utskrift i form av alternative analyser. Det er også laget en enklere, interaktiv versjon av denne parseren beregnet på undervisning. Hensikten er at studenter skal få øvelse i lingvistikk ved å legge inn egne ordbøker og grammatikker.

For tiden arbeides det med en tredje parser. Det er meningen at denne skal koples sammen med et kunnskapsrepresentasjonssystem som også er under utvikling ved instituttet.

Datamaskinen – språkforskernes store utfordring i vår tid

ANLA-symposiet 1984

Bergen 12. og 13. oktober 1984

Jostein H. Hauge

Norsk forening for anvendt språkvitenskap arrangerte 12. og 13. oktober sitt årlige symposium i Bergen med hovedtemaene *Datalingvistikk* og *Prosodi i fremmedspråkundervisningen*. Ca. 40 deltakere fra universiteter og høgskoler deltok på symposiet. Nedenfor vil jeg gi en omtale av de foredrag som ble holdt om datalingvistiske emner.

Professor Helge Dyvik, Universitetet i Bergen, konstaterte at det i dag skjer en eksplosiv faglig utvikling i skjæringsfeltet mellom lingvistikk og datavitenskap. Interessen manifesterer seg i utvikling av nye datalingvistiske teorier, prosjekter for automatisk analyse av ulike språk, og utarbeidelse av praktiske anvendelsesområder av datateknikk der lingvistisk viten inngår. Også konferanse- og publiseringsvirksomheten øker i raskt tempo.

I sitt foredrag kom Dyvik inn på teorier for automatisk morfologisk, syntaktisk og semantisk analyse som er utviklet, og det forskningsarbeidet som pågår i Bergen gjennom et samarbeid mellom Institutt for fonetikk og lingvistikk og NAVFs EDB-senter for humanistisk forskning.

Etter Dyviks oppfatning har humanistene – det vil i dette tilfelle si språkviterne – i dag en viktig oppgave i å påvirke utformingen av de allmenne, brukerrettede datasystemer i samfunnet.

Det vises for øvrig til Dyviks foredrag fra konferansen som blir gjengitt i noe bearbeidet form i dette nummer av Humanistiske Data.

Kolbjørn Heggstad, firmaet Logos, Bergen konkretiserte i sitt foredrag hvilke forventninger de kommersielle datafirmaer i dag har til lingvister og spesielt datalingvister.

I raskt tempo nærmer vi oss den situasjon at datamaskinen finnes på de aller fleste arbeidsplasser og i de fleste hjem. Parallelt med denne utvikling finner vi at datamaskinene er gått over fra å være elektroniske regnemaskiner til å bli ordbehandlingsmaskiner. Alt i dag finnes det maskiner som styres ved hjelp av kommandoer i form av tale, og datamaskiner som gir resultatet av databehandlingen i form av syntetisk tale. Denne utviklingen vil fortsette. Over hele verden arbeides det i dag med et bredt spekter av støtteprogram for ord- og tekstbehandling og nye former for kontakt mellom brukerne og maskinene.

Datamaskinenes store fortrinn i arkiv- og kontorarbeid fører til at man i offentlig og privat virksomhet sitter med svære tekstmengder som man også skal gjenfinne informasjon i for ulike formål. For å klare det, må vi ha lingvistisk baserte tekstsøkemetoder og tekstkompresjonsmetoder i tillegg til nye metoder for kunnskapsorganisasjon og informasjonssøking. Her stiller man i dag i datafirmaene store forventninger til forskning innen fagfeltet kunstig intelligens, men også til mer direkte datalingvistisk fagarbeid.

Den økende internasjonalisering generelt og den amerikanske (og engelskspråklige) dominans på det teknologiske området gjør det nå mer og mer viktig å utvikle metoder for datastøttet oversettelse og databaserte løsninger for terminologiarbeid.

Heggstad kom inn på hvordan det kan opparbeides en bedre kontakt mellom datafirmaer og akademiske språkmiljøer i vårt land. Han la vekt på at datafirmaenes utgangspunkt er å være *resultatorientert*. Det betyr at språkmiljøene, for å etablere kontakt og vinne tillit, *først* må vise at de kan bistå datafirmaer og programvarehus med løsningen av (teoretisk sett) enkle språkproblemer som bransjen sliter med i dag. Eksempler kan her være lingvistisk baserte orddelingsprogram og ulike støtte- og diagnoseprogram i forbindelse med tekstbehandling. Først når relevansen av datalingvistisk arbeid på denne måten er demonstrert på enkle problemområder, vil datafirmaene satse på et utvidet samarbeid med datalingvistikkmiljøene. Da kan også grunnen være beredt for et økonomisk samarbeid om mer langsiktig utviklingsarbeid i de akademiske forskningsmiljøene.

Fremveksten av fagfeltet anvendt språkvitenskap var temaet for foredraget til *Lars S. Evensen, Universitetet i Trondheim*. Anvendt språkvitenskap kan ses på som enten en *teoridrevet* aktivitet eller en *problemdrevet* virksomhet. Ifølge foredragsholderen var utgangspunktet for anvendt språkvitenskap at en ønsket å finne frem til praktiske anvendelsesområder for de lingvistiske teorier, jfr. fokuseringen på fremmedspråksundervisning. Slik sett er anvendt språkvitenskap, iflg. Evensen, å anse som «det rørledningssystemet som lingvistisk teori bruker for å nå bakken».

Men mange kom i årenes løp i opposisjon til det syn at praktiske problemer alltid kan/skal løses ved hjelp av lingvistisk teori. Slike synspunkter førte til at de språklig definerte problemene i seg selv kom i fokus. Målet ble her å hente teoretisk og annen fagkunnskap fra lingvistikken, men også utenfor, for å løse de konkrete språklige problemer. Teoretisk basis ble derved en nødvendig, men ikke tilstrekkelig betingelse.

I dag er det naturlig å se på datalingvistikken som en del av anvendt språkvitenskap. I dette faget er en ikke minst opptatt av performansstudier, d.v.s. studier av språkproduksjon. Målet er å produsere - og analysere - automatisk naturlig språk i en datamaskin. Datalingvistikken vil ha stor betydning for viktige lingvistiske arbeidsfelt i fremtiden,

f.eks. datamaskinstøttet undervisning og produksjon av læremidler, leksikografisk arbeid og utvikling av ekspertsystemer.

Også innenfor et klassisk område for anvendt språkvitenskap som språklaboratoriene, vil datalingvistikken få stor betydning. Fremtidens språklaboratorier vil i høy grad være basert på datateknikk. De vil gi anledning til datamaskinbasert fremstilling av både bilde, lyd, tale og skrift i ulike kombinasjonsløsninger og for ulike undervisningsformål.

Evensen beklaget sterkt den defensive holdning som humanister viser i sin kontakt (eller mangel på kontakt) med de datatekniske miljøene. Etter hans mening har språkvitenskapen kunnskap som også er av stor verdi for å kunne skape bedre programsystemer for databehandling i fremtiden.

Helge Lødrup, Universitetet i Bergen viste bl.a. til at japanerne i sitt konsept for den såkalte 5. generasjonsmaskinen har prioritert høyt felt som maskinoversettelse og ekspertsystemer og har satt seg som mål innen 1990 å utvikle datamaskiner som også kan foreta logiske slutninger.

Tradisjonelt har en kommunisert med og utnyttet datamaskinens ressurser ved hjelp av kunstige språk (programmeringsspråk), som er strukturert etter mønster fra naturlige språk. Fremover vil naturlige språk bli utnyttet direkte for å styre maskinene og for å hente ut informasjon som er lagret i dem.

Lødrup ga et oversiktsbilde av datalingvistisk forskning og la vekt på at utgangspunktet for slik forskning er å etterprøve de hypoteser man implementerer i datamaskinen i form av eksplisitte grammatikker som bygger på bestemte lingvistiske teorier. I denne forskningen blir det sentralt å studere språket som *prosess* ved hjelp av et analysesystem, ofte kalt en parser. Målet er å foreta en automatisk syntaktisk, morfologisk og semantisk analyse. Lødrup ga en rekke eksempler på hvordan en må lage eksplisitte regler for å dekke de leksikalske, semantiske og strukturelle tvetydigheter som finnes i naturlig språk. Ofte vil en analyse av en setning bryte sammen dersom en blir nødt til å foreta en ord-for-ord analyse. Analysestrategier for å kunne vurdere mange tolkingsmuligheter parallelt, og strategier som setter en i stand til å revurdere tidligere analyseresultater i lys av øket kunnskap, blir derfor sentrale i datalingvistikken.

I tillegg til de mer teoretiske presentasjoner ble det gitt en serie praktisk orienterte foredrag:

Gulbrand Alhaug, Universitetet i Tromsø gjennomgikk en prosjektplan for datastøttet oversettelse mellom bokmål og nynorsk. Forfatterens utgangspunkt var at forskjellen mellom bokmål og nynorsk var så liten (på ordnivå) at det vil være mulig med enkle metoder å ta datamaskinen til hjelp i oversettelser fra bokmål til nynorsk og omvendt. Studier av de 600 mest frekvente ord i en løpende tekst på 700.000 ord viser at av disse 600 ord (som utgjør 60% av hele tekstmengden) er bare 21% forskjellige i bokmål og nynorsk. Utgangs-

punktet er da de tradisjonelle former. Dersom en legger de såkalte tilnærmingsformene til grunn, vil prosenten bli 12.

Alhaug skisserte et datastøttet oversettelsessystem der oversetteren overvåker og griper inn i den oversettelsen som datamaskinen utfører. Ved også å legge inn i systemet enkle regler for syntaktisk omforming vil en kunne oppnå en betydelig gevinst. Et system som det skisserte, vil ikke minst ha relevans for læremiddelproduksjon der det er behov for parallellutgaver i begge målformer. Den ønskverdige situasjon hadde for øvrig, iflg. foredragsholderen, vært at lærebøkene først ble skrevet på nynorsk og så oversatt til bokmål. Nynorsk syntaks er som oftest gangbar i bokmål også, men det motsatte er ikke alltid tilfelle.

I diskusjonen knyttet til foredraget ble det bl.a. uttrykt frykt for at slike oversettelsessystemer kunne føre til en tilstivning av forfatterspråket dersom en både i syntaks og ordvalg tar hensyn til de syntaktiske strukturer og ord som lettest lar seg oversette i en datamaskin.

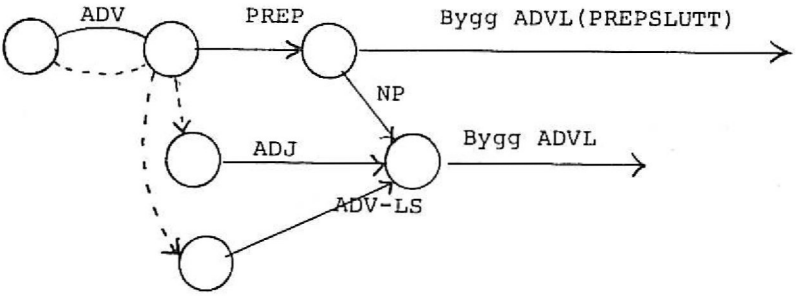
Anne Golden og *Anne Hvenekilde, Universitetet i Oslo* redegjorde for hvordan brukere med liten edb-erfaring kan få mye hjelp selv av enkle, generelle databehandlingsprogrammer og programmer for språkbehandling. Golden viste hvordan en ved hjelp av ulike ordliste- og konkordansprogrammer hadde kunnet analysere et korpus med innvandrerspråk på nye måter i forhold til en konvensjonell fremgangsmåte.

Hvenekilde tok for seg hvordan hun hadde utført språkdiagnostiske studier av tyrkiske elever i Norge ved hjelp av datamaskin. Ikke minst her kom det frem at brukernes fantasi ved anvendelse av programmene kan gi uventede, verdifulle resultater. For eksempel nevnte foredragsholderen hvordan hun selv hadde utnyttet programmer opprinnelig utviklet for regnskapsformål i sitt arbeid med analyse av tyrkiske elevers norske stiler.

Stig Johansson, Universitetet i Oslo ga en orientering om de metoder som er utviklet i forbindelse med grammatisk merking av det én million ord store LOB korpus som består av britisk-engelske tekster av ulike slag. Arbeidet er et samarbeidsprosjekt mellom University of Lancaster, England, Britisk institutt, Universitetet i Oslo og NAVFs EDB-senter for humanistisk forskning i Bergen. Arbeidet med automatisk merking av ordklasser til ord i løpende tekst har vist hvilke store problemer som er forbundet med å lage formaliserbare regler for formålet og hvor vilkårlig grensedragingen mellom ulike ordklasser er i de grammatiske tradisjoner.

Målet med den grammatiske merkingen er bl.a. å kunne utgi frekvensordlister med utgangspunkt i grunnformer av ordene, konkordanser av ord med grammatikkoder, oversikter over ordsammenstillinger, og statistiske oversyn over grammatiske kategorier. Det grunnlagsmaterialet som skapes, vil også ligge til rette for syntaktiske studier, avanserte ordstavingsprogrammer og stilistiske diagnoseprogrammer.

Knut Hofland, NAVFs EDB-senter for humanistisk forskning hadde fått i oppgave å vise hvordan en datalingvistisk syntaktisk analyse av



Et eksempel fra Svein Lie og Knut Hoflands parser for norsk. Eksempelet viser en del av en grammatikk (for adverbialer) framstilt som et nettverk.

norsk foregår. Utgangspunktet var den parser som han har utviklet sammen med Svein Lie, Universitetet i Oslo og som bygger på prof. Martin Kays såkalte Chartparser fra 1974. Gjennom eksempler gjennomgikk Hofland hvordan syntaktiske regler i største detalj må foreligge for at datamaskinen skal kunne utføre en automatisk analyse av selv enkle setninger.

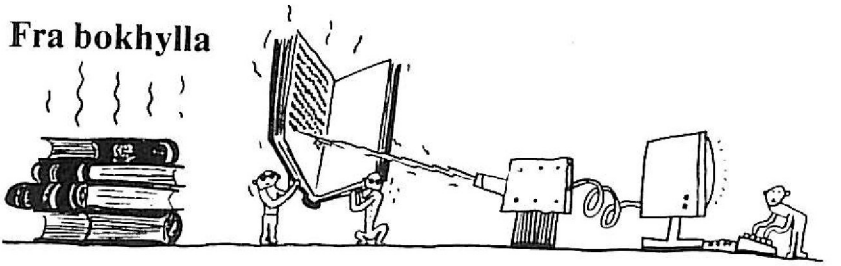
For tiden er leksikalsk-funksjonell grammatikk den grammatikkmodellen som det arbeides med i Bergen. LFG-grammatikken er utviklet av lingvister og informatikere ved Stanford University og Xerox' forskningssenter i Palo Alto. Et av fortrinnene ved denne grammatikken er at den er et velegnet hjelpemiddel for lingvister uten edb-kunnskap som ønsker å skrive grammatikkfragmenter som kan testes i en datamaskin.

Gjennom eksempler anskueliggjorde Hofland til slutt hvordan en med moderne, én-bruker arbeidsstasjoner, f.eks. såkalt LISP-maskiner, kan kommunisere direkte med grammatikken, parseren og ordboken i et slikt system.

Symposiet kan tas som et klart tegn på at datalingvistiske emner for tiden får øket oppmerksomhet også i vårt land. Til i dag har bruk av datamaskiner i språkforskning i hovedsak hatt karakteren av datamaskinstøttet språkforskning. Det nye er at det også er i emning forskningsmiljøer som setter datamaskinen i sentrum. Målet er å få datamaskinen til å simulere viktige deler av den menneskelige, språklige adferd. Det skal bli spennende å se om den økende faglige interessen som vi nå ser kan videreutvikles til å bli kompetente fagmiljøer for forskning og undervisning i datalingvistikk i vårt land.

Interesserte vil senere kunne få kjøpe et kompendium med foredragene eller resymeer av dem. Henvendelse til *prof. Bjørn Stålhane Andrésen, Engelsk institutt, HF-bygget, Sydnesplass 9, 5000 Bergen.*

Fra bokhylla



Robert L. Burke: CAI SOURCEBOOK Background and Procedures for Computer Assisted Instruction in Education and Industrial Training. Prentice-Hall, 1982. Ca. 200 s. Pris: 164 kroner.

Burkes bok om CAI (Computer Assisted Instruction), på norsk skal det visst hete DAL (DatamaskinAssistert Læring), gir en svært god innføring i emnet. Boken bærer preg av at forfatteren har lang pedagogisk erfaring. Hans engasjement i CAI begrunner han slik: «One of the reasons I am in the field of individualized instruction is that I have witnessed the failure of traditional educational methods in the lives of people close to me. I think CAI is one of the most elegant alternatives to traditional methods we have.» Han legger her vekt på den nye mulighet som gis lærerne for å tilpasse stoffet til den enkelte elevs situasjon.

Boken er imidlertid ikke ment som en diskusjonsbok om fordeler og ulemper, men som en opplæringsbok for forfattere av kurs i datamaskinassistert læring. Burke drøfter mulighetene som gis for en forfatter av opplæringsprogrammer. Han setter «CAI» inn i en historisk og en pedagogisk/-teoretisk sammenheng, han drøfter det avgjørende vendepunktet for «CAI» i og med mikromaskinene og han gir det terminologiske begrepsapparat som er nødvendig for å diskutere de forskjellige teknikker og muligheter innen «CAI». Kapitlene om «CAI Frame Protocols» og «CAI Lesson Design» fant jeg ytterst informative og stimulerende. Bokens praktisk/økonomiske tilsnitt kommer tydelig frem i nest siste kapittel, «Cashing In». Her står en del nyttig om kursforfatteres «kontraktsmessige forhold».

Boken sier lite om kjente systemer som PLATO og PILOT, men har en instruktiv redegjørelse for de tre produksjonsteknikkene som er tilgjengelige for kursforfattere. Man kan programmere selv i et *generelt programmeringsspråk*, man kan bruke et *forfatterspråk*, som de nettopp nevnte, eller man kan prøve et *«CAI authoring system»* som har den fordel at systemet v.h.a. menyer leder kursforfatteren gjennom hele prosessen slik at forfatteren bare gir inn de faglige opplysninger. Vedkommende behøver ikke selv forstå eller anvende noe eget språk. Ulempen er at forfatteren blir bundet til den modell som systemet bygger på. Slike systemer er der få av, Burke nevner CDS 1, men han regner med at flere vil komme og at der vil bli «giant improvements in

these systems in the immediate future.» Datamaskinassistert læring er bare i sin begynnelse, og enten man selv vil produsere kursvare, skal evaluere kursvare eller diskutere teknikken som sådan så har man nytte av Burkes bok. Den er lettlest, oversiktlig og informativ og bærer preg av å være skrevet av en mann med egenerfaring.

Av norsk programvare på området kan nevnes DAMBU prosjektet ved Agder distriktshøyskole med *forfatterspråket* ALPRO. Et forfatterspråk antas å redusere utviklingstiden for et kurs med en faktor på 6. Forfattersystemer med en faktor på 12. Siden det tar ca. 300 timer å utvikle et én-times kurs så er det klart at slike systemer vil få avgjørende innflytelse på kursutviklingen. Bokens drøftelse av de tekniske muligheter (lyd, grafikk, bilde m.m.) er perspektivrik i denne sammenheng, men jeg ble godt skremt av beskrivelsen på s. 176f av occulometeret som skulle måle elevens fysiologiske aktivitet under kursgjennomgåelsen og måleresultatene skulle tjene som grunnlag for å utdele kursbevis!

Roald Skarsten
Edb-seksjonen ved Hf-fakultetet, Universitetet i Bergen

Gunnel Engwall: Vocabulaire du roman francais (1962-1968). Almqvist & Wiksell, Stockholm 1984. 427 pp. + 43 microfiches. Price: SEK 250.

This book presents the results of an investigation into the vocabulary of twenty-five best-selling novels published in France during the period 1962-1968. A systematic random sample was drawn from these novels, resulting in a corpus of 500,096 running words. The introduction describing material and principles is followed by six extensive word lists and five tables. The dictionary also includes 43 microfiches containing a complete concordance of the corpus in KWIC format. A summary of the introduction in English is provided.

K.N.

MELDINGER

Office for Humanities Communication

The British Library har opprettet Office for Humanities Communication ved University of Leicester. Formålet for denne institusjonen er å orientere humanistiske forskere om relevante utviklinger innen kommunikasjon, med særlig vekt på ny teknologi og dens bruksområder innen humaniora. Konkrete oppgaver består av undersøkelser om f.eks. universitetskurs i dette feltet og bruk av edb hos humanister, arrangement av seminarer, demonstrasjoner m.m., og utgivelse av meldingsbladet «Humanities Communication Newsletter.» Hovedvekten blir lagt på utviklingen i Storbritannia, men medarbeiderne er også interessert i kontakt med utenlandske enkeltpersoner og institusjoner.
Adresse: *Office for Humanities Communication, University of Leicester, Leicester LE1 7RH, England.*

Oxford University Computing Service

TEXT ARCHIVE

Oxford Text Archive

I 1976 etablerte Oxford University Computing Service (OUCS) et tekstarkiv som i dag er et av verdens største. Nylig utga OUCS en revidert katalog over tekstene som er lagret både i Oxford og ved Literary and Linguistic Computing Centre i Cambridge. Foruten engelsk er 34 språk representert i disse samlingene, som også er tilgjengelige for utenlandske forskere. Katalogen inneholder i tillegg en liste over de viktigste tekstarkivene i de andre europeiske land.

Oxford Text Archive har dessuten en oversikt over maskinleselige tekster som er lagret eller tilrettelegges i andre land, og tar del i forberedelsene til en internasjonal database over denne typen materiale.

Katalogen og ytterligere informasjon fås ved henvendelse til: *Oxford Text Archive, Oxford University Computing Service, 13 Banbury Rd, Oxford OX2 6NN, England.*

Senterrapport nr. 34

NAVFs EDB-senter for humanistisk forskning har nylig utgitt nr. 34 i den eksterne rapportserien: *Tutorial on Machine Translation. Rapport fra en konferanse i Lugano 2.-6. april 1984,* av *Jostein H. Hauge.* Vi sakser fra forordet: «Hensikten med rapporten er å øke kunnskapen i Norge om det arbeid som i dag foregår innenfor datamaskinstøttet oversettelse. Målet har vært å få frem hovedtrekk i den mangesidige virksomhet som pågår, uten bruk av vanskelig tilgjengelig terminologi.

Fremstillingen bygger både på skriftlige presentasjoner som ble lagt frem av foredragsholderne og supplerende opplysninger gitt under forelesningene.»

I rapporten drøftes utviklingen i maskinstøttet oversettelse fra starten til i dag og hovedtrekk i nyere systemer. Det gis i tillegg eksempler på slike systemer og en oversikt over foredrag og deltakere på konferansen.

Rapporten, som koster kr 60, kan bestilles fra Senteret. En kortfattet versjon av rapporten ble gitt i Humanistiske Data 2-84.

Konferanser

CALICO'85

CALICO - Second Annual Symposium

Computer Assisted Language Learning and Instruction Consortium (CALICO) skal holde sitt andre årlige symposium i Baltimore, Maryland, USA 29.1.-2.2. 1985. Formålet med symposiet er å samle bl.a. lærere, forskere, programutviklere, offentlige myndigheter og representanter fra edb-firmaer til en diskusjon om hvordan teknologi kan tilpasses en mer effektiv undervisning i og læring av språk. Foruten foredrag vil det bli holdt «workshops» og en utstilling av både program- og maskinvare. Flere opplysninger fra:
CALICO '85, 3078 JKHB, Brigham Young University, Provo, Utah 84602, USA.

Second Infoterm Symposium

International Information Centre for Terminology (Infoterm) skal avholde sitt andre symposium i Wien 14.-17. april 1985. Hovedtema for

konferansen blir «Networking in Terminology» – International Co-operation in Terminology Work. Representanter for internasjonale tekniske og vitenskapelige organisasjoner som arbeider med terminologi, spesialiserte språkformidlere og vitenskapelige redaktører vil både holde foredrag som peker på terminologiske problemer og løsninger og rapportere om samarbeidstiltak som søker å imøtekomme terminologiske behov. En egen sesjon vil dreie seg om edb-hjelpemidler for terminologi-nettverk og datastøttet terminologisk arbeid. En demonstrasjon av terminologiske databanker blir også arrangert.

I tilknytning til symposiet vil det bli holdt et møte i TermNet (17.-18. april) og et «workshop» om dokumentasjon av terminologi (18.-19. april).

Flere opplysninger om disse arrangementene og en liste over dokumenter utgitt av både Infoterm og TermNet kan fås fra: *Infoterm, Postfach 130, A-1021 Wien, Austria.*

Kurs i menneske-maskin kommunikasjon

INRIA (Instiut Nationale de Recherche en Informatique et en Automatique) organiserer kurset «Fundamentals in Computer Understanding: Speech, Vision and Natural Language». Kurset vil finne sted i Les Premontrés, Frankrike i tida 28. mai-7. juni 1985. Arrangørene opplyser at følgende temaer blir tatt opp: Automatic speech recognition, Principles of computer vision and natural language understanding, Multi-media man-machine dialog, Knowledge-based and expert systems – application to speech, vision and natural language. Flere opplysninger kan fås fra:
INRIA, Service des Relations Extérieures, Bureau Cours-Écoles, Domaine de Voluceau – Rocquencourt, B.P. 105 – 78153 Le Chesnay Cedex – France.

Cognitiva 85

Målet med Cognitiva 85, som skal finne sted 4.-6. juni 1985 i Paris, er å samle medarbeidere innen de kognitive vitenskapene, nevrovitenskapene og edb-teknologi til en konferanse om kunstig intelligens. Konferansen vil bestå av tre deler: et vitenskapelig symposium, et forum for forskere og en utstilling av maskinvare. Følgende temaer blir dekket: kognitivitet, kunnskapsrepresentasjon, arkitekturer, persepsjon, språk, læring, hukommelse, problemløsning, menneske-maskin kommunikasjon, ekspertsystemer og robotikk. Påmelding til og flere opplysninger fra:
Secrétariat Cognitiva 85, Marie-France Chicanne, CESTA, 1 rue Descartes, 75005 Paris, France.

International Workshop on the Creation, Connection and Usage of Large-scale Interdisciplinary Source Banks in the Historical Disciplines

Göttingen, July 15th through 18th 1985

NAVFs EDB-senter har ei 10-siders orientering om denne verkstaden. Kontaktmann: Per Vestbøstad, tlf. (05) 212959

NORD IoD 6

Den 6. nordiske informasjons- og dokumentasjonskonferansen skal avholdes i Helsingfors 19.-23.8. 1985. Konferansens målsetting er å skape et forum for utveksling av erfaringer og planer mellom informatikere og dokumentalister i de nordiske land.

Konferansens motto er «Information Resources Management (IRM)» - Informasjon som ressurs: kilder - system - tjenester - produkter. Disse fire elementene skal utgjøre den røde tråden i fire plenumssesjoner. Innledere blir invitert fra USA, EF-landene og samtlige nordiske land. Demonstrasjoner og en utstilling av moderne utstyr vil bli arrangert i tilknytning til konferansen. Ytterligere opplysninger fås fra:
NORD IoD 6, Sekretariatet, c/o Tekniska högskolans bibliotek, Otnäsvägen 9, SF 02150 Esbo 15, Finland.

De nordiska datalingvistikdagarne 1985

Institutionen för allmän språkvetenskap ved Helsingfors universitet inviterer til de femte nordiske datalingvistikkdagene i tida 11.-12. desember 1985. Eksakt tema for dagene er ikke fastsatt ennå. De som ønsker ordnet innkvartering, må melde seg før 15. januar, ellers er påmeldingsfristen 1. mars. Flere opplysninger fra:
Fred Karlsson, Institutionen för allmän språkvetenskap, Helsingfors universitet, Regeringsgatan 11, SF-00100 Helsingfors, Finland.

Symposium om datastøttet terminologi- og leksikografiarbeid

I forbindelse med de nordiske datalingvistikkdagene i Helsingfors 1985 arrangeres et symposium i datastøttet terminologi- og leksikografiarbeid 13. og 14. desember 1985. Et avgrenset tema for symposiet er ennå ikke fastsatt. Referat av foredrag må sendes innen 1. september. Påmeldingsfrister: For ordnet innkvartering - 15. januar, ellers - 1. mai. Flere opplysninger fra:
Centralen för teknisk terminologi, Elisabetsgatan 16 B 13, 00170 Helsingfors, Finland.

SUMMARY

Hva er datalingvistikk?

What is computational linguistics?

Professor Helge J. Jakhelln Dyvik at the Dept. of Phonetics and Linguistics, University of Bergen, characterizes computational linguistics as the development of software for basic research purposes and practical applications, which incorporates linguistic insight or simulates linguistic competence. Features of the main types of systems developed for practical purposes are discussed in detail. Among these are question-answer systems, expert systems, and systems for automatic translation, text retrieval and language recognition. All systems of this kind require an element of simulated language understanding.

Analysis of natural language for computer systems must be conducted on many levels: morphological, lexical, syntactic, semantic and pragmatic. Dyvik discusses the linguistic theories and models this work is based on.

Computational linguistics has also influenced traditional linguistics, in that the former makes it possible to study not only language structure (competence), but also language processes (performance).

Dyvik draws a line between computational linguistics and computer-assisted linguistics. In his opinion the former should be given higher priority in Norway. One of the reasons for this is that it is necessary to develop and adapt systems for the analysis of Norwegian language.

Datalingvistikk i Norge

Computational linguistics in Norway

Research Fellow Helge Lødrup at the University of Bergen gives an overview of computational linguistics in Norway – a fairly new but expanding field of research. Even engineers and natural scientists have been engaged in research of this kind. Work has been carried out in Trondheim and Oslo on natural language for practical purposes, topics on the borderline between logics, linguistics and informatics, and semantic interpretation within Lexical Functional Grammar (LFG).

Recently interest in computational linguistics has become more wide-spread among linguists – a natural consequence of increased work on formal syntax and semantics. Pioneering work was carried out by Svein Lie at the University of Oslo, who in co-operation with Knut Hofland at the Centre designed a parser for Norwegian.

Studies at different levels and new positions in computational linguistics have been/will be established at the universities of Tromsø, Trondheim and Bergen. In Bergen joint research is being carried out by the Department of Phonetics and Linguistics, the Department of Scandinavian Language and Literature and

the Centre on computational linguistics within the LFG model. So far this work has dealt with automatic syntactic and semantic analysis of Norwegian.

Finländsk datalingvistik

Computational linguistics in Finland

The point of departure of this article by Professor Fred Karlsson is the fact that the current theories and models of computational linguistics are not suitable for the analysis of languages such as Finnish and Hungarian. At the Department of General Linguistics at the University of Helsinki a project called «Automatic analysis of Finnish» was started in 1981. The primary goal of this project is to establish which theoretical demands a universally valid, integrated morpho-syntactical parser should meet in order to analyze Finnish as well as e.g. English. So far, the project's most important result is Kimmo Koskeniemmi's language independent, two-level model for morphological analysis and synthesis, which Karlsson describes.

The syntactical part of this project has just been started. With the assistance of Jouko Lindstedt, Karlsson has constructed a morphologically orientated, semi-automatic tagging program, FINTAG, which has been used in the work of designing a parser for the basic sentence structure of Finnish. Karlsson both describes this tagging program and outlines the principles of the parser. Between 1985 and 1990 the principles of a *general* parser will be investigated in detail.

Several aspects of this work have turned out to have practical applications, e.g. in information retrieval systems.

Datalingvistlinjen i Göteborg

Studies in computational linguistics in Gothenburg

University Lecturer Lars Ahrenberg reports on the new four-year course in computational linguistics established at the University of Gothenburg. This course is a joint venture between the departments of linguistics, computational linguistics, informatics and philosophy. Students spend equal time studying the first three fields. The last term consists of project work leading to a final exam. Ahrenberg gives an overview of the subjects taught in each term.

Sprogbeskrivelse til flersproglig maskinoversættelse

Language description for multilingual automatic translation

Lecturer Hanne Ruus at the Institute for Scandinavian Philology, University of Copenhagen starts this article by describing the phases of linguistic analysis necessary for the construction of an automatic translation system. The special requirements of a multilingual system are outlined, exemplified by the EUROTRA project (introduced in Humanistiske Data 2-82), the goal of which is an automatic translation system between each of the seven EEC languages. For this purpose an analysis module, a synthesis module and twelve transfer modules will be developed for each language.

So far the bulk of the work has dealt with determining what linguistic

knowledge should be accessible in the transfer stage and how it should be specified. At present the project is in a preparatory stage which will result in a precise description of the software to be developed for the system. Ruus describes the planned contents of this system. In the next phase of the project a preliminary prototype will be developed that will be able to translate between each of the languages within a vocabulary of 2500 words.

Edb og talemålsforsking

ADP and research in spoken language

Lecturer Helge Sandøy at the Department of Scandinavian Languages and Literature, University of Bergen discusses a project started in 1977 on the development and variations of the spoken language of young people in Bergen. Conversations with 104 informants (220.000 words) have been recorded on tape and data processed in order to facilitate various types of analyses.

Sandøy outlines the principles for the transcription of the conversations and the preparation of the material for the production of a concordance. The next phase of the project involved morphological tagging, a process which Sandøy describes in detail. The morphologically tagged concordance formed the basis for the analyses, which have been undertaken with the aid of specially written programs for selection and sorting. In its data processed form this material can be used for other purposes than its original goal – in teaching, for grammatical analyses, and for work on speech synthesis.

Sandøy gives an evaluation of the use of ADP in this project. In his view, although data processing of the material was costly, it involved many practical advantages, such as a high degree of precision and a wide range of possibilities for use. However, the data preparation was so time-consuming that research work proper became rather frustrating due to the long time gap between hypothesis formation and results.

Personregisterloven og behovet for datavern

The Personal Data Registers Act and the need for data security

The Norwegian Personal Data Registers Act entered into force on January 1st, 1980. The act imposes applications to the Data Inspectorate for concession for establishing machine-stored personal data registers. Executive Officer Thore Gaard Olaussen at the Norwegian Social Science Data Services (NSD) summarizes the main features of this law. It pertains to both machine-stored and manually stored registers, but does not cover the problems posed by the rapid technological development.

Olaussen also outlines the negative consequences of this act for the use of data in research. In order to protect the researchers' right to gain access to information regulated under this law, the Norwegian Research Council for Science and the Humanities has established a secretariat for data protection affairs, located at NSD in Bergen.

As a result of negotiations with the Data Inspectorate the Secretariat now reports to and advises the Inspectorate on research projects that need concession. In addition, sensitive information can be transferred to NSD for storage, with the permission of a special committee for data archiving.

Olaussen states that the term «data security» also pertains to unauthorized

use and destruction of data. By defining the term in this way, a constructive discussion may take place between parties with differing interests to attend to.

Spørjeundersøking om bruken av statistiske metoder i språk- og litteraturforskinga

Survey on the use of statistical methods in linguistic and literary research

Computing Officer Ole Lauvskar at the Centre was responsible for this survey carried out last Spring. The goals of the survey were to chart both the use of and the need for statistical methods.

Out of a total of 470 researchers in linguistics and literature, 203 or 43% filled out the form sent to them - an equal number from both groups. However, 40% of the linguists use/had used statistical methods, compared to only 16% of the literary researchers.

Half of these researchers had used both advanced methods and data processing in their work. Two thirds had acquired their statistical knowledge on their own initiative. The greatest problems encountered when using statistics were in connection with methodical questions and the preparation of data.

Among those who had not used statistical methods, most of the linguists but few of the literay researchers felt that such methods are relevant to their respective fields. Of the whole sample, many have a wish for greater possibilities of training in and of using statistical methods.

Nordiske arkivdager i Ebeltoft 2.-5. august 1984

Nordic Archives Seminar in Ebeltoft

The 14th Nordic Archives Seminar, held in Ebeltoft, Denmark in August was divided into eight sessions. Senior Archivist Anne Hals at the National Archives of Norway reports on the session called «ADP and the Archives.» In this session the following topics were dealt with: reception of data processed archive materials, ADP-based journal systems, the use of ADP in archives institutions, and training in ADP for employees at these institutions. In connection with each of these topics papers were given on the present situation in the Nordic countries, and problems and future tasks were discussed.

On methods for using poulation registers in historical research

Computing Officer Eirik Lien from the University of Trondheim reports on this international conference at the University of Umeå, Sweden. It was held in connection with a series of conferences on historical data bases in the Nordic countries.

One part of the conference consisted of project reports. These dealt with, among other topics, Japanese population registers, the reliability of public registers, record linkage, and simulation in demographic research.

The participants were also given a guided tour of the Demographic Data Base in Umeå. Here emphasis is put on mediation - the main topic of next year's conference in Stockholm.

ECAI 84 - 6th European conference on artificial intelligence

Senior Computing Officer Øystein Reigem from the Centre attended this conference, which was held in Pisa, Italy, September 5-7, with a two-day tutorial on the 3rd and 4th. In addition an industrial liaison session and a session on ESPRIT and artificial intelligence were held on the 4th. (ESPRIT is the European counterpart to the Japanese Fifth Generation Project.) An exhibition lasting the whole week presented artificial intelligence software and hardware.

The tutorials addressed four topics - languages for artificial intelligence, expert systems, natural language processing and robotics and vision. The conference also covered cognitive modelling, planning and search, system support, knowledge representation, automatic programming, theorem proving, logic programming, industrial applications, learning and philosophical implications. 160 lectures were held in 5-6 parallel sessions. 700-800 people attended the whole or parts of the conference.

In his report Reigem concentrates on the tutorial on expert systems, especially on the speech by Bob Wielinga, Senior Lecturer at the University of Amsterdam, Departments of Psychology and Social Science Informatics.

Toward a Computer Ethnology

Director Jostein H. Hauge at the Centre reports on an international symposium in Osaka, Japan, September 16-23. The theme of this symposium - arranged by the National Museum of Ethnology - was computational methods in ethnological research.

The Japanese hosts spoke on topics related to the advanced developmental work being carried out at the museum. The four participants from the West (England, Germany, Norway, USA) gave talks on the development of computer-assisted research in the humanities and new methods for data base design and semantic analysis of natural language.

Visits were arranged to factories that use robots to produce electronic equipment and the participants met with humanistic researchers in the fields of industrial design and automatic translation. Work was also presented on the standardisation of the Japanese sign system and specially designed hardware for kanji, and on computational methods in archaeological research.

The National Museum of Ethnology is one of the most technically advanced in the world. One of its features is a videotek, an audiovisual mediation system, steered by robots, for the storage, treatment and presentation of some 800 video- and audiocassettes. The museum also has an impressive range of computational facilities. Between 1977 and 1983 computational work was mainly aimed at organizing catalogues and reference data in data bases. Since then a program has been initiated for the solution of problems connected with the storage, treatment and presentation of non-written primary sources. It is hoped that this work will result in a fully integrated system in which differentiated information on primary material is integrated with data bases containing depictions of primary sources in digitalized form.

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

This conference took place in Pisa September 25-27 1984. There were approximately 300 participants from the whole world, but only three from Norway: Svein Engelstad, Britt Kroepelien and Espen Ore, who report on the conference.

Most of the papers read at the conference discussed the need for international standards in the processing of art history data. In addition to the lecture sessions there were demonstrations of various projects and pieces of equipment.

Among the papers read, many were of general interest to participants who work with information storage and retrieval. Some projects were, however, of particular interest to art historians. One of the more important of these projects has resulted in an iconographic system for classifying western art. The system, called ICONCLASS, has been developed at the University of Leiden in the Netherlands over a period of about 35 years.

ICONCLASS is used by many projects and institutions. At the conference one such project was presented: The Marburger Index. This index consists of a catalogue with 560.000 illustrations and 30.000 texts. The Marburger Index is available on microfilm and as a data bank.

Besøk ved Istituto di Linguistica Computazionale - CNR, Pisa

Visit to Istituto di Linguistica Computazionale

The Istituto di Linguistica Computazionale in Pisa applies computer technology to linguistics in general. Computing Officer Espen Ore at the Centre visited the Institute in September.

A. Bozzi from the classics department showed how work is progressing on a system for automatic lemmatisation of Latin. Since Latin is a highly inflected language, the system makes use of lists of pre-, in- and suffixes. The lists also contain diacronic variants so that Latin texts from different periods can be handled by the system.

The institute has a group that works on parsing of Italian and knowledge-based retrieval systems. Among the members of this group are G. Ferrari and I. Prodanof. So far most of the work has been concentrated on parsing. The group has produced three ATN-based parsers. One of the parsers is implemented in a simplified version that is used by students of linguistics as part of their training.

Datamaskinen - språkforskernes store utfordring i vår tid

The computer - today's biggest challenge to language researchers

In October the Norwegian Association of Applied Linguistics held their annual symposium in Bergen. The main themes this year were computational linguistics and prosody in foreign language teaching. Director Jostein H. Hauge at the Centre gives summaries of some of the speeches held at the symposium, including Helge Dyvik's, which is printed in full from page 4.

Kolbjørn Heggstad from the firm Logos spoke on commercial computing firms' expectations of computational linguists. The theme for the speech made by Lars S. Evensen, University of Trondheim, was the role of computational

linguistics within the field of applied linguistics. Helge Lødrup, University of Bergen, gave an overview of research in computational linguistics.

More application-orientated papers were also given. Gulbrand Alhaug, University of Tromsø, outlined a project plan for computer-assisted translation between the two variants of the Norwegian language. Anne Golden and Anne Hvenekilde, University of Oslo, gave an account of how inexperienced users can take advantage of simple, general programs for computing and language analysis. Stig Johansson, University of Oslo, explained the methods developed in connection with the grammatical tagging of the British English LOB corpus, whereas Knut Hofland at the Centre showed how a compuational syntactic analysis of Norwegian can be developed.

Fra bokhylla

From the bookshelf

Roald Skarsten of the Faculty of Arts' ADP section, University of Bergen, reviews *CAI SOURCEBOOK Background and Procedures for Computer Assisted Instruction in Education and Industrial Training, by Robert L. Burke.* In this book CAI is placed in a historical and pedagogical/theoretical context, and the necessary terminology for evaluating different techniques is presented. Burke also includes an instructive account of the three production techniques available to course authors: programming in a general language, using an author language, and using a special «CAI authoring system.» Technical possibilities (sound, graphics, etc.) are also discussed. Burke stresses the fact that CAI enables teachers to adapt material to each pupil's needs.

Meldinger

News

The British Library has established an Office for Humanities Communication at the University of Leicester. The Office will «... alert research workers in the humanities to developments in communication of relevance to their activities, with particular emphasis on new technologies and their applications in the humanities.» The Office conducts surveys, organizes seminars etc., and publishes «Humanities Communication Newsletter». Address: *Office for Humanities Communication, University of Leicester, Leicester LE1 7RH, England*

In 1976 Oxford University Computing Service established Oxford Text Archive, one of the world's largest. Recently OUCS published a catalog of texts stored both at Oxford and at Literary and Linguistic Computing Centre in Cambridge. This catalog and further information can be obtained from: *Oxford Text Archive, Oxford University Computing Service, 13 Banbury Rd, Oxford OX2 6NN, England.*

Forthcoming conferences:

CALICO (Computer Assisted Language Learning and Instruction Consortium) - Second Annual Symposium - Baltimore, Maryland, USA, 29 January-2 February.

Second Infoterm Symposium - «Networking in Terminology» - Vienna, 14-17 April.

«Fundamentals in Computer Understanding: Speech, Vision and Natural Language» - Les Premontrés, France, 28 May-7 June. Course arranged by INRIA.

Cognitiva 85 - conference on artificial intelligence - Paris, 4-6 June.

Forts. fra 2. omslagsside.

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

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

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

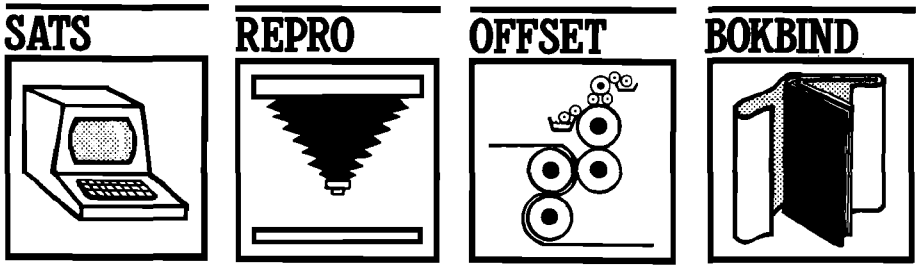
nr. 31 STAR III Archaeology for statisticians. ISBN 82-7283-035-3 Pris kr. 60.

nr. 32 STAR IV STAR algorithms. ISBN 82-7283-036-1 Pris kr. 30.

RAPPORT nr. 33. *Årsmelding 1983.* NAVFs EDB-senter for humanistisk forskning. ISBN 82-7283-038-8 Gratis.

RAPPORT nr. 34. *Jostein H. Hauge: Tutorial on Machine Translation.* Rapport fra en konferanse i Lugano 2.-6. april 1984. ISBN 82-7283-039-6. Pris kr. 60.

C



Vår garanti –
trykksaker av høy faglig kvalitet

A.S JOHN GRIEG

Grafisk produksjon · Vaskerelven 8 · Postsboks 248 · 5001 Bergen
tlf.: (05) 23 39 00 · telefax: (05) 32 01 06

Returadresse:

NAVFs EDB-senter for humanistisk forskning
Boks 53
5014 Bergen Universitetet