NAVF'S EDB-SENTER FOR HUMANISTISK FORSKNING
VILLAVEI 10, POSTBOKS 53
5014 BERGEN-UNIVERSITETET

Oktober 1979

NORSK TEKSTARKIV

Jostein H. Hauge

1. FORHISTORIE

Datamaskinell språkbehandling er kanskje det feltet som har stått mest sentralt siden datamaskinene gjorde sitt inntog i de humanistiske fag på 60-tallet i Norge.

Ved Nordisk institutt, Universitetet i Bergen ble det tidlig bygget opp en avdeling, Prosjekt for datamaskinell språkbehandling, som særlig skulle arbeide med oppgaver av betydning for norsk språkvitenskap og norsk språklig utviklingsarbeid. Senere har NAVF's EDB-senter i Bergen og EDB-tjenestene for de filologiske fag ved universitetene i Oslo og Trondheim vært aktive i arbeidet med språklig databehandling.

NAVF's EDB-senter har arbeidet med flere større språkprosjekter i de siste årene. I tillegg til arbeid med et stort Ibsen-prosjekt er det foretatt revisjon og nyutgivelse på mikrokort av Brown Corpus, og vi avslutter i disse dager det tilsvarende Lancaster/Oslo/Bergen Corpus i samarbeid med dosent Stig Johansson, Britisk institutt, Universitetet i Oslo. Senteret er videre utøvende EDB-organ for organisasjonen International Computer Archive of Modern English. Dette har vi sett på som viktige og lærerike oppgaver, som også har fått stor internasjonal oppmerksomhet, men utførelsen har også kostet oss ganske mye i tid og penger.

I vårt styre har derfor spørsmålet vært reist om ikke tiden nå var moden til å avslutte servicearbeidet for engelsklingvistene og vende øynene mot vårt eget språk, som få andre enn nordmenn kan tenkes å interessere seg for. Når

Norsk tekstarkiv
Jostein Hauge
Proceedings of NODALIDA 1979, pages 7-13

styret påpekte dette, var det også fordi en generelt kan merke en stadig økende interesse for og et stigende behov for norske språkdata, særlig fra moderne tid.

2. KONFERANSEN OM ET NORSK DATAMASKINELT TEKSTKORPUS

For å få drøftet spørsmål i tilknytning til norske språkdata, inviterte NAVF's EDB-senter i oktober 1978 til en konferanse i Bergen om et norsk datamaskinelt tekstkorpus. På konferansen deltok 32 representater fra universitets- og høyskolesektoren og fra andre miljøer hvor en arbeider med norske språk. I tillegg til norske deltakere var det også invitert gjester fra Sverige og Danmark (Sture Allén, Rolf Gavare og Bente Maegaard). Hensikten med konferansen var i første rekke å få i stand en drøfting av behov for norske språkdata i datamaskinell form og de prinsipper en bør legge til grunn for arbeidet her. Det viktigste innslaget på konferansen kom til å bli plenumsdrøftingene, men forut for disse ble det gitt en serie innlegg om det arbeid med større tekstsamlinger som pågår i Danmark, Sverige og i vårt eget land. Det foreligger en omfattende konferanserapport som gir et detaljert innsyn i det som foregikk på denne konferansen (Et norsk datamaskinelt tekstkorpus. Rapport nr. 2. Februar 1979). I vår sammenheng skal bare følgende punkter nevnes:

1. Det var stor interesse for å få intensivert arbeidet med å skaffe fram, lagre og utnytte tekstmateriale fra moderne norsk i undervisning og forskning.
2. Det er mange ulike brukerinteresser knyttet til dette arbeidet.
3. Det var en stor skepsis mot å lage et norsk standard tekstkorpus.
4. Heller gikk drøftingene i retning av å planlegge et tiltak som hadde karakter av en språkbank. På konferansen, hvor mange knapt nok anerkjente verdien av de tradisjonelle korpus, mente de fleste at det ville være bedre i dag å satse på et tiltak som hadde som hovedmål å samle inn,

tilrettelegge og gjøre bruksklare tekster slik at brukerne selv kan velge ut de tekstene eller tekstdelene de ønsker for ulike spesialoppgaver.

5. Det ble vedtatt å nedsette en gruppe som kunne arbeide videre med de tankene som kom fram på konferansen.

3. PLANLEGGINGSGRUPPEN

Kort tid etter konferansen i Bergen kom det i sving en planleggingsgruppe som skulle arbeide videre med det grunnlagsmateriale som var fremskaffet gjennom konferansen og utforme et tiltak langs de retningslinjer som var trukket opp i plenumsdrøftingene. Følgende personer og institusjoner har vært med i planleggingen:

Jostein H. Hauge, NAVF's EDB-senter
Kolbjørn Heggstad, Nordisk institutt, PDS, UiB
Aagot Landfald, Norsk språkråd
Eirik Lien, EDB-tjenesten for humanistiske fag, Tr.heim
Egil Pettersen, Nordisk inst., UiB (formann)
Jarle Rønhovd, Nordisk inst., UiTrheim
Dagfinn Worren, Norsk leksikografisk institutt

Som en vil se av dette, la en vekt på allerede i utgangspunktet å markere dette tiltaket som et nasjonalt prosjekt og også å sikre at ulike brukerinteresser kom fram i planleggingsarbeidet. I løpet av noen møter høsten 1978 og våren 1979 greidde planleggingsgruppen å konkretisere et tiltak som ble kalt Norsk tekstarkiv.
De planene som planleggingsgruppen kom fram til ble sendt alle deltakerne fra Bergenskonferansen til uttalelse, og det viste seg at det ikke innkom noen merknader til de planene som var lagt fram. Dette velger planleggingsgruppen å tolke som tilslutning til planene, heller enn som manglende interesse for dem. Når vi tror at det er slik, skyldes det det heftige engasjement og den store interesse som ble vist denne saken på konferansen i Bergen.

4. NORSK TEKSTARKIV

Norsk tekstarkiv har som mål å koordinere og øke innsatsen i arbeidet med å samle inn og tilrettelegge tekstmateriale fra moderne norsk til bruk i forskningsog utviklingsarbeid. Tiltaket vil fra starten av bli nasjonalt orientert. En forutsetning for at Norsk tekstarkiv kan resultere i en vitenressurs om norsk språk, er at arbeidet med tekstinnsamling legges opp etter en nasjonal koordinert plan og at materialet tilrettelegges på en standard måte. De datamengder som legges opp, må kunne utnyttes datamaskinelt i alle interesserte miljøer med et minimum av ekstra tilretteleggingsarbeid. Hovedtyngden av de data som samles inn, vil referere seg til moderne norsk skjønnlitteratur og bruksprosa. Det sier seg selv at det her bare kan bli tale om å samle inn en meget liten del av alt som utgis. I størst mulig utstrekning er det derfor ønskelig å oppnå enighet på nasjonal basis om de prioriterte oppgaver. Det er tatt spesielt hensyn til dette ved oppbygging av organisasjonsstrukturen for Norsk tekstarkiv.

Selv om arbeidet med norsk datamaskinelt tekstmateriale bør ses i et langt perspektiv, er det naturlig i første omgang å definere en prosjektramme på 5 år for Norsk tekstarkiv. I løpet av denne perioden bør ulike sider av prosjektet kunne prøves ut. Vi tenker i første rekke på nødvendig metodeutvikling på EDB-siden og organisasjonsformer for samvirke mellom mange EDB-organer. Like viktig vil det bli å få kunnskap om hvilke typer tekstmateriale som bør prioriteres og finne fram til egnede presentasjons- og utnyttelsesformer av tekstdata. Innenfor dette tidsrom vil det bli mulig å gjennomføre ulike typer prosjekter knyttet til spesielle tekstsamlinger. Sist i prøveperioden blir hovedoppgaven å finne fram til permanente organisasjons- og finansieringsformer for et slikt tiltak.

5. ORGANISASJON

Grunnlaget for Norsk tekstarkiv (NT) vil være et formalisert samarbeid mellom PDS, Nordisk institutt, UiB og NAVF's EDB-senter i Bergen, som igjen har EDB-tjenestene ved

HF-fakultetene ved universitetene som sine faste samarbeidspartnere. Siden et tiltak som NT vil ha et betydelig innslag av dataadministrasjon og administrative rutiner generelt, la planleggingsgruppen vekt på å finne fram til et organisasjonsmønster som gir klare ansvarsforhold og også å foreta en markert funksjonsfordeling mellom de ulike medvirkende parter i prosjektet.

6. FAGLIG INNHOLD

Som tidligere nevnt er hovedmålet med NT å kunne intensivere arbeidet med å samle inn, tilrettelegge og presentere data for språkvitenskapelig forsknings- og utviklingsarbeid. Når NT kommer i gang, vil den første oppgaven være å fastlegge et standardformat for tekstmateriale som skal inngå i arkivet. Det vil dessuten være behov for å utvikle program for konvertering og justering til dette standardformatet. De fleste av oss vet at det foreligger en rekke typer utstyr i den grafiske industri, som NT vil forsøke å samarbeide med, og det trengs en god del systemerings- og programmeringsarbeid for å kunne utvikle konverteringsprogram for å ta de datamaskinlesbare tekstene som man kan få fra tykkerier, forlag, aviser etc. på hullbånd eller magnetbånd. Det bør også utvikles et datamaskinelt lagringssytem for tekstmateriale som er effektivt og som kan ta hånd om den stadig økende tekstmengde. Målet er å kunne ta vare på mest mulig informasjon som ligger i tekstene, enten i originalkilden eller i den versjon som den grafiske industri presenterer, men samtidig å ha tekstene slik organisert at hver enkelt bruker kan undertrykke informasjon som er lite interessant i den enkelte forskningsoppgave, og også legge til informasjon som er nødvendig for den aktuelle oppgave. En fordel med et standard dataformat vil være at brukerne av NT alltid vet hvordan de vanligste typer av tekster ser ut, og det vil også ha som fordel at de som arbeider med utviklingsoppgaver på EDB-siden, vil kunne legge til rette sin programutrustning for å kunne behandle data fra NT.

Ved fastsetting av vårt eget standardformat vil vi selvsagt skjele til det som er utført i andre land, ikke minst her i Danmark og i Sverige. Det er etablert en styringsgruppe som

skal lede arbeidet til en medarbeider som skal utgreie spørsmålet om standardformat, og vi håper til slutt å kunne stå med et framlegg som vi kan foreslå som norsk norm innenfor humanistisk databehandling i alle fall.

Det er til i dag ikke foretatt en omfattende vurdering av hvilken ende en skal starte i, dvs. hva slags type tekster en først skal forsøke å få inn i NT. Der er en del planer lansert, f.eks. at NT bl.a. burde ta sikte på å få med all ny norsk skjønnlitteratur eller et utvalg av denne hvert år. Denne oppgaven og mange andre ønsker som kom fram på konferansen i Bergen, må gjennomgås nøye før vi setter i gang. Det bør her nevnes at det i Norge allerede i dag foreligger en ikke ubetydelig del tekstmateriale som ligger tilrettelagt for datamaskinell behandling, i første rekke ved PDS, Universitetet i Bergen, noe ved NAVF's EDB-senter og en god del materiale spredd rundt om ved universitetene, materiale som har vært brukt i konkret prosjektarbeid. Et viktig mål, etter mitt syn, må i første omgang være å få samlet det mest verdifulle av dette materialet og lagt det til rette etter de prinsipper som er nevnt tidligere.

7. STATUS FOR NORSK TEKSTARKIV I DAG

Det er forutsetningen at NT skal komme i gang fra 1980 av. Noe støtte til tiltaket kan trolig gis fra grunnbevilgningen til PDS og NAVF's EDB-senter, men for å komme noe vei er det nødvendig å ha eget personale knyttet til NT. Planleggingsgruppen søkte derfor NAVF om støtte til tiltaket for 1980, bl.a. støtte til å opprette en hel stilling for en vitenskapelig assistent som kan spesialisere seg for oppgavene med NT. Denne søknaden ble innvilget ved Rådets budsjettbehandling i oktober i år. Dermed vil forholdene ligge godt til rette for etablering av Norsk tekstarkiv neste år.

8. ARBEIDET FREMOVER

Det vil derfor i høst bli aktuelt å innkalle det faglige råd for NT for å diskutere de bevilgninger som er gitt til NT, den organisasjonsplanen som er utviklet, og på den bakgrunn

forme planer om den konkrete oppstarting av tiltaket i vinter. Like viktig vil det være å komme fram til en faglig prioritering mellom de ulike arbeidsoppgaver som er stilt opp, d.v.s. velge ut hvilket tekstmateriale en skal ta fatt på i første omgang, og også finne fram til praktiske samarbeidsformer mellom samarbeidspartnere i Bergen, EDB-konsulentene ved universitetene og andre brukermiljøer som har spesielle interesser knyttet til NT, f.eks. Norsk språkråd og Norsk leksikografisk institutt. Det vil i løpet av høsten også bli tale om å forhandle med Norsk språkråd om betalte oppdrag i forbindelse med undersøkelser som Norsk språkråd ønsker utført vedrørende utviklingen av norsk språk slik den fremkommer i litterær prosa fra 1937 til i dag.