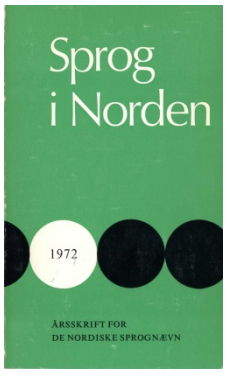
Sprog i Norden

Titel:	Datamaskinell språkbehandling – og nordisk samarbeid
Forfatter:	Kolbjørn Heggstad
Kilde:	Sprog i Norden, 1972, s. 103-107
URL:	http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive



Søgbarhed
Artiklerne i de ældre numre af Sprog i Norden (1970-2004) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Datamaskinell språkbehandling — og nordisk samarbeid

Av Kolbjørn Heggstad

Det er ikkje lenger berre numerisk informasjon som datamaskinane rundt om ved dei ulike forskningssentra i Norden arbeider med, språklege data er i høg grad med og krev "CPU-time", som det enno heiter på norsk. "Output" strøymer ut av dei store linjeskrivarane med ein fart på over 1 000 linjer i minuttet, og kanskje ser ein der ordindeksar og konkordansar over gamle greske tekster utskrivne med fullt gresk teiknsett. Datamaskinell språkvitskap har etter kvart blitt eit veletablert arbeidsområde med ein regelbunden internasjonal kongressserie.

Bruk av datamaskin i språkvitskapen vart svært tidleg tatt opp til drøfting i språknemndene. Alt i 1963 på det nordiske språkmøtet i Stockholm gav professor Carl Ivar Ståhle ei grundig utgreiing om "möjligheterna att åstadkomma en morfemordbok med hjälp av datamaskiner". Professor Paul Diderichsen gjekk sterkt inn for vidare utgreiing om arbeidet, han ville ha eit større nordisk samarbeid på dette området. Han foreslo at språknemndene skulle prøve å organisere ei nordisk arbeidsgruppe for å undersøkje korleis nordisk språkforskning med hjelp av datamaskinar skulle leggjast opp.

På det nordiske språkmøtet i København 1964 var databehandling eitt av hovudpunkta og ein finn referat frå drøftingar om Svenska Akademiens ordlista (SAOL) og den danske Retskrivningsordbog som "magnetbandordbøker". Det var alt gjort kostnadsoverslag og ein arbeidsplan for overføringa av SAOL. Etter drøftingane på dette møtet gjekk Nämnden för svensk språkvård i gang med overføring av SAOL til magnetband. Første mål var å utarbeide ei baklengsordbok

(finalalfabetisk) på grunnlag av ordtilfanget i ordlista. Vidare ville ein ha ei førehandsredigering av ordlista der ein sette inn markeringar for alle morfemgrenser. Denne informasjonen skulle også overførast til magnetband med tanke på utarbeiding av ei morfemordbok. Overføringa er no ferdig og baklengsordliste, morfemordliste og ei rad andre spesialsorteringar er tilgjengelege. (Erik Kristensen: SAOL-LISTER 15/1—71. Interim report no. 33, Forskningsgruppen för kvantitativ lingvistik (KVAL). Sth. 1971.)

Situasjonen i Noreg var den at ei høveleg ordbok til dette føremålet låg ikkje føre, og ein hadde dessutan omsynet til dei to språkformene å ta vare på. Etter ein del drøftingar som ordboksredaktør Alf Hellevik hadde med representantar for Universitetet i Bergen, vart det etablert eit samarbeid mellom universitetet og Norsk språknemnd, og ved Nordisk institutt, Prosjekt for datamaskinell språkbehandling, vart det sett i gang eit arbeid med å setje opp eit dobbeltspråkleg ordarkiv i maskintilgjengeleg form. Etter ein grundig analyse av norske bøyingskategoriar vart ei klasseinndeling fastsett der ein både tok omsyn til bokmål og nynorsk og valfridomen av former innan desse. Etter dette vart orda i nokre skoleordlister kodifiserte, morfemdelte og puncha. Ved sida av dette ordtilfanget er heile bøyingsverket som låg til grunn for klasseinndelinga, skjematisert og overført til datamaskinen.

Dette dataarkivet over norsk ordtilfang er kalla *Norsk ordregistrant*. Føremålet er at ein i dette dataarkivet skal kunne registrere kva som til kvar tid er vanleg, normert norsk. Frå arkivet skal ein kunne ta ut ord i kva for bøyingsformer ein vil, ein kan setje opp morfemordlister, ordlister ordna etter bøyingskategoriar, eller vanlege baklengsordlister. Ein kan vidare få analysert likskapar og ulikskapar mellom bokmål og nynorsk. For tida er Norsk ordregistrant på om lag 40 000 oppslagsord, men eit nytt stort tillegg er under arbeid.

For å registrere nytt ordtilfang som kjem inn i dei ulike nordiske språka, har språknemndene i Danmark, Noreg og Sverige i fleire år arbeidd med innsamling av materiale. (Arnulv

Sudmann: Nordisk språksamarbeid. Språk i Norden 1970, s. 94 ff.) Utdrag frå desse samlingane kan ein finne i tidlegare årgangar av Nordiske språkspørsmål.

I Noreg valde ein også her å dra nytte av moderne teknikk. Då Språknemnda likevel skulle setje opp eit maskinskrive setelarkiv, tok ein opp til vurdering om det ikkje var mogeleg i same arbeidsprosessen å gjere materialet maskintilgjengeleg. Ein fann då at med å skaffe ein spesialskrivemaskin, kunne ein skrive på setel og punche samstundes. Etter ei stund gjekk ein frå dette med å ta vare på setelen til bruk i eit handarkiv, og i samarbeid med Universitetet i Bergen arbeider ein no datamaskinelt ut lister over dette nyordstilfanget, som kan vere grunnlaget for vurderinga når det gjeld den endelege publisering av materialet. Samstundes byggjer ein på denne måten opp eit dataarkiv over ord med kontekst og kjeldetilvisingar som mellom anna vil vere ei viktig kjelde for det vidare oppbyggingsarbeidet av Norsk ordregistrant.

Det er ei heil rad med andre datamaskinelle språkprosjekt rundt om i Norden som er interessante og som fortente ein grundigare omtale. Særleg er det i denne samanhengen grunn til å nemne det viktige arbeidet som i fleire år har vore i gang ved Forskningsgruppen för modern svenska, Universitetet i Göteborg. Ein har der arbeidd med utforskninga av moderne svensk avisspråk. (Sture Allén: Nusvensk frekvensordbok. 1 og 2.) Nemnast bør også at ved Universitetet i Århus har ein skaffa Nudansk ordbog på magnetband og arbeider med å framstille ei baklengsordbok. Også ved Universitetet i Helsinki er det i gang eit prosjekt med å lage ei slik baklengsordbok for finsk.

Ved Universitetet i Bergen er ein i ferd med truleg for første gong å databehandle eit materiale frå moderne islandsk avisspråk. Vidare er det også i gang større arbeid som gjeld nyare norsk avisspråk og skjønnlittteratur. Eit prosjekt som kanskje særleg bør nemnast er innsamling og databehandling av alt stoffet sendt ut over Norsk Telegrambyrå i tidsrommet 1.2—30.4 1971. Her er m.a. registrert dei ulike kjeldene telegrambyrået nyttar og får omsett meldingar frå.

Fleire av dei arbeida som er omtala her, både dei ved språknemndene, universiteta og andre forskningsinstitusjonar, kunne vere nyttige når det gjeld eit samarbeid om nordiske språkspørsmål. Det verdfulle materialet for studiet av ordlagingsprinsipp i nordisk ei magnetbandordbok ville vere, har ofte tidlegare vore omtala på dei nordiske språkmøta.

I kva utstrekning hindrar dei ulike ortografiske bileta at vi les kvarandre sine bøker i Norden? Frekvensordlistene kan med klåre tal syne kor mykje like ord med ulike skrivemåtar dominerer ein tekst, og viser kva som kanskje kunne gjerast.

På det nordiske språknemndmøtet i 1954 peika professor K. G. Ljunggren på at ei av dei viktigaste årsakene til at dei nordiske språka har skilt lag i ordtilfanget, er at ein ofte har fått ulike nemningar på nye omgrep. Derfor var det viktig å halde auge med dei nye orda som dukkar opp i språket. Ein av dei viktigaste innfallsportane i språket for slike nye omgrep er gjerne avisspråket, og med nyordstilfanget og dei datamaskinelle avisspråkprosjekta skulle ein få materiale for både å studere orda sjølv og omfanget, frekvensen, av dei ulike typane.

Den første innfallsporten for framand påverknad er kanskje likevel telegrambyråa. Språket her kunne vere interessant å studere for å finne kor opent for framand påverknad eit "hastverksspråk" er, og kvar påverknaden kjem frå. Telegrambyråa fargar som kjent avisene sterkt i våre dagar.

Teknisk fagterminologi og eit mogeleg nordisk samarbeid om nye termar har vore drøfta i språknemndene. Kvart av dei nordiske landa har sine eigne organ som arbeider med desse spørsmåla, men det synest klart at også språknemndene har eit ansvarsområde i dette arbeidet. I Noreg har ein hatt drøftingar mellom dei interesserte partane i terminologiarbeidet, og ein vurderer no om ein skal prøve å lage ein datamaskinell fagord-registrant. Ein kan tenkje seg at terminologiske ordlister etter kvart som dei blir utarbeidde, skal gå inn med alle opplysningar i eit dataarkiv. Frå dataarkivet kunne ein då ta ut selektive ordlister, gjere jamføringar for å skape konsekvens i terminologien, og ha eit effektivt hjelpemiddel i arbeidet med

nye termar.

Kunne ein leggje dette opp som eit nordisk samarbeid gjennom dei nordiske språknemndene, ville mykje vere vunne. Eit fullstendig ordfellesskap er det vanskeleg å tenkje seg, men berre det å ha ein nordisk termbank som kunne sikre at termar lett kunne jamførast, ville vere eit stort framsteg.

Av denne korte gjennomgangen av datamaskinelle språkprosjekt skulle det gå tydeleg fram at ein ny teknikk er i ferd med å gje oss høve til eit betre og nærare språkleg samarbeid i Norden. Dei nordiske språknemndene tok dette tidleg opp til vurdering, noko er blitt gjort på dei ulike stadene, men med tanke på eit nordisk samarbeid er enno det meste ugjort.