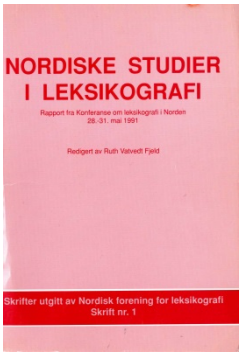
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Dokumentasjonsprosjektet ved Det historisk-filosofiske fakultet, Universitetet i Oslo

Forfatter: Christian-Emil Ore

Kilde: Nordiske Studier i Leksikografi 1, 1992, s. 403-408
Rapport fra Konferanse om leksikografi i Norden, 28.-31. mai 1991

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) 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.

Christian-Emil Ore

Dokumentasjonsprosjektet ved Det historisk-filosofiske fakultet, Universitetet i Oslo

Det historisk-filosofiske fakultet ved Universitetet i Oslo har det siste året gjennomført et forprosjekt og drevet annet forberedende arbeid i forbindelse med det såkalte dokumentasjonsprosjektet. Dette prosjektet har som mål å bygge opp databaser over arkivmaterialet ved fakultetets samlingsavdelinger. Samlingsavdelingene er oldsaksamlingen (arkeologiske samlinger), folkemusikksamlingen, avdelingene for etnologi og folkloristikk samt leksikografi-målføre- og navnegranskingavdelingen ved Institutt for nordistikk og litteraturvitenskap.

Prosjektets idé er å lage databaser for de enkelte fagområdene, men på en slik måte at det er mulig å samkjøre dem. Åpenbare felles innganger til dataene er tid og sted. Gjenstandene i de arkeologiske samlingene har et funnsted og kan som oftest dateres til en periode. Avdeling for navnegranskning har arkiv over norske stedsnavn som er sted- og tidfestet. Navn kan også gi verdifull informasjon for arkeologer og historikere, og særlig for språkforskere. Det skulle ikke være nødvendig å drive videre utbrodering av nytten ved muligheten av en krysskobling av arkivene.

Den oppgaven vi har gitt oss i kast med, er ikke av de minste. Samlet er det tale om 12-13 millioner arkivkort og dokumenter, hvorav det meste ligger i de leksikografiske avdelingenes skuffer. Bokmåls- og nynorskavdelingen har hver i overkant av tre millioner ordbokssedler, mens målføre- og gammelnorskavdelingen har omlag to millioner sedler tilsammen. Det meste av arkivmaterialet er uegnet for optisk lesing og må rett og slett skrives av. Registreringsarbeidet vil samlet kreve 600-700 årsverk. Det er klart at det ikke er mulig å få ekstraordinære midler til å ansette registreringspersonell. Tanken har derfor hele tiden vært å bruke arbeidsledige til registrering, og å ansette studenter og ferdige kandidater til å overvåke arbeidet. Denne modellen er nå under utprøving. Sommeren 1991 ble det ansatt fire forskningsassisstenter, og i august ble de første 8 arbeidsledige satt i virksomhet.

Oppbygging av de leksikografiske databasene

For å få erfaring har vi begynt dataregistreringen ved de leksikografiske avdelingene samt ved myntkabinettet. Elektroniske myntkataloger er et interessant felt, men siden temaet er leksikografi, skal vi la myntene hvile og konsentrere oss om ordene. I resten av denne artikkelen vil jeg prøve å forklare hvordan vi har tenkt å løse den formidable oppgaven vi har gitt oss i kast med. Gammelnorskavdelingen og nynorskavdelingen er brukt som eksempler, idet deres materiale representer hver sin ytterkant av arkivene. Til slutt vil jeg belyse en del problemer og valg forbundet med de datatekniske løsninger. Drøftingen er holdt på et generelt nivå og burde være av interesse for alle som ønsker å bruke datateknikk av noe omfang.

Gammelnorskavdelingens arkiv; et seddelarkiv over et avgrenset korpus

Gammelnorskavdelingens arkiv består av 700 000 ordsedler basert mest på litterære tekster, men også noe på diplomer. Sedlene har allminnelig ekserptseddelutforming. Et eksempel er vist i fig. 1.

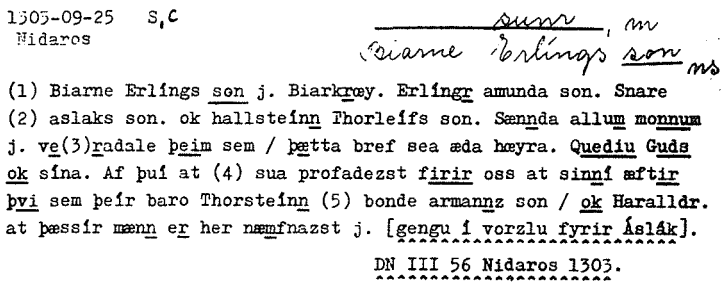


Fig. 1 Original seddeltekst fra gammelnorskarkivet

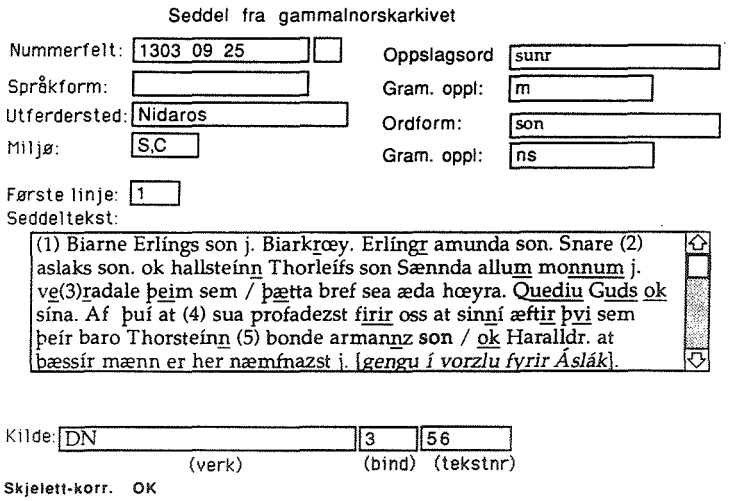


Fig 2 Seddelen skrevet inn i registreringsskjemaet

Før man starter registreringen av en slik seddelmasse, er det viktig å foreta en grundig analyse av sedlenes form og innhold. Det er vel og bra å finne hvilke felter som er nødvendig. Men det er like viktig å undersøke hvilke data som deles, dvs. er de samme på mange sedler. For å få dette klart frem lønner det seg å bruke et datamodelleringsverktøy. Vi har valgt NIAM (Nijssen 1988) og vil forlange at alle databasene i prosjektet beskrives i dette systemet. Se

også (Hjort 87) for bruk av formelle datamodelleringsverktøy innen leksikografi. I det følgende blir det gitt en noe uformell analyse av gammelnorsksedlene.

En kontekst er brukt til 30-40 sedler. Det er derfor bare rundt 30 000 forskjellige kontekster blant de 700 000 sedlene. Det er altså et stort rasjonaliseringspotensiale. Videre finner man ut at en slik seddelkontekst er entydig bestemt ut fra den trykte kildens tittel, hvilket bind og hvilken side konteksten er hentet fra, samt hvilken linje på denne siden konteksten begynner med. Tallene i parenteser i konteksten angir linjenumrene i den trykte kilden (se fig. 1). Ved å legge inn en test i databaseprogrammet er det dermed mulig å sikre seg mot at en kontekst blir skrevet mer enn en gang. Innskriverne vil først gå gjennom sedlene og bare skrive av kontekstene. Når dette er gjort, kan den sammenhengende teksten i et verk rekonstrueres ved at kontekstene sorteres etter sidenummer og linjenummer. I databasen vil imidlertid kontekstene få en fortløpende intern nummerering i henhold til den rekkefølgen de registreres i. I tillegg denne tabellen lages det en tabell over kildehenvisninger (som kunne rasjonaliseres ytterligere) som vist i tabell 4 under.

Tabell 1:

Gram.oppl	Ordkl.	Ordform	Norm.form
.	.	.	.
.	.	.	.
ns	m	son	nei
.	.	.	.
ns	m	sunr	ja
.	.	.	.

Tabell 2:

Norm.form	Ordform
.	.
sunr	son
.	.
sunr	son
.	.
.	.

Tabell 3:

Ordform	Kontekstnr	B.pos.	S.pos.
.	.	.	.
son	12342	122	124
son	845	17	19
.	.	.	.
.	.	.	.

Tabell 4:

Kontekstnr	Verk	Tekst	F.linje
.	.	.	.
.	.	.	.
1	Thomas	1	1
12342	DN 3	56	8
.	.	.	.

Fig. 3 Skissemessig fremstilling av noen av tabellene i gammelnorskdatabasen

Kontekstene på sedlene er hentet fra trykte utgaver, så man kan spørre om hvorfor vi ikke rett og slett skriver av de trykte kildene. Svaret er at det er lagt et stort arbeid i å justere kontekstene i henhold til originalmanuskriptene.

Når alle seddelkontekstene til et verk, f.eks. en saga, er skrevet inn, starter innskrivningen av ordopplysningene på de enkelte sedlene. Disse står i øvre høyre hjørne (se fig.1 og 2) og er "koblet" til konteksten ved at ordformen er understreket (med rødt på originalsedlene). Denne koblingen ønsker vi å bevare. I innskrivingsskjemaet (fig. 2) markerer innskriverne ordformen i teksten på skjermen på Macintosh-manér. I databasen er det så en tabell som til hver ordform forteller hvor ordformen forekommer (tabell 3 i fig. 3). Her lagres kontekstnummeret og start- og sluttposisjon i konteksten. Ved å koble tabell 3 og tabell 4 i fig. 3 kan man få frem verk, bind, side og linje som i en alminnelig konkordans, men også

ordformens nøyaktige posisjon i teksten. De grammatiske oppplysningene og forbindelsen ordform og normalisert form kan lagres som vist i tabell 1 og 2. i fig. 3.

Registreringen av sedlene gir en database over sedlene, altså en elektronisk variant av det tradisjonelle arkivet. I tillegg åpnes en rekke andre innganger: søking etter ordformer etter grammatiske opplysninger samt fritekstsøking. Det fullstendig nye er at tekstene er koblet til sin egen ordliste. På dertil egnede datamaskiner (Mac, PC med Windows) kan man nå markere et ord og umiddelbart få vist frem seddelinformasjonen om ordforekomsten. Denne såkalte hypertekst-muligheten kommer rett og slett som en bivirkning av registreringen av sedlene.

Nynorskavdelingens arkiv; et mangeartet seddelarkiv

Nynorskavdelingens arkiv er svært forskjellig fra gammelnorskavdelingens både når det gjelder størrelse (3,2 millioner sedler) og innhold. Sedlene er basert på ekserpter fra litteratur, presse og eldre ordbøker, men også på opplysninger om muntlig og skriftlig bruk av ord fra informanter spredd rundt i Norge. Det er således ikke mulig å velge samme løsning som for gammelnorsk. Det er også liten hensikt i å starte på A og skrive av hele arkivet. I stedet har vi valgt å behandle sedlene i henhold til innhold og opprinnelse. Som nevnt, faller sedlene i tre grupper: sedler med ekserpter fra ordbøker, sedler ekserpter fra litterære kilder samt målføre- og informantsedler. Vi har begynt med sedlene med ekserpter fra ordbøkene til Aasen og Ross. Disse finnes allerede som deler av et grunnmanuskript for Norsk ordbok. Manuskriptet er fra 1930-årene og er på 13 000 maskinskrevne sider. Ordboksteksten blir for tiden tagget på en tilsvarende måte som *Oxford English Dictionary* (OED)(Tompa 1988) og skrevet inn i med et vanlig tekstbehandlingsprogram. Dette er tilstrekkelig siden teksten ikke inneholder posisjonskoblinger og kopier av like tekstfragmenter slik som gammelnorskarkivet (og alle andre seddelsamlinger basert på sammenhengende tekster). Et utsnitt av den taggede teksten er vist i fig 4.

> /OPPF hamp-åker /GRMr m, /DEFI åker der det veks hamp, /KjFd Å. /OPPF hamra /HONr I. /GRMr v. /BØYr (a) /DEFI arbeida på med hamar. /MÅLF òg "hambra" /HMFm (sumst.), /KJFd Å. /OPPH Jfr. hamar I. Jfr. nisl.

Fig 4 En bit av det taggede manuskriptet til *Norsk ordbok.*

Når ordboksteksten er ferdig innskrevet, blir den analyseres og senere lagt opp som en database. Til innskrivningen er det i skrivende stund engasjert en forskningsassistent og fire personer på sysselsettingsmidler.

Det neste steget i registreringsarbeidet er å gå løs på registreringen av informant-sedlene. Hvordan sedlene med ekserptene fra de litterære kildene skal behandles, er ennå ikke bestemt. Her gjelder det å tenke som om oppgaven skulle gjøres på nytt. Det er mulig at den beste løsningen vil være å lese optisk hele referansebibliotektet og analysere dette maskinelt.

Det endelige målet for både nynorsk- og bokmålsdelen av prosjektet er å lage en autoritativ database over det norske skrift- og talemålet som støtte for ordboksproduksjon, men også som et hjelpemiddel for annen språkforskning.

Fritekstsøking kontra databaser

For å gi våre avdelinger en viss følelse av hvilke muligheter datateknikken kan gi, har vi tatt for oss de elektronisk tilgjengelige versjoner av manuskriptene til *Bokmålsordboka* og *Nynorskordboka*. De manuskriptene vi har fått, er stort sett løpende tekst full av settekoder, men det har vært foretatt noe analyse og feltinndeling. Videre er det i manuskriptene en del tilleggsinformasjon om godkjente sideformer. Disse opplysningene er ikke med i de trykte utgavene. Vi har analysert tekstene og tilrettelagt dem for fritekstsøk. I tillegg har vi laget en sterkt redigert versjon som danner basis for våre databaser.

Det å lage en fritekstsøkeversjon av manuskriptene innebærer en en viss tagging. Dette ble i vårt tilfelle gjort automatisk. Resultatet ble et system der man kan søke etter vilkårlige tekststrenger i ordbøkene og få vist resultatet i et "pent" format på skjermen. Vi har her brukt programvare utviklet ved University of Waterloo, Canada, til bruk i forbindelse med OED (Tompa 88). Det var ganske enkelt å få laget søkeversjoner av de norske ordbøkene. Metoden må sies å være lite kostnadskrevende hva angår informatikerinnsatsen. Man får altså et slagkraftig søkeverktøy uten stor innsats. Men resultatet er lite fleksibelt i den forstand at det ikke er mulig å få skrevet ut statistikker og tabeller uten innsats fra profesjonelle programmerere.

Å legge ordbøkene opp som databaser, dvs. plassere opplysningene i tabeller som vist for gammelnorsk, er mer krevende. Tekstene er ordboksmanuskripter. De er ment å skulle leses som løpende tekst. De inneholder også en del inkonsistens både på formatplan og i det innholdsmessige. Noe annet er ikke å vente, og som manuskripter betraktet er de svært ryddige. Men de interne særegenhetene gjør det vanskelig å omforme manuskriptene til fullverdige databaser. Vi har likevel kommet langt i databaseoppbyggingen, men ideelt sett burde en leksikograf gå gjennom databasene og rette inkonsistenser. Et annet punkt er hvor langt man kan fjerne seg fra ordboksformen og likevel kunne rekonstruere en ordbok automatisk fra databasen. Fordelen med slike databaser er at det åpnes for å foreta søk og lage tabeller hinsides de mulighetene som ligger i en fritekstdatabase.

Når en ordbok skal legges opp som en tradisjonell database, kan imidlertid selve databaseverktøyet bli en kompliserende faktor. Det synes å være mindre informatikk-krevende å velge en enbrukerløsning på en PC eller Macintosh enn den flerbrukerløsningen vi har valgt. Databasene er lagt opp i INGRES, som er et profesjonelt flerbrukersystem. Fordelen med dette systemet er at mange brukere kan bruke databasene (både lese og skrive) samtidig, samt at ulike brukere kan gis ulike privilegier. Ulempen er at opplegg, drift og vedlikehold av en slik database krever spesialopplært personell. Tross de større informatikkmessige kostnadene mener vi likevel at det er best å velge denne siste løsningen. Det gir oss muligheten til å knytte leksikografene til én database der databasesystemet automatisk kontrollerer tilgangsprivilegier og takler at flere brukere gjør forandringer i databasen samtidig. Orddatabasene kan dermed også bli tilgjengelige for brukere utenfor de leksikografiske avdelingene samtidig som det er mulig å kontrollere adgangen.

Litteratur

Hjort, Ebba. 1987. The Danlex-Group, *Descriptive Tools for Electronic Processing of Dictionary Data* Lexicographica Series Maior 20 Max Niemeyer Verlag, Tübingen 1987

G.M. Nijssen, T.A.Halpin. 1988. *Conceptual Schema and Relational Database Design*, Prentice Hall, New York 1988

Tompa, F. Wm., D.L.Berg, G.H.Gonnet. 1088 *The New English Dictionary Project at the University of Waterloo.*

Centre for the NOED. University of Waterloo, Waterloo, Canada