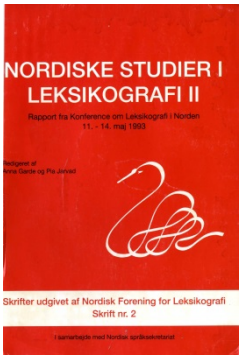
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Blant fire millioner sedler - En situasjonsrapport fra Dokumentasjonsprosjektet

Forfatter: Christian-Emil Ore

Kilde: Nordiske Studier i Leksikografi 2, 1993, s. 243-247
Rapport fra Konference om leksikografi i Norden, 11.-14. maj 1993

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.

Blant fire millioner sedler
En situasjonsrapport fra Dokumentasjonsprosjektet

Christian-Emil Ore

Det store arbeidet som er lagt ned i oppbygningen og vedlikeholdet av seddelsamlinger vitner om hvilket viktig hjelpemiddel ordsedlene har vært i ordboksarbeid. I de senere år har datateknikken muliggjort en mye mer effektiv oppbygning av den informasjonen som et seddelarkiv representerer. Elektronisk lesning av tekst (OCR), konkordansprogrammer og hjelpemidler for (halv-) automatisk markering av grammatisk informasjon på ord i løpende tekst kan nevnes [se Atkins 1992]. Tilstedeværelsen av denne effektive teknikken gjør det fristende å konsentrere seg om kilder som for en relativt billig penge kan gjøres elektronisk tilgjengelig eller som allerede er påelektronisk form. En slik dreining kan lett medføre at verdifull informasjon i de manuelle arkivene blir neglisjert fordi den er tungt tilgjengelig.

Dette problemet er selvfølgelig ikke bare begrenset til leksikografien, men finnes innen en rekke humanistiske fag. Dokumentasjonsprosjektet er et fellesprosjekt ved de fire universitetene i Norge som skal bøte på noe av problemet. Det går i korthet ut på å omforme papirarkivene ved en rekke samlingsavdelinger (muséer o.l.) ved de humanistiske fakultetene til elektroniske form. Prosjektet vil gå frem til 1998, og målet er å bygge opp "Universitetenes databaser for språk og kultur". Basene vil i første omgang omfatte de 12-14 millioner arkivkort og dokumenter ved samlingsavdelingene, men vil senere omfatte store bildebaser samt elektroniske kart for å lette koblingen av de forskjellige databasene. Prosjektet vil forhåpentligvis muliggjøre en helt ny krysskobling av data ved de ulike samlingene.

Leksikografidelen av Dokumentasjonsprosjektet

Innen leksikografi finnes det i Dokumentasjonsprosjektet tre delprosjekter, et for hvert av feltene bokmål, gammelnorsk og nynorsk. Dette avspeiler den tilsvarende tredelingen av avdeling for leksikografi. I alle de tre underavdelingene er det store ordseddelsamlinger, til sammen syv millioner sedler. Selv om den opprinnelige målsetningen for hele Dokumentasjonsprosjektet var "å gjøre seddelarkivene tilgjengelig på elektronisk form" er det ikke tale om å behandle mer enn fire millioner av disse sedlene. Dette skyldes at det finnes store mengder sedler som bare har den samme informasjonen som man vil finne i en vanlig KWIC-konkordans. Det ville være uansvarlig bruk av ressurser å skrive av disse sedlene. I stedet vil det bli brukt optiske lesere for å gjøre de opprinnelige

tekstene elektronisk tilgjengelige. Disse tekstene vil bli tagget i henhold til den SGML-baserte "Text Encoding Initiative" (Goldfarb 1991, Sperberg-McQueen and Bernard 1990)

Konverteringsarbeidet er nå i full gang for alle de tre avdelingene. Til sammen er det engasjert ni assistenter og omlag 100 ufaglærte innskrivere på hel- eller deltid. De sistnevnte er alle arbeidsledige som i enten deltar "Arbeid for trygd" tiltak eller er tilknyttet en av våre spesiallagde studie- og registreringssentraler i Nord-Norge.

Det aller meste av Bokmålsavdelingens sedler er av den nevnte KWIC-konkordanstypen, riktignok med ordklassemarkering. I tillegg til disse har avdelingen et seddelarkiv med opplysninger om omlag 350 000 ord i norsk sakprosa. Dette ekserptarkivet er nå skrevet inn og skal slåes sammen med et tilsvarende, allerede elektronisk tilgjengelig arkiv påbegynt i 1968 over nyere norske ord i aviser og tidsskrift. Det arbeides også med å gjøre et utvalg verker av norske forfattere elektronisk tilgjengelig.

Gammelnorskavdelingens sedler er også av konkordanstypen, men skiller seg fra de litterære sedlene for moderne norsk i minst to henseender. Teksten sedlene er basert på, er en korrigert utgave av de tilgjengelige trykte tekstutgavene av manuskriptene. Det er derfor ikke mulig å laste inn de trykte tekstene. Det kompliserer oppgaven. På den annen side inneholder sedlene både linje og sidehenvisninger til den trykte utgaven. Det er derfor mulig å legge sedlene inn i et databasesystem på en slik måte at den originale teksten kun blir skrevet en gang (se også Ore 1991). Idag er de 65 000 sedlene basert på "Thomas saga erkebiskups" skrevet inn i dette systemet. Vi har dermed en løpende tekst på omkring 65 000 ord der hvert ord er påført grammatiske opplysninger og normalisert grunnform. Innskrivningsarbeidet ble utført av fire ufaglærte i løpet av et halvt år. Dette er en beskjeden innsats når man tenker på hvilket arbeid som ligger bak det å lage ordsedlene. Når hele Gammelnorskavdelingens seddelsamling på i alt 700 000 sedlene er skrevet inn, vil vi ha en interessant base over grunnord og deres realisasjon i en rekke tekster. I tillegg til seddelarkivet arbeides det for tiden med å legge inn det samlede gammelnorske korpuset.

Det nynorske seddelarkivet representerer en blanding av det meste. Her finnes alt fra litterære belegg til sedler som kun gir opplysninger om at ord er hørt et sted i Norge. Arkivet er bygd opp over flere generasjoner og mange hundre informanter har bidratt. Å forsøke å laste alt dette inn i et formelt skjema kan synes vel dristig. Til dette er det å si at ikke alle sedlene vil bli skrevet inn. Vi frasorterer de fleste litterære sedlene og også alle sedler basert på dialektordlister. Disse vil enten erstattes av hele, elektroniske tekster eller vil bli skrevet inn under ett.

Innlegging av materialet

For gammelnorsksedlene har vi laget et innskrivningskjema som har vist seg meget effektivt. Ved de andre avdelingene har vi valgt å skrive inn hver enkel seddel som en løpende tekst med koder. Bokmålsavdelingen hadde allerede et ferdig kodeoppsett utviklet på 70-tallet. Så her var oppgaven også ganske enkel. For for den nynorske seddel-

samlingen eksisterte hverken en modell eller et kodeskjema. Da formaliseringsarbeidet startet våren 1992 hadde avdelingen en del erfaring i å kode ordboksmanuskripter. I samarbeid med Bokmålsavdelingen var det blant annet utviklet et feltskjema og et enkelt registreringsprogram (Felted) i forbindelse med utgivelsen av håndordbøkene "Nynorskordboka " og "Bokmålsordboka". I dette systemet finnes det lagret omkring 30 000 sedler med opplysninger om ekserpter fra sakprosa. Men materialet i det store arkivet viste seg å være mer sammensatt og uensartet enn de enkle formaliseringssystemene kunne takle. Formaliseringsarbeidet har strukket seg over vel ett år og det har vært mye prøving og feiling underveis. Men assistentgruppen på nynorskavdelingen har gjort et kjempearbeid (se Hagen og Ragnsæter 1993). Vi har nå både et SGML-taggesystem for arkivet og en formell datamodell. Det er også utarbeidet et hundresiders instruksjonshefte for koding av sedlene (Hagen, Haukaas og Ragnsæter 1993).

Datamodellen

Det første skrittet mot en database er å foreta en grundig analyse av de dataene som skal legges inn. Analysen bør resultere i en datamodell beskrevet i et formelt datamodelleringsverktøy. Dette kan riktignok gjøre jobben vanskeligere idet en formell beskrivelse må være pinlig korrekt. Men det gir en større sikkerhet mot inkonsistenser enn bruk av et uformelt feltskjema. Vi bruker datamodelleringsverktøyet NIAM (Nijssen 1988), men det finnes en lang rekke metoder (se Hjort & al 87).

I figuren er det tegnet et forenklet bilde av den formelle datamodellen for nynorskmaterialet. Ovalene representerer hovedkategoriene. Linjene indikerer forholdet mellom dem. "Buntene" betyr mange-til-mange relasjoner. Et oppslagsord kan være forbundet med mange ordformer, og en ordform kan stå i forbindelse med flere oppslagsord (f.eks. andre hovedformer eller sidestilte former). En enkelt linje forteller om en en-til-en forbindelse, mens de sprikende linjene angir en-til-mange forbindelser. Boksene er brukt for å antyde hva slags tilleggsinformasjon som er lagret om elementene i de enkelte kategoriene.

Den venstre delen av diagrammet viser sammenhengen mellom data i et tagget tekstkorpus der ordene er kodet med normert form og morfologiske opplysninger. Den nederste ovalen i diagrammet representerer samlingen av alle tekstene, mens den nest øverste ovalen representerer de enkelte ordformene slik man finner dem i tekstene. Forbindelsen mellom disse er gitt ved ordformens aktuelle plassering i en tekst. Den øverste ovalen er de normerte oppslagsordene slik man vil finne dem i en ordbok. I verdenen av ordsedler vil forbindelsen mellom denne og den nest øverste ovalen svare til forbindelsen mellom det normerte ordet øverst på ordseddelen og det understrekte ordet i seddelteksten. Sånn sett beskriver den venstre delen av diagrammet hvordan opplysningene på sedlene i gammelnorsk- eller bokmålsarkivet henger sammen.

Nynorsksedlene er derimot mer komplekse. De kan ha illustrasjoner, det kan stå ekstra kommentarer osv. Dette er illustrert i diagrammet ved at vi har en egen kategori "seddel". Nynorskarkivet skal dokumentere både talemål og skriftspråk. Dette betyr at

det finnes mange sedler som viser at et ord er brukt et sted i Norge. På en slik seddel vil det normalt være et eller flere sitater som viser ordet i bruk, samt en unormert grunnform av det. Denne målføre-grunnformen vil skille seg fra både den normerte oppslagsordet og fra den aktuelle (bøyde) ordformen ieksempelet. Vi har derfor introdusert en tredje ordkategori kalt "målføre-grunnformer". Til denne knyttes opplysninger om sted, hva ordet betyr, bøyning

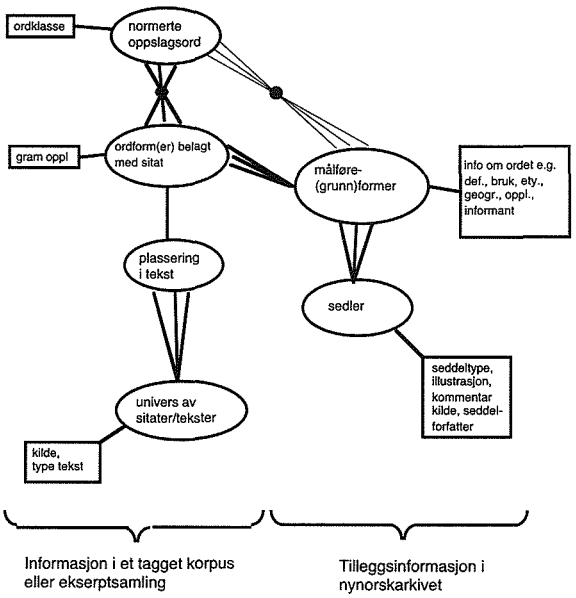


Fig. 1 En forenklet skjematisk fremstilling av informasjonen i seddelbasene

og liknende. Det hender også at det ikke står noe sitat på seddelen. Det er derfor en direkte forbindelse mellom målføre-grunnformen og den normerte utgaven av ordet. Kategorien "målføre-grunnformer" inneholder også sammensetninger og ordsamband. I tillegg finnes det sedler der målføre-grunnformen ikke er ført opp, men hvor det står opplysninger om bruk. Slike sedler blir kodet som om grunnformen eksisterte, men feltet vil bli stående tomt i databasen.

Seddelvirkeligheten er adskillig mer kompleks enn det denne korte gjennomgangen kan gi inntrykk av. Men figuren gir hovedtrekkene i databasene som bygges opp på grunnlag av de eksisterende seddelarkivene (også bokmål- og gammelnorskmaterialet). Diagrammet kan også illustrere at et (morfologisk tagget) tekstkorpus kan settes mer eller mindre rett inn i systemet. For folk som ikke er vant til å tenke i bokser og piler er dette kanskje ikke selvinnlysende. Tekstene i et slikt korpuset havner under kategorien "Univers av sitat/tekster", mens de enkelte ordformene med grammatiske opplysninger går inn i ovalen "ordform(er) belagt med sitat". Elektronisk tilgjengelige ordbøker vil også kunne føyes til ved at oppslagsordene grupperes under "normerte oppslagsord".

Ordartiklene vil da bli informasjonspakker forbundet med disse. Slik kan man ad libitum fortsette å bygge opp en kompleks dataverden.

Videre arbeid

Leksikografidelen av Dokumentasjonsprosjektet har som hovedmål å lage elektroniske hjelpemidler som kan erstatte avdelingenes bruk av sedler. Idag er konverteringen av de eksisterende seddelsamlingene godt igang, mens oppbygningen av elektroniske tekstsamlinger er i startfasen. Slik sett går prosjektet etter planen. Det mangler imidlertid noe "å vise frem". Det meste av vår tid har vært brukt til å bygge opp innskrivningskapasiteten og metoder for formalisering av materialet. I de neste to årene vil vi arbeide mer med verktøyer og systemer for å bruke dataene i det praktiske, leksikografiske liv. Vi håper å kunne gi en positiv rapport om dette på konferansen i Reykjavik i 1995.

Litteratur

Atkins, S. *The Hector Project*, i "Proceedings of Complex '92, Budapest 1992

Goldfarb, C. *The SGML Handbook*, Oxford University press 1991

Hagen,K., Haukaas,J., Ragnsæter, O. *Instruksjonshefte for innskriving og tagging av ordsetlar*, Oslo 1993

Hagen,K., Ragnsæter, O. *Nynorskdelen av dokumentasjonsprosjektet* Språklig samling 2 Oslo 1993

Hjort 87 The Danlex -Group, *Descriptive Tools for Electronic Processing of Dictionary Data* Lexicographica Series Maior 20 Max Niemeyer Verlag, Tübingen 1987

Ore, C.-E. *Dokumentasjonsprosjektet ved Det historisk-filosofiske fakultet, Universitet i Oslo*, i Nordiske studier i leksikografi, konferanserapport, Oslo 1991

Sperberg-McQueen, C.M., Bernard, L. (eds) *Guidelines for the Encoding and Interchange of Machine-Readable Texts (TEI P1)*, Chicago and Oxford Noveber 1990