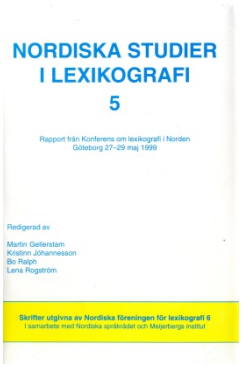
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Metaordboken - et rammeverk for Norsk Ordbok?

Forfatter: Christian-Emil Ore

Kilde: Nordiska Studier i Lexikografi 5, 2001, s. 250-270
Rapport från Konferens om lexikografi i Norden, Göteborg 27.-29. maj 1999

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

Metaordboken – et rammeverk for Norsk Ordbok?

Innledning

På slutten av 1980-tallet vokste det frem en ide ved Det historisk-filosofiske fakultet ved Universitetet i Oslo om å la data-maskinene overta for arkivskapene i en rekke samlingsavdelinger og museer. Dette noe uspesifiserte, men på det daværende tidspunkt svært ambisiøse og fremtidsrettede utopiske målet ble til det noe mer jordnære Dokumentasjonsprosjektet. For de leksikografiske avdelingene skulle prosjektet lette ordboksarbeidet ved å erstatte ordsedler og annet materiale med databaser og ikke minst den måten gjøre informasjonen i arkivene tilgjenglig for et større publikum.

De teknologiske forutsetningene i verden rundt oss har utviklet seg voldsomt i de 7 årene prosjektet varte. Det som kunne virke utopisk i 1990, er i dag hverdagslige realiteter. Vi er alle knyttet til Internett der vi kan hente ut informasjon om de fleste emner. Det foregår også en stadig økende elektronisk publikasjon av ordbøker, tekstarkiv og tekstkritiske kommentarutgaver både via Internett og via CD-ROM. Men det som kanskje er viktigere for leksikografer og forfattere av andre typer oppslagsverker, er at man nå ser de de første økonomiske konsekvensene av Internett. Det foregår for tiden vesentlige endringer i måten de store forlagene publiserer sine store ordbøker og andre oppslagsverk på. Verken *Oxford English Dictionary* (OED) eller *Encyclopaedia Britannica* vil lengre komme i papirform. Sistnevnte finnes nå bare i en reklamefinansiert gratisutgave på Internett. Tilsvarende strukturendringer har nå også kommet i Norge der vårt største forlag for oppslagsbøker, Kunnskapsforlaget, av økonomiske grunner legger om sin utgivelsesstrategi for sitt største leksikon fra papir til Internett.

Hvordan kan så et stort og tradisjonelt ordboksverk som *Norsk Ordbok* klare seg i denne moderne heksegryten? *Norsk Ordbok* er et nasjonalt ordboksverk som redigeres og utgis i tradisjonen fra begynnelsen av 1900-tallet. Internett-alderen vil neppe ha noen direkte økonomiske konsekvenser slik den har for de kommersielle oppslagsverkene idet salgsinntektene ikke utgjør noen vesentlig del av finansieringen. Men indirekte vil det kunne ha alvorlige konsekvenser siden særlig yngre mennesker forventer å finne informasjon gjennom elektroniske kanaler. Dette kan forverres ytterligere av den langsomme publiseringstakten. Arbeidet med *Norsk Ordbok* har til nå vart i nesten 70 år, og med dagens ordartikkelvolum vil man ikke bli ferdig før rundt år 2060. Sett i et historisk lys er dette lange tidsperspektivet langt fra enestående for et nasjonalt ordboksverk. Men en slik trøst setter strenge krav til den idealistiske tålmodighet hos oss nålevende mennesker. Det er derfor ikke urimelig om man ønsker å søke etter løsninger som ikke bør være til hinder for en realisering av det store planlagte verket, men samtidig kan formidle mer stoff, om enn i en noe ubearbeidet form, til dem som venter.

I denne artikkelen prøver jeg å belyse en mulig løsning på problemet gjennom å etablere det jeg vil kalle en metaordbok. Dette er riktignok ingen ordbok eller bok i det hele tatt. Metaordboken er en blanding av en leksikalsk database, et elektronisk seddelarkiv, et historisk ordboksbibliotek og et ordboksredigeringsverktøy med et voksende manuskript. En slik hybrid vil etter min mening være nødvendig for å kunne omplassere de store nasjonale ordboksverkene i en nåtidig omgivelse.

Ideen til metaordboken har vokst frem gjennom arbeidet med Dokumentasjonsprosjektet (1991–1997) og har hentet mye spesielt fra arbeidet med elektronisk utgivelse av eldre tekster som middelalderdiplomer og 1800-talls litteratur i Dokumentasjonsprosjektet. Det største enkeltprosjektet innenfor Dokumentasjonsprosjektet var likevel å lage en elektronisk erstatning for seddelarkivet og annet bakgrunnsmateriale for *Norsk Ordbok*. Dette arbeidet startet i august 1991 og ble i hovedsak avsluttet i 1998.

Metaordboken ble i sin første form foreslått som en mulighet i en artikkel skrevet i forbindelse med et innlegg på seminaret "Norsk Ordbok – nynorskens kanon" i 1996 (se min artikkel i Vikør 1997). Denne artikkelen er en viderutvikling og utdypning av seminarartikkelen. Det har vist seg at utviklingen igjen har gått fortere enn ventet. Metaordboken er også kommet nærmere en realisering siden vi i 1999 har hatt penger til et mindre pilotprosjekt gående for å prøve å bygge opp ryggraden i metaordboken. Det og mulighetene for videre arbeid vil bli behandlet til slutt i denne artikkelen.

Siden Dokumentasjonsprosjektets arbeid er vesentlig for muligheten av å realisere en metaordbok, har jeg gjort kort rede for prosjektet og hva databasene inneholder. Hovedvekten er lagt på en gjennomgang av forholdet mellom arkiv, kommenterte tekstutgaver og annen vitenskapelig publikasjon i en elektronisk tidsalder og hvilke innvirkninger dette kan og bør ha for et prosjekt som *Norsk Ordbok*.

Norsk Ordbok og seddelarkivet

Arbeidet med *Norsk Ordbok* har til nå vart i nesten 70 år. Prosjektet har vokst langt ut over de forestillingene man hadde i 1930 om å sammenstøpe ordtilfanget i de da eksisterende ordbøkene for nynorsk og utvide dette med nyere ord til et firebindsverk. I dag er målet "ei vitskapeleg ordbok over dei norske målføra og det nynorske skriftspråket" planlagt i 12 store bind.

Det store skiftet i utgivelsesplanen synes å ha kommet rundt andre verdenskrig da man erkjente at det da ferdige Grunnmanuskriptet ikke uten videre egnet seg som grunnlag for et trykkmanuskript, og at det tilfanget av nyere ord utgjorde en adskillig viktigere og større del av ordboken enn det man i utgangspunktet hadde forestilt seg. En del av denne erkjennelsen kan muligens ha kommet fra arbeidet med å bygge opp et leksikografisk seddelarkiv i forbindelse med prosjektet. Et typisk kort i et slikt arkiv inneholder et ord i grunnform, en liten tekstbit som gir et eksempel på en bruk av ordet, hvor ordet er brukt, opp-

lysninger om grammatikk, uttale og eventuelle andre forhold rundt dette eksempelet. Arkivet er sortert alfabetisk etter ordenes grunnform. Om man leser forordet til de ulike heftene som til nå er kommet av *Norsk Ordbok*, slår det en at oppbygningen av seddelarkivet har vært svært sentral hele tiden helt til det siste tiåret. Det har vært et uttalt ønske om å få arkivet størst mulig for nettopp å kunne ha mange eksempler for hvert eneste ord, også de mer sjeldne. Seddelarkivet rommer nå omlag 3,2 millioner sedler. Dette er ikke spesielt mye. Tilsvarende nasjonale prosjekter i Sverige og Danmark har mer en 10 millioner hver, mens arkivet til *Oxford English Dictionary* rommer mer enn 30 millioner sedler.

Seddelarkivet er altså den tradisjonelle systematiske metoden for å samle belegg og opplysninger om ord til bruk i redigeringen av en ordbok. I de siste 40–50 årene har imidlertid datateknologien muliggjort alternative metoder for å fange inn og lagre tilsvarende informasjon. Den nye teknikken har skapt et skille i synet på bruk av materialet i ordboksredigeringen. For noen står seddelarkivet som en tilfeldig samling opplysninger eller deler av en konkordans, og arkivet er kun et eksempel på hva gårsdagens teknikk kunne produsere. For andre representerer seddelarkivet en skattkiste der hver seddel er valgt med omhu, mens det med et korpus kan være vanskelig få med det spesielle og sjeldne språket.

Dokumentasjonsprosjektet

Formålet med Dokumentasjonsprosjektet var å bidra til å omstille samlingsavdelingenes behandling av informasjon til moderne datateknikk og dermed effektivisere det interne samlingsarbeidet, eksternt samarbeid og utveksling av informasjon samt innhenting av ny informasjon. På denne måten ønsker man både å utløse et forskningspotensiale, men også å avdekke svakheter i rutiner og systemer og forbedre disse. Det var også et viktig mål å tilgjengeliggjøre informasjonen for andre forskere, for studenter, for undervisning, for offentlig forvaltning og for allmennheten

så langt dette var forsvarlig ut fra personvern, sikkerhet, opphavsrettigheter og eventuelle kommersielle hensyn.

Dokumentasjonsprosjektet kan grovt deles inn i en museumsdel og en språklig orientert del. I den museumsrettede delen av prosjektet arbeidet vi med å bygge opp forskningsdatabaser for de arkeologiske museene i Bergen, Oslo og Tromsø. I Tromsø arbeidet vi også med nyere kulturhistorisk materiale. Den språklige delen bestod i hovedsak av tilrettelegging av bakgrunnsmateriale for ordboksavdelingene (bokmål, gammelnorsk og nynorsk) ved Avdeling for Leksikografi i Oslo og for Trønderordboka ved Norges teknisk-naturvitenskapelige universitet (NTNU) og for navnegranskingsmiljøene i Oslo og Tromsø.

Dokumentasjonsprosjektet ble avsluttet i 1997. Museumsarbeidet ble i 1998 forsatt gjennom et større Museumsprosjekt. Arbeidet med de språklige samlingene blir videreført av en liten etterorganisajon i perioden 1998–2001. Etter den tid vil antakeligvis samlingene bli ivaretatt av et "Senter for språklige datasamlinger" ved Universitetet i Oslo.

Dokumentasjonsprosjektet og de leksikografiske arkivene

Den opprinnelige målsetningen for delprosjektene ved den daværende Avdeling for leksikografi ved Institutt for nordistikk og litteraturvitenskap på Universitetet i Oslo var "å gjøre seddelarkivene tilgjengelig på elektronisk form". Til sammen er det omlag 9 millioner sedler i avdelingens arkiver. Ved en mer inngående analyse av seddelmaterialet ble det besluttet bare å konvertere nyordsarkivet til Bokmålsavdelingen, seddelarkivet til Gammelnorskavdelingen og hele nynorskarkivet (Seddelarkivet til *Norsk Ordbok*), i alt 4 millioner sedler. De resterende sedlene var stort sett laget ut fra hele forfatterskap (Henrik Wergeland, Bjørnstjerne Bjørnson mm.) og inneholder den samme informasjonen som man vil finne i en vanlig KWIC-konkordans. Det aller meste av Bokmålsavdelingens sedler er av denne typen. Istedet for å arbeide med ordsedlene har vi her brukt optiske leseprogrammer (OCR) for å gjøre de opprinnelige tekstene

elektronisk tilgjengelige. Disse tekstene ble kodet i overensstemmelse med anbefalingene til det SGML-baserte "Text Encoding Initiative" (Goldfarb 1991, Sperberg-McQueen and Burnard 1994). Tekstene kunne da også legges ut til almenheten i et lite elektronisk nettbibliotek på en tilsvarende måte som vi finner i det svenske Runeberg-prosjektet eller det amerikanske Project Gutenberg.

Seddelarkivet er som nevnt tidligere, den tradisjonelle systematiske metoden for å samle belegg og opplysninger om ord til bruk i redigeringen av en ordbok. Det store arbeidet som er lagt ned i oppbygningen og vedlikeholdet av seddelsamlinger vitner om hvilket viktig hjelpemiddel ordsedlene har vært og fremdeles er 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 til ord i løpende tekst kan nevnes (se for eksempel Atkins 1992).

I et leksikografisk seddelarkiv er det for et utvalg ord fra et utvalg tekstbrokker (en til 20 linjer) gitt opplysninger om bøyningsformen, ordets rot og annet. Et tagget tekstkorpus er derimot en samling større tekstfragmenter (fra 25 sider løpende tekst og oppover) der hver ordform i tekst(fragment)ene har fått markert grunnord, ordklasse og aktuell bøyningsform. Informasjonen i et seddelarkiv kan altså sammenlignes med den vi finner i et tagget tekstkorpus. Denne observasjonen lå til grunn da vi planla konverteringen av de leksikografiske arkivene og valgte å erstatte bokmålsarkivet med løpende elektronisk tekst, riktignok uten grammatisk informasjon.

Gammelnorskavdelingens arkiv er også et arkiv av "KWIC-konkordans" typen. Men sedlene er så systematisk bygd opp at vi her har valgt å bruke informasjonen på dem til å bygge opp hele tekster der hver ordform har fått markert grunnord, ordklasse og aktuell bøyningsform. Vi har altså laget et lite tagget tekstkorpus for det gammelnorske materialet.

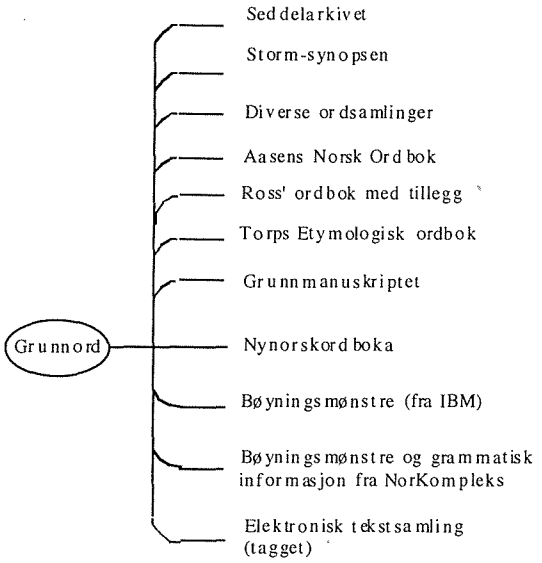
Arkivet til *Norsk Ordbok* skiller seg ut fra de to andre arkivene ved at det er bygd opp av mange hundre frivillige uten spesiell

leksikografisk utdannelse over en lang periode (omlag 60 år). Dette har resultert i et heterogent arkiv som både består av rene ekserptsedler og av sedler med mye ekstra informasjon om blant annet bruk og uttale. En effektiv konvertering av et slikt arkiv fordrer at de rene ekserptsedlene frasorteres og erstattes med elektronisk tekst. De resterende sedlene kan skrives inn og SGML-kodes (Sperberg-McQueen and Burnard 1994) slik at de kunne lastes inn i et databasesystem. Dette var også den opprinnelige planen for vårt arbeid. Etterhånden viste det seg at det var ekstremt ressurskrevende å foreta en sortering av arkivet. Det viste seg også at avskrift og koding av sedlene byr på de samme prinsipielle og metodologiske problemene som enhver transskripsjon av håndskrevet materiale. Vi valgte derfor gå bort fra avskriften av sedlene. I stedet satset vi på en database med faksimiler av samtlige nynorsksedler. Denne samlingen av faksimiler har oppslagsord, ordklasse og uttømmende kildeopplysninger som søkenøkler. Man mister på denne måten mulighetene til å søke i den løpende teksten på sedlene, men har fremdeles muligheten til å finne sedler etter grunnord, ordklasse, sted i landet, kildetype og hvem som har skrevet seddelen. Se (Ore 1996) for en inngående diskusjon av fordeler og ulemper ved denne løsningen.

Dokumentasjonsprosjektet og leksikalske databaser

Med leksikalske databaser menes databaser som inneholder informasjon om ord som deres oppbygning og bøying (morfologi), grammatisk funksjon, mening, relativ frekvens og så videre. En leksikalsk database er altså datateknikkens svar på en ordbok. Den skiller seg fra en tradisjonell ordbok også ved at det ikke er meningen at all informasjonen skal leses av mennesker. I mange tilfeller kan leksikalske databaser inneholde informasjonen kodet til bruk i ulike språkteknologiske verktøy så som morfologiske analysatorer, syntaksanalysatorer og oversettelsesstøttesystemer. Men en leksikalsk database vil typisk kunne inneholde teksten

fra en eller flere tradisjonelle ordbøker som hjelp og supplement til menneskene som bruker basen.



FIGUR 1. *En skisse av den leksikalske databasen.*

Dokumentasjonsprosjektets nynorske orddatabase (leksikalske database) inneholder informasjon fra mange kilder, seddelarkiv, ordbøker og ordsamlinger. Blant ordbøkene kan vi nevne Aasens *Norsk Ordbok* (1872), *Nynorskordboka* (1994) og Grunnmanuskriptet som er et første utkast til Norsk Ordbok fra 1930-tallet (se senere for en nærmere beskrivelse). I tillegg finner vi 20–30 ordsamlinger fra 1600-, 1700- og 1800-tallet. Man arbeider også med å legge inn den såkalte Storm-synopsen som er en oversikt over uttale og bøyning av omlag 1000 ord 1000 steder i Norge. I tillegg til dette historiske materialet inneholder orddatabasen også bøyningsinformasjon og annen grammatisk informasjon hentet fra IBM og fra arbeidet med et "Komputasjonelt leksikon (NorKompLeks)" ved Lingvistisk institutt ved NTNU (Nordgaard 1995). Dokumentasjonsprosjektets nynorske

ordbaser vil dermed være en blanding av en leksikalsk database i ordets mer tekniske betydning og av en samling av elektroniske versjoner av ordbøker, ordsamlinger og annet leksikografisk bakgrunnsmateriale. Figur 1 gir en skjematisk oversikt over den nynorske leksikalske databasen slik denne kan bli.

Et viktig punkt som krever norskfilologisk og helst leksikografisk ekspertise, er å etablere en kobling av grunnord mellom de ulike delbasene. Dette kompliseres av de ulike rettskrivingstandardene som er brukt. De eldre ordsamlingene og ordbøkene følger sin egen rettskriving. Seddelarkivet og Grunnmanuskriptet følger ideelt sett 1938-rettskrivingen, mens det nyere materialet følger moderne rettskriving. Sammenkoblingsarbeidet krever altså at alle oppslagsordene i de ulike samlingene og verkene også får påført en variant som følger den moderne rettskrivingen eller eventuelt 1938-rettskrivingen. Dette er et ikke-trivielt problem siden ingen av de nevnte normene dekker alle grunnordene i databasen. Fagkonsulentene i Dokumentasjonsprosjektet begynte standardiseringsarbeidet. Dette arbeidet har en nå tatt opp igjen i pilotprosjektet til metaordboken. Men med dagens utgivelsesstrategi for *Norsk Ordbok* vil den leksikalske databasen de facto være den eneste samordnede presentasjon av grunnlagsmaterialet fra A til Å i de neste 50 årene.

En frigjøring fra Gutenberg

Mange forskningsprosjekter vil vanligvis bygge på store mengder data som er samlet inn gjennom eksperimenter, gjennom feltarbeid eller som resultat av besøk i biblioteker og arkiver. På bakgrunn av disse dataene utarbeides det artikler og rapporter som presenterer de resultatene man har kommet frem til. Prosjektgenerert bakgrunnsmateriale blir så enten ødelagt eller bevart i et aller annet arkiv eller bibliotek. Et slikt arkivmateriale kan selv bli gjenstand for (meta)forskning ved en senere anledning der det har vokst frem et ønske om å finne ut hvordan prosessen mot det ferdige resultatet egentlig var.

Som den motsatte ytterlighet til formidling av forskningsresultater gjennom artikkelen, kan vi sette temaarkivet eller temabiblioteket. Det er en lang tradisjon innen arkiv og biblioteksverden å samle arkiver og bibliotek over spesielle temaer eller over enkeltstående personers etterlatte papirer. Her kan for eksempel nevnes Wittgensteinarkivet i Bergen, Kildeskriftsavdelingen ved Riksarkivet i Norge eller ved Stiftinga Nynorskkultursentrum (tidligere Ivar Aasen-stiftinga) der man vel originalt tenkte seg å samle alle Ivar Aasens etterlatte papirer, hans bibliotek og ideelt sett alt som er skrevet om eller på nynorsk.

Tekstkritiske utgaver av ulike verk og verk som prøver å etablere "en beste tekst" ut fra eksisterende versjoner, vil på denne tenkte skalaen kunne plasseres nærmere den vitenskapelige artikkel enn bibliotek og arkiv. Som eksempler på slike arbeider kan jeg nevne så ulike ting som utgivelser av Homer og bibelforskning. Den tradisjonelle publiseringen av større tekstsamlinger ved bruk av mikrofilm ligger derimot nær bibliotek eller arkiv på denne skalaen.

Denne klassifikasjon kan kanskje virke noe uinteressant om ikke introduksjonen av ny teknologi hadde snudd opp ned på en lang rekke tilvante publiseringsmetoder. Det har til nå vært høyst uvanlig at forskningsprosjekters bakgrunnsmateriale er blitt publisert selv i de tilfeller der institusjonen som har huset prosjektet, ikke har noen kommersielle eller sikkerhetsmessige grunner til å holde materialet for seg selv. Mye av årsaken ligger selvfølgelig i kostnadene. Men i en tid hvor det meste av ny informasjon finnes i elektronisk form, og hvor lagringsmediene (f.eks. CD-ROM) er svært billige, vil en slik publisering av bakgrunnsmaterialet sammen med resultatene av forskningen kunne bli alminnelig.

Innenfor de humanistiske fag har det i løpet av 1990-tallet blitt mer allminnelig å gi ut selve kildemateriale i elektronisk form utstyrt med søkeverktøy og/eller et elektronisk note- eller kommentarapparat. Denne trenden var tydelig allerede i 1996 (se Ore 1997). Slike publikasjoner er på mange måter svært like mikrofilmversjoner av et arkiv eller bibliotek, men synes likevel å være nærmere intensjonen bak tekstkritiske utgaver. Den publi-

seringsmetoden er også et svært godt alternativ til det tradisjonelle filologiske arbeid der kildematerialet kun refereres gjennom noteapparatet. Ved å legge selve bakgrunnsmaterialet sammen det vitenskapelige arbeidet, endres forholdet mellom kildematerialet og den vitenskapelige artikkelen. Som George Landow påpeker, får fotnoten eller henvisningen en annen rolle. Fotnotene er ikke lenger underordnede tekstfragmenter, men vil bli pekere inn i andre komplette tekster (Landow 1991). I stedet for å stå som en egen overordnet enhet vil på mange måter den vitenskapelige artikkelen kunne bli et kommentarverk som eksisterer i parallell med sine kilder. Artikkelen vil således bevege seg i retning av en spesialisert kommentert tekstutgave.

Muligheten for å publisere hele arkiver med et noteapparat vil på den annen side kunne viske ut forskjellen mellom arkiv/bibliotek og de filologiske tekstkritiske samlingsutgavene. Dette vil stoppe utgivelsen av de store trykte kommentert tekstutgavene med omfattende tekstkritisk noteapparat. Av åpenbare økonomiske årsaker, men også av praktiske årsaker vil de bli erstattet av utgivelser av arkiver der grunnmaterialet vil være mer eller mindre bearbeidede rådata i form av avskrifter eller faksimiler supplert med ulike fortolkede utgaver av de samme tekstene. Man kan også stille seg spørsmålet om i hvilken grad de svært forseggjorte noteapparatene i de tekstskritiske utgavene virkelig brukes av forskere til å rekonstruere alternative tekstvitner, eller om disse forskerne heller vil søke til (faksimiler av) originalmanuskriptene i det de ikke stoler på sin egen evne til å gjennomføre dette puslespillet nøyaktig nok (se Vanhoutte 1999 for en interesant debatt om dette). Vanhoutte opererer også med en arkiv/museums-modell, der arkivdelen svarer til arkivklassifikasjonen over, museumsdelen, eller i min tolkning utstillingsdelen, er den aktuelle tekstkritiske utgave.

I dag finnes det uttallige eksempler på elektronisk utgitte tekstarkiv og tekstarkiv koplet mot et kommentarapparat. Det fantes allerede i 1996 tre ulike Shakespeare utgivelser på CD-ROM. Det første store og gode eksempelet på elektroniske kommenterte tekstutgaver var "The Wife of Bath's Prologue" publisert som CD-ROM gjennom "The Canterbury Tales Project"

(Robinson 1996). Her er omlag 40 ulike manuskripter lemmatisert og sammenstilt med felles kommentarapparat og søkesystem. I Norden vil vel både den nye August Strindberg-utgaven, Søren Kierkegaard-utgaven og Henrik Ibsen-utgaven komme som elektroniske produkter.

Norsk Ordbok og de nye publikasjonstrendene

I det foregående avsnittet ga jeg en kort diskusjon om forholdet mellom edisjonsfilologi og elektroniske metoder for publisering. Har dette edisjonsfilologiske aspektet noen relevans for historisk orientert leksikografi?

Prosjektet "Norsk Ordbok" slik vi kjenner det i dag, er et tradisjonelt nasjonalt ordboksprosjekt med sterk vekt på historisk leksikografi. Verk av denne typen blir oppfattet som en autoritativ og uttømmende beskrivelse av betydning og bruk av ordene i et språk. Enhver filologisk skolert person vet at dette ikke er tilfellet. Men de fleste vil akseptere påstanden om at disse historiske nasjonalensyklopediene er det nærmeste man kommer en slik altomfattende beskrivelse av et språk. Mange moderne ordbøker er laget mot en bestemt målgruppe som for eksempel skoleelever eller (fremmedspråklige) studenter. Målgruppen for de store nasjonalensyklopediene er ofte noe mer diffus. En ordartikkel i de store verkene representer resultatet av redaktørens analyse av sitt kildemateriale der de ulike betydninger og ordets utbredelse er underbygget med kildehenvisninger på en tilsvarende måte som noteapparatet i et hvilket som helst annet vitenskapelig filologisk arbeid. For å forstå mange av artiklene fullt ut forutsettes det derfor et relativt høyt kunnskapsnivå hos leserne. *Norsk Ordbok* presenterer seg selv som en "vitskapleg ordbok" som åpenbart henviser både til metode for redigering og for anvendelsesområde. På dette grunnlaget kan det være riktig å sjangerbestemme *Norsk Ordbok* som vitenskapelig publikasjon.

Hvis man nå kan betrakte *Norsk Ordbok* som en slags antologi av vitenskapelige artikler, hvilke resultater vil vi få dersom vi appliserer de påståtte trendene fra forrige avsnitt? Prosjektet slik

det er drevet i dag, baserer seg direkte på et seddelarkiv, en målføresamling (Storm-synopsen se over) samt en rekke eksisterende ordbøker. Indirekte gjennom seddelarkivet baserer det seg på tusenvis av skriftlige kilder. Det er ikke urimelig å betrakte seddelarkivet som resultatet av den datainnsamling som foregår i ethvert empirisk prosjekt. Man skulle dermed kunne tenke seg en samlet publisering av både bakgrunnsmaterialet og artiklene som presenterer resultatene, det vil si ordboksartiklene. I dette tilfellet vil dermed artiklene være systematiseringer og kommentarer til bakgrunnsmaterialet. Kildehenvisningene vil være pekere inn i det originale materialet som er brukt i redigeringen. Man ville dermed få en enestående mulighet til å kunne studere hvordan ordboksredaktøren har arbeidet frem artikkelen. Siden Dokumentasjonsprosjektet har gjort hovedmengden av kildematerialet elektronisk tilgjengelig, er forutsetningene til stede for en slik kobling av ordboken mot kildematerialet.

Den andre påståtte trenden ovenfor er at det vil skje en økt publisering av arkivmateriale i elektronisk form utstyrt med søkeverktøy og et kommentarapparat. Slike utgivelser vil etter all sannsynligvis overta for de store mangebinds kommenterte tekstutgavene fordi de vil være billigere å utgi, kunne nå et større publikum, inneholde mer informasjon og være mer hensiktsmessig i bruk. Nå kan vel neppe *Norsk Ordbok* sies å være en kommentert tekstutgave, men den har klare islett av edisjonsfilologi. Men bakgrunnsmaterialet for arbeidet er et arkiv som det vil være interessant å publisere i sin egen rett. Jeg tenker her både på seddelarkivet som eksempler på bruk av ord, men også på alle de ordlister, ordsamlinger og ordbøker som direkte eller indirekte gjennom seddelarkivet brukes i arbeidet med *Norsk Ordbok*. Her ligger det klart i dagen et stort utgivelsesarbeid og venter: en fortrinnsvis kommentert utgave av alle kjente ordlister, ordsamlinger og ordbøker over (ny)norsk og norske dialekter supplert med arkivet til *Norsk Ordbok*. Mesteparten av grovarbeidet er allerede gjort gjennom Dokumentasjonsprosjektet.

Et eksempel på en slik blanding av ordbok og arkiv er "OED On Line" (OED 1996). Lik utgivere av andre gigantverk har Oxford Univeristy Press forsøkt å finne alternative utgivelses-

former. "OED On Line" ble i 1996 presentert som en abonnementtjeneste der man får tilgang til ordboken, men også til deler av bakgrunnsmaterialet som er brukt. Oxford University Press har også satt i gang et gigantisk revisjonsarbeid av OED, anslagvis 800–1000 årsverk. I 2010 vil det foreligge en fullstendig ny utgave av ordboken. Denne ordboken vil *foreligge bare i elektronisk form* og med muligheter for å se dagens ordartikler og utvalgt bakgrunnsmateriale som et supplement.

En ny strategi for utgivelsen av Norsk Ordbok – metaordboken

Norsk Ordbok-prosjektet ble startet i 1930. Arbeidet går jevnt fremover, men verket vil med dagens hastighet neppe være fullført før om 60 år. Arbeidet er svært tidkrevende. I andre land har det vært alminnelig å bruke mellom 50 og 100 år på tilsvarende verk. *Norsk Ordbok*-prosjektet er således på ingen måte unormalt i sin klasse, og den lange utgivelsestiden kan synes nødvendig. Jeg mener imidlertid at man ikke kan slå seg til ro med at det er nødvendig å vente i 60 år på fullføringen av *Norsk Ordbok*. At det er ressurskrevene å redigere en slik ordbok, er hevet over enhver tvil. Oxford University Press bruker hundrevis av årsverk på sin nye utgave. Men det mangler i Norge i dag en stor ordbok over nynorsk og dialektene. Jeg tror at det vil styrke nynorsksaken og dialektene å ha mest mulig tilgjengelig informasjon. Jeg mener videre at det er viktig å nå yngre mennesker gjennom de samme kanalene der de finner annen informasjon nemlig Internett. Spørsmålet er om en ikke kan kombinere det elektronisk tilgjengelige bakgrunnsmaterialet med modern teknologi og metoder hentet fra elektronisk edisjonsfilologi til både å øke produksjonstakten av den tradisjonelle utgaven, men samtidig lage et (elektronisk) verk som kan gjøre stoffet mer tilgjengelig.

Den nåværende redigeringsmetoden er den tradisjonelle sekvensielle gjennomgang av alfabetet der hver artikkel gjøres helt klar for trykk og hvor det er viktig å gjøre artikler mest

mulig stenografisk for at antall sider ikke skal bli for stort. De to største ulempene er den lange tidsperioden mellom første og siste hefte (100 år) og at artiklene kan være tunge å lese for legfolk. Metoder for å presse mest mulig informasjon inn på minst mulig plass har vært et sentralt tema i tradisjonell leksikografi. Til tider kan en spørre seg om en i leksikografien har evaluert denne komprimerte presentasjonen mer ut fra trykkøkonomiske hensyn enn ut fra pedagogiske hensyn. Satt litt på spissen: Det er en stor prestasjon å skrive Koranen på et frimerke, men det er dårlig formidlingskunst. Det er grunnlag for å diskutere om det er riktig å fortsette å bruke en redigerings og utgivelsesmetode som er så tett knyttet til papiret og trykkerkunsten og den skriftlige presentasjonmetoden slik vi har kjent den helt siden de hellenistiske filologene.

For å unngå misforståelser vil jeg understreke at det ikke et spørsmål om man skal utgi *Norsk Ordbok* i sin nåværende form, men heller om hvorledes utgivelsesprosessen best kan organiseres for at flest mulig nålevende skal få adgang til mest mulig informasjon om norsk på en måte som vil bli brukt. I mitt foredrag i 1996 brukte jeg som et 'memento mori' at det prestisjetunge *Encyclopedia Britannica* hadde en så sterk nedgang i salget av papirutgaven på 1980-tallet og begynnelsen av 1990 tallet at utgiverne besluttet å satse utelukkende på en nettbasert abonnementsordning. Høsten 1999 ble verket til og med relansert som en reklamefinansiert gratis nettjeneste. OEDs tredje utgave kommer som nevnt bare som elektronisk abonnementsutgave. Her i Norge ser det ut til at vårt største konversasjonsleksikon ender som et elektronisk abonnementsprodukt. De bindstore verkene nyter åpenbart ikke lenger den respekt de før fikk bare i kraft av sin fysiske størrelse. De ansees vel rett og slett for å være for uhåndterlig for folk flest.

Er denne dreiningen vekk fra de bindsterke verker et problem, og i så fall hvilke konsekvenser har det for utgivelsen av *Norsk Ordbok*? Man kan vel neppe forestille seg dette eller andre tilsvarende nordiske verk skal ende som reklamefinansiert nettjenester. Til det er nok den tradisjonelle lesergruppen for smal. Men burde man ikke vurdere om redigeringsstrategi kan endres

slik at den bedre utnytter de muligheter teknologien gir oss til å nå et større publikum samtidig som en sikrer det forsvarlig vitenskapelig nivå i redigeringen?

Ovenfor hevder jeg at vi ser en begynnende trend til en elektronisk publisering av kildemateriale med kommentarer og en elektronisk publisering av vitenskapelige essays og monografier sammen med kildematerialet. Kunne man for *Norsk Ordbok* foreta et bevisst valg og publisere verket i bredde først istedenfor som nå i dypde først, det vil si å tilby en ordbok fra A til Å i en gradvis økende finhetsgrad for hver utgave, istedet for å starte på A og lage hver artikkel helt ferdig? Kan man bygge opp en norsk nasjonalordbok ved å flette sammen hovedkildene til et verk for deretter å legge til systematiserte målføreopplysninger, etymologi og definisjoner?

Tanken virker besnærende. Men det kan innvendes at ideen har vært prøvd i 1930-årene under oppstarten av prosjektet *Norsk Ordbok* og i 1990-årene i forbidelse med prosjektet "Dansk Ordbok". Ingen av forsøkene ga særlig oppløftende resultater og planene ble oppgitt. Er det da noen hensikt å prøve på nytt? La oss først se på hva som ble gjort på 30-tallet og hvorfor planen ble oppgitt.

I innledningen til det første heftet av *Norsk ordbok* gir redaksjonen en kort skisse av den originale planen. Her heter det: "I samsvar med den planen som var lagd for ordboka, vart arbeidet skipa såleis at det fyrst vart laga eit grunnmanuskript, der alt tilfanget i Aasens og Ross' ordbøker (med alle tillegga), Schjøtts "Norsk Ordbok" og sume mindre ordbøker og ordsamlingar vart samanstøypt. Samstundes vart ordtydingane overførde til norsk. Grunnmanuskriptet var det så tanken å fylla ut med det nye tilfanget frå målføra og litteraturen, og såleis nå fram til det endelege prentemanuskriptet" (Førebels innleiing til *Norsk Ordbok*, 1950). Grunnmanuskriptet ble også laget og finnes i en maskinskrevet versjon på 13 500 sider. Når man studerer Grunnmanuskriptet, slår det en også at mange av definisjonene bærer preg av de danske originalene, samt at mye av særlig Aasens stringente oppstilling av ordartiklene har gått tapt. Stilistisk er vel heller ikke Grunnmanuskriptet helt på topp. Det

er åpenbart at store deler av manuskriptet måtte ha blitt renskrevet og nyere ord føyd til. Det var dermed ikke store spranget til dagens bruk av Grunnmanuskriptet. Det brukes nå som en viktig kilde når en ny ordartikkel skal skrives.

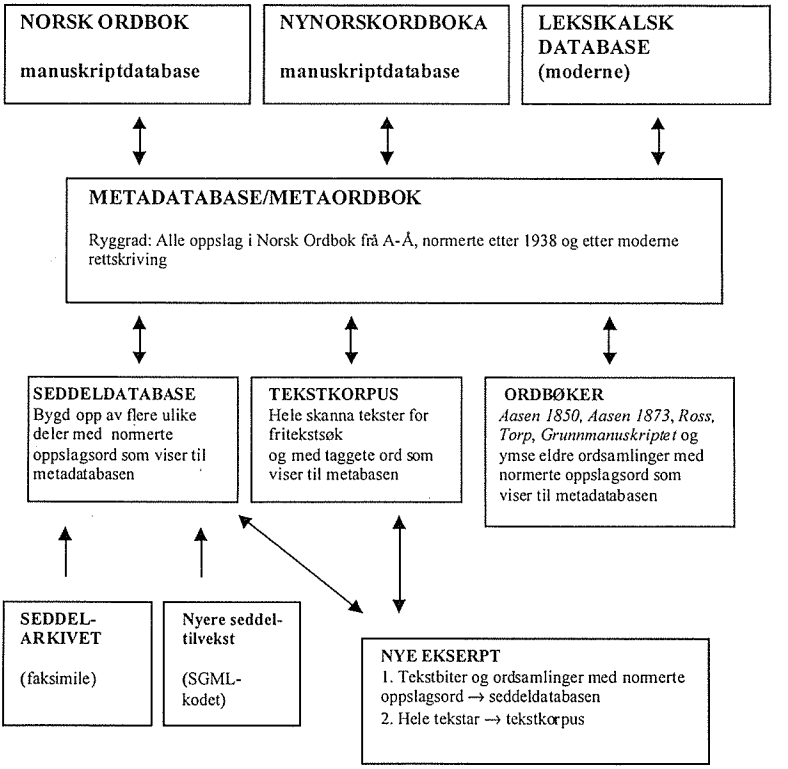
Prosjektet "Den danske ordbog" ble i 1991 (Hjort 1999) ble på sin side planlagt som et gjenbruksprosjekt eller sammenstøpingsprosjeket der man skulle gjenbruke ordboksdata fra ulike ordbøker for hurtig å lage en ordbok. Her kom redaksjonen til den erkjennelse at ordutvalg, uttale og retskrivning var foreldet og dels feilaktig i forhold til det innsamlede korpusmaterialet. Man ga i følge Hjort opp gjenbruksideen og bestemte seg for å skrive ordboken fra grunnen. Den optimistiske ideen bak "Den danske ordbog" viste seg altså å fungere like dårlig som den tilsvarende tanken bak "Grunnmanuskriptet til Norsk Ordbok" selv med moderne datateknologi.

I tillegg til disse to eksemplene på mislykket "resirkulering" av leksikografisk materiale synes det også å være en god del skepsis eller usikkerhet rundt bruken av datateknologi i leksikografisk arbeid. I *Nordisk leksikografisk ordbok* presiseres det under oppslaget "leksikalsk database" at "slike programmer [knyttet til en leksikalsk database] innebærer selvsagt ikke at ordboksarbeidet blir automatisert. Tvert imot er leksikografens aktive innsats svært viktig". Videre formulerer en stor kapasitet som Lars Vikør at "...for hovedsaka for oss [NOs redaksjon] er det intellektuelle grunnarbeidet, som ikkje all verdas datateknologi kan rasjonalisere bort om det skal bli ei ordbok" (Vikør 1999).

Det er altså veldokumentert skepsis til å bruke eldre ordboksmateriale til å lage nye ordbøker særlig ved hjelp av datamaskinelle metoder. Med metaordbokideen griper man imidlertid saken noe anderledes an enn i de nevnte eksempler. Metaordboksideen springer ut fra de nye trendene i edisjonsfilologien og vil nok ikke være egnet til arbeidet med en moderne definisjonsordbok. Den er ment som et hjelpemiddel til å gradvis kunne publisere materialet det Lars Vikør kaller en "flergenerasjonsordbok" (Vikør 1999) eller Matti Vilppula kaller enn "evighetsordbok" (Vilppula 1999). Som nevnt tidligere, har det de siste 5–

6 årene begynt å vokse frem en forståelse for at datateknologiens tilnærmet ubegrensede lagringskapasitet kan brukes til å publisere (elektronisk) grunnlagsmaterialet for et vitenskapelig prosjekt sammen med den vitenskapelige publikasjonen. Dette kan være interessant dels fordi man da kan se hva forskerne, i vårt tilfelle ordboksforfatterne, har brukt som grunnlagsmateriale. Dette gjør det mulig for andre for å forske i det samme materialet eller etterprøve resultatene. Denne tanken har særlig fått aksept innen edisjonsfilologien i de senere årene. Det store Henrik Ibsen-prosjektet som startet i 1998, er lagt opp på denne måten. OED har i sin nettversjon lagt inn en mulighet til å søke i bakgrunnsmaterialet. Tanken om å gjøre noe tilsvarende for *Norsk Ordbok* ble lansert allerede våren 1996 i et foredrag jeg holdt på et jubileumsseminar for Norsk Ordbok i Oslo, se Ore 1997.

Ryggraden i metaordboken vil være en liste med normerte (oppslags)ord. Den gjør det mulig å skaffe seg oversikt over ordtilfanget og å planlegge størrelsen på *Norsk Ordbok*. Ved hjelp av denne ryggraden vil det også være mulig å henge på de ulike kildene: Dette vil i og for seg svare til sammenstøpingstanken fra 1930 tallet. Men med "sammenstøping" i forbindelse med Grunnmanuskriptet mente man en sammenskriving og oversettelse av høyst ulike forfatteres verker. Dette er en vanskelig oppgave dersom man ikke kan stoffet svært godt og klarer å frigjøre seg fra språket i de ulike originalene. "Sammenstøping" eller påhektingen av originalkilder i metaordboken vil være en helt annen ting. Alle de originale kildene vil være med komplette, i sin opprinnelige form. Metaordboken vil i sin mest basale form bestå i at denne samlingen av kilder får felles inngang gjennom grunnordene. For hvert grunnord vil man få tilgang til alle de ordartikler eller avsnitt der ordet er behandlet eller forekommer. I tillegg vil det selvfølgelig være andre søkemuligheter som datateknikken gir oss så som tid, sted, type og generelt fritekstsøk. Dette er selvfølgelig ikke en ordbok – det er mange. Men i motsetning til Grunnmanuskriptet danner den et fundament for en ordbok. Det trenges i seg selv ikke renskrives til et trykkmanuskript.



FIGUR 2. *Skjematisk fremstilling av metaordboken.*

Å sette opp listen over ønskede oppslagsord for *Norsk Ordbok* er imidlertid ingen triviell oppgave. Som et pilotprosjekt er det i 1999 satt av 9 månedsverk til dette. Pilotprosjektet går ut på å gå gjennom det elektroniske seddelarkivet fra og med bokstaven *i* og etablere en foreløpig oppslagsordliste på bakgrunn av oppslagsordene på sedlene. Pilotprosjektet er ennå ikke avsluttet: Men arbeidet er interessant også fordi det setter de eksisterende redaksjonsreglene på en hard prøve og avslører problemer man ikke har tenkt på eller ikke ønsket å ta opp (ennå) i den tradisjonelle A-til-Å-redigeringsmetoden. Vi håper nå å få finansiering i 2000 og 2001 til å fullføre oppbyggingen av denne.

Ordartiklene i *Norsk Ordbok* slik den i dag redigeres, vil komme i tillegg til metaordboken og vil eksistere som et eget separat verk i parallell med sine kilder. Ordartiklene kan nå skrives i det omfang og i den rekkefølgen som er ønskelig. Metaordboksprosjektet gjør det også mulig å benytte maskinenes ubegrensede lagringskapasitet til å skrive ikke fullt så kompakte ordartikler. Man kan altså legge noe mer vekt på formidlingsaspektet og ikke så mye på antallet spaltemillimetere man har til disposisjon. Det er også mulig for en enkelt redaktør å vie seg til enkeltfelter som målføreopplysninger eller etymologi. Det bør være mulig å kunne vurdere utvalget av opplysninger i den trykte versjonen av *Norsk Ordbok* idet man kan lettere henvise til den elektroniske metaordboken for supplerende opplysninger og/eller legge inn for eksempel detaljerte dialektopplysninger i en senere utgave. En kan dermed øke tempoet gjennom alfabetet slik at en (første) gjennomgang kan gjøres i løpet av en 10–15 års periode.

Det store prosjektet "Norsk Ordbok" kan dermed deles opp i overkommelige delprosjekter som utgjør noen titalls årsverk hver. Et avsluttet delprosjekt vil resultere i en ny og større utgave. Men i motsetning til dagens situasjon vil hver utgave vil dekke hele alfabetet. Ordboken vil på denne måten vokse frem og alltid være temmelig ajour. Denne gradvise utviklingen av ordboken vil dermed være i samsvar med det den skal beskrive, nemlig et levende språk som hele tiden endres gjennom bruk.

Henvisninger

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

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

Delary P. and Landow G. (eds.). 1991. *Hypermedia and Literary Studies*. The MIT Press, London.

Nordgaard T. 1995. "NordKompLeks – et norsk komputasjonelt leksikon." Prosjektbeskrivelse. Trondheim.

Oxford English Dicitonary On Line. 1996. Oxford University Press, http://www.oed.com.

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

Ore, C.-E. 1996. "Korpus og seddelarkiv, fredelig sameksistens mellom det beste og det gode?" I: *Nordiske studier i leksikografi*. 3. Reykjavik.

Robinson P. (ed.). 1996. *The Wife of Bath's Prologue*. Cambridge. University Press, Cambridge.

Sperberg-McQueen, C.M., Burnard, L. (eds.). 1994. *Guidelines for the Encoding and Interchange of Machine-Readable Texts (TEI P3)*. Chicago and Oxford.

Vanhoutte, E. 1999. "Where is the Editor? Resistance in the creation of an electronic critical edition." I: *Human IT* nr 1, 1999. Institutionen Bibliotekshögskolan, Högskolan i Borås.

Vikør, L.S. 1999. "Fleirgenerasjonsordbøker og tida." I: *Nordiska studier i lexikografi*. 4. Helsingfors.

Vilppula, M. 1999. "Ordbok över finska dialekter och evigheten." I: *Nordiska studier i lexikografi*. 4. Helsingfors.