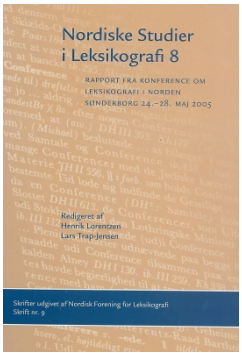
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Digital sats eller digital satsing?

Forfatter: Christian-Emil Ore og Lars Jørgen Tvedt

Kilde: Nordiska Studier i Leksikografi 8, 2006, s. 315-322
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 og Lars Jørgen Tvedt

Digital sats eller digital satsing?

Presentation or content? The Unit for Digital Documentation at the University of Oslo has been making dictionary writing systems (DWS) for more than ten years. In this paper we will present some of our ideas on how a modern DWS should deal with the source material, other dictionaries and cross references inside the dictionary, and what implications this will have on the implementation of such a system.

1. Innleiing

1.1. Om Eining for digital dokumentasjon

Eining for digital dokumentasjon (EDD) ved Universitetet i Oslo er direkte underlagt det Humanistiske fakultet (HF) og har som mål å hjelpe fageiningane ved HF med å lage digitale dokumentasjonssamlingar, tekstsamlingar, korpus og andre databasar.

EDD er eit produkt av Dokumentasjonsprosjektet som var eit stort samarbeidsprosjekt mellom dei humanistiske fakulteta ved dei fire universiteta i Noreg. Oppgåva til Dokumentasjonsprosjektet var å digitalisere dei store samlingsarkiva ved dei ulike HF-einingane.

Etter at Dokumentasjonsprosjektet vart avslutta, vart det naudsynt å etablere ein driftsorganisasjon som skulle vedlikehalde og vidareutvikle databasane som vart bygd opp i prosjektperioden, samt å utvikle nye verktøy for brukarane. EDD vart oppretta som ei permanent eining under HF. EDD har i tillegg til dei faste oppgåvene for HF også ansvaret for gjennomføringa av Museumsprosjektet, eit nytt stort nasjonalt digitaliseringsprosjekt som skulle bygge på dei røynslene som var gjort under Dokumentasjonsprosjektet.

Som ei følgje av dette har EDD ein forholdsvis liten fast tilsett stab som har hovudansvaret for oppgåvene som ligg til HF, og ein stor prosjektstab knytt til Museumsprosjektet. Trass i dette skiljet, utviklar EDD bevisst løysingar som er felles for HF og Museumsprosjektet, og ein prøver å utnytte kompetansen mellom dei to stabane til beste for begge partar.

1.2. Leksikografi

Arbeidet med dei leksikalske arkiva har vore ein sentral del av EDD sine oppgåver heilt frå tida i Dokumentasjonsprosjektet. Setelarkivet til Norsk Ordbok var mellom dei første arkiva som vart digitalisert tidleg på 1990-talet, og skanning og registrering av 3 millionar setlar tok mykje ressursar.

EDD var tidleg ute med å lage løysingar for å bruke dei leksikalske arkiva, og

utvikla også tidleg applikasjonar for redigering av ordbøker. Taggarprosjektet som var eit samarbeid mellom Tekstlaboratoriet ved HF og EDD, utvikla tidleg ein taggar basert på data frå Nynorskordboka og Bokmålsordboka. EDD laga også verktøy for redigering av dei første einspråklege ordbøkene i Zimbabwe (ALLEX-prosjektet), og har også stått bak redigeringssystema for Nynorskordboka og Bokmålsordboka. I dag jobbar EDD mellom anna med redigeringssystemet til Norsk Ordbok, og eininga har utvikla integrerte korpussystem knyta til redigeringsapplikasjonane basert på CQP/Corpus Workbench.

1.3. EDD sin filosofi

Gjennom arbeidet med samlingane har EDD fleire gonger påvist små og store manglar både ved måten arkiv er organisert på, og ved innhaldet i arkiva. Arkiva har ofte vore utforma på ein slik måte at dei berre kan brukast av ei lita gruppe menneske, og ikkje sjeldan er det store problem med oppdatering av arkiva. Ofte kjem ein over arkiv der tidlegare tolkingar av eit materiale vert kasta når det vert gjort nye tolkingar. Andre arkiv kan vere organisert slik at det ikkje er rom for nytolkingar. Eit døme på det siste er setelarkivet ved Norsk Ordbok som i papirutgåve var sortert etter ei normalisert skriftform basert på 1938-rettskrivinga av nynorsk.

Dei første digitale arkiva EDD laga, var nok på mange måtar prega av dei same problema, til dømes i høve til det å kunne ha rom for nytolking av materialet. Etter kvart har EDD lagt stor vinn på å forbetre dette, og ein har i fleire år vore aktiv deltakar både i nasjonale og internasjonale fora der ein prøver å lage standardar for datautveksling, samt diskutere løysingar på ulike problem knytt til forskingsdatabasar.

EDD har på bakgrunn av dette lagt vinn på å utvikle arkivløysingar som på alle vis kan karakteriserast som vitskaplege. Med dette meiner vi mellom anna at ein må lagre grunnlagsmaterialet i så lite bearbeidd form som mogleg. Dei mest atomære storleikane i eit arkiv bør vere så fri for tolking som mogleg. Dernest må alle hypotesar som er basert på dette grunnlagsmaterialet handsamast med like stor respekt. Arkivsystema må ta omsyn til at nye tider og nye forskarar vil ha nye tolkingar av kva som ligg i eit materiale. Og sist, men kanskje viktigast, arkivsystema må kople saman grunnlagsmaterialet og hypotesane slik at det er enkelt for nye forskarar å etterprøve resultata.

Ut over dette har EDD ønskje om å kunne knytte saman informasjon i ulike arkiv slik at forskarane kan dra nytte av kvarande sine resultat i langt større grad enn i dag.

2. Produksjon av tradisjonelle ordbøker

Produksjon av ordbøker vart tidleg ei viktig oppgåve for EDD, og EDD har vore med på å flytte fleire store prosjekt inn i framtida. Norsk Ordbok er absolutt det største, og vi vil bruke denne ordboka som eit døme på korleis dette har gått til.

2.1. Fokus på utsjånad

Før EDD kom inn i biletet, vart artiklane i Norsk Ordbok skrive i WordPerfect og Word. Ein nytta eit sjølvutvikla kodesystem for å merke opp teksten, og all oppmerking hadde som mål å styre utsjånaden på trykk.

Problemet med denne framgangsmåten er at ordboka sjølv ikkje gjer det lett for ein brukar, eller ein forskar, å etterprøve informasjonen. Rett nok har ein i alle vitskaplege ordbøker eit referansesystem til litteratur eller arkivmateriale som underbygg påstandane, men reelt sett er det ikkje mogleg for ein brukar å sjekke desse kjeldene med mindre brukaren sit i arkivet til ordboksredaksjonen. Sjølv om ein skulle sitte i arkivet, ville det vere omtrent umogleg å finne fram til dei setlane som har danna grunnlaget for ein artikkel, og ikkje minst finne ut kva setlar redaktøren ikkje brukte. Ordboka ville med andre ord formelt sett tilfredstille vitskaplege krav, men reelt sett ville ho i avgrensa grad bli utsett for vitskapleg kritikk fordi det ville vere eit alt for stort arbeid å gå til kjeldene.

Frå redaksjonen sin synsstad har også denne gamle produksjonsmåten mange ulemper. Mellom anna er handteringa av krysstilvisingar mellom artiklar eit evig problem. Det er ingenting som garanterer at ein definisjon som ein viser til, faktisk eksisterer. Stringens i form og innhald vert også meir komplisert enn det ein gjerne vil. Det einaste som styrer dette er redaktørane sin godvilje, og redaksjonen sin vilje til å overprøve einskildredaktørar sin måte å skrive artiklar på. Handtering av mange redaktørar vert også vanskeleg. Dess fleire redaktørar, dess vanskelegare vert det å halde styr på dei to føregåande punkta.

Gjenbruk av data frå ordboka er heller ikkje særleg lett med denne måten å redigere artiklar på. Det er svært lite struktur i innhaldet, og den som måtte ønskje å bruke ordboka til noko anna enn å slå opp definisjonar, måtte bruke mykje tid på å finne det ein leita etter, trass i at teksten var tilgjengeleg digitalt.

2.2. Fokus på struktur

Ut over 1990-talet vart det opplagt for dei fleste at dei gamle måtane å redigere ordboka på, ikkje gav samfunnet nok attende i høve til innsatsen. Krav til tilgjenge for ålmenta, gjenbruk av data, vitskapleg kvalitetssikring, og ikkje minst effektiv produksjon av ordboksmanuskript, sette fart i planane om å endre måten Norsk Ordbok vart redigert på. Alle krava ein stilte til den nye ordboka, peika i retning av meir fokus på struktur, mindre på utsjånad, samt at det måtte vere mogleg å kople ordboka sine artiklar mot grunnlagsmaterialet, og at ein trong administrative rutinar

bygd inn i systemet for å auke effektiviteten. I dag er det to teknologiar som peiker seg ut, strukturert tekst (SGML/XML) eller ei eller anna form for databaseløysing.

2.2.1. SGML/XML

EDD har lang erfaring i bruk av strukturert tekst. Sidan starten av 1990-talet har EDD arbeidd med SGML i oppmarkeringa av tekst. Då XML overtok for SGML, vart dette ein teknologi som vart tilgjengeleg for eit større publikum, og absolutt ein mogleg teknologi å bruke i utforminga av ordbøker.

Den største styrken til XML som format for eit ordboksmanuskript er kombinasjonen av stringens og fleksibilitet. XML gjer det mogleg å lage klare reglar for korleis strukturen skal vere, men kan tillate stor grad av valfridom i innbyrdes plassering av tekstelement. XML kan også på ein enkel måte opne for at einskilde element kan dukke opp på fleire plassar i eit dokument utan at strukturen lir under dette.

Det største problemet med XML er at dette ikkje er eit databaseformat. I utgangspunktet er eit XML-dokument ei samling tekstar som alle har fast struktur. For store dokument vil ein rein XML-struktur gjere referanseintegritet mellom delar av dokumentet tungt og tidkrevjande, og kopling til element som ligg utanfor dokumentet sjølv vil vere enno tyngre. Med XML er det ikkje mogleg å garantere at referansar til grunnlagsmaterialet i form av til dømes ordsetlar, er korrekte. XML kan heller ikkje hjelpe til med administrasjonen av ein stor redaksjon. Eit større ordboksprosjekt er avhengig av desse funksjonane.

2.2.2. Databaseløysingar

Hovudalternativet til XML er å lagre artikkelteksten i ein database av eit eller anna slag. Databasen sin største styrke er innebygde rutinar for dataintegritet. Dataintegriteten gjer det mogleg å sikre at krysstilvisingar til ein kvar tid er korrekte, at kopling til kjeldematerialet er på plass, osb. Databasesystem utan slik integritetskontroll er til lita nytte for eit ordboksprosjekt.

Databasen tilbyr struktur, men kan tidvis også stille for store krav til strukturering av artiklane i ei ordbok. Dette er den største innvendinga mot å bruke ein database som basis for sjølve artikkelteksten. Databasen har ikkje innebygde system for å markere at delar av ein laupande tekst har ei spesiell meining.

2.2.3. Kombinasjonsløysingar

Konklusjonen er at korkje ei rein XML-løysing eller ei rein databaseløysing fungerer for dei fleste større ordboksprosjekt. Ein må ty til kombinasjonsløysingar.

Det finst mange verktøy som i ulik grad løyser problema med XML ved å innføre ein eller annan form for databasestruktur rundt XML-dokumenta, og dette kan ofte vise seg svært fruktbart. Ofte nyttar ein XML som internt databaseformat i eit slikt

databasesystem, og dette kan gjere desse databasesystemet ope og lett å dokumentere, men dette er på langt nær opplagt. Mange av desse systema er også svært nye og lite utprøvde, dei er laga for mindre organisasjonar, og skalerer ikkje godt i store miljø.

Mange databasesystem i dag kan handtere XML eller XML-fragment i tekstfelt i databasen. Dette løyser litt av problema knytt til databasen sine ekstreme strukturkrav.

Om ein vel å basere seg på ei utvida XML-løysing, eller ei utvida databaseløysing, kjem an på kva eigenskapar ein meiner er viktigast. Eit spesialutvikla databasebasert system vil vere mest teneleg dersom ein skal ha store redaksjonar og god kontroll med referansar og tilvisingar i ordboka. Men dette krev at ein kan definere ein streng struktur for artiklane utan at dette går ut over den faglege kvaliteten.

Ei mindre ordbok med ein mindre redaksjon og mindre krav til stringens, vil tene på å bruke eit enklare, meir generelt XML-basert verktøy.

3. Den nye ordboka

I dei føregåande avsnitta presenterte vi dei krava vi meiner ein må stille til verktøy for produksjon av dei tradisjonelle ordbøkene. Men teknologien fører ikkje berre til at ein kan gjere tradisjonelle oppgåver på ein betre måte. Det kan like gjerne føre til at oppgåva eller det endeleg resultatet totalt endrar karakter. Tenkjer ein visjonært kan ein ofte med same arbeidsinnsats, få eit anna og mykje meir nyttig resultat. Dette ser ein tydleg innanfor mange humanistiske fag i dag, og ikkje minst innanfor språkforskinga.

Teknologien kan og gje eit omgrep eit heilt nytt innhald, og innhaldet knytt til omgrepet endrar krava publikum stiller. Dette ser ein tydeleg i samband med bøker generelt, og ordbøker spesielt. I Norsk Ordbok, band 1, som vart utgjeve i 1966, er ordet "bok" definert som ei samling innbundne, prenta ark. I dag kan vi lese på websidene til Norsk Ordbok at ein har som mål å kome med ei elektronisk utgåve av ordboka. Både Norsk Ordbok, men også store delar av det norske folk, meiner at ei samling av ord og definisjonar som ein kan finne på internett, er ei ordbok. I daglegtale er det ikkje lenger uvanleg å høyre om elektroniske bøker, eller lydbøker. Teknologien har gjort at ei bok i dag er noko heilt anna enn det ho var for 40 år sidan, og våre krav til bøkene speglar teknologien det er mogleg å bruke for å presentere innhaldet av boka. Den beste ordboka er ikkje berre den med dei mest korrekte definisjonane, men også den som utnytter den eksisterande teknologien best til å formidle innhaldet i definisjonane. Leksikografen har ikkje gjort jobben sin dersom han ikkje har utnytta teknologien optimalt når han formidlar kunnskapen sin om eit ord.

Dette legg eit stort press på dagens leksikografar når dei skal lage vitskaplege ordbøker. Frå Gutenberg og fram til 1990 måtte leksikografen ta stilling til

sidestorleikar, skrifttypar, spalter, osb. Dette var nokre av dei parametrane som styrte formidlinga. I tillegg hadde formidlingskanalen klare grenser for kva ein kunne ta med. Mangel på teknologiske løysingar førte til at ein sjølvsagt ikkje kunne prente alt kjeldemateriale, og ein kunne heller ikkje legge ved lyd eller video. Dersom ein kunne gjort dette ville den vitskaplege verdien av ordboka auka.

I dag set ikkje teknologien slike grensar. Det er fullt mogleg å lenke oppslagformer og dialektformer direkte til lydkjelder. Definisjonar og døme kan peike til tekst i eit korpus og ein kan legge inn formell syntaktisk informasjon i tillegg til definisjonar og døme. Det er mogleg å kople mot kjelder for ensyklopedisk informasjon, og ein kan legg inn informasjon om bøying av ord, informasjon om samansetting av ord, osb. Tekst, lyd, bilete og film i uavgrensa menge kan knytast til artiklane. Leksikografen si oppgåve blir å organisere all denne informasjonen slik at det er mogleg å finne fram i han.

Utfordringa er at brukarane av ordboka no venter å finne denne typen informasjon. Forskarar vil døme ordboka etter kor lett det er å etterprøve informasjon ved å følgje ordboka sin eigen dokumentasjon av påstandar. Oppdragsgjevarar og andre venter at det skal vere mogleg å trekke informasjon ut av ordboka, og bruke dette til talesyntese, maskinomsetting og dikteringssystem. Når det er teknisk mogleg å lage ei slik ordbok/leksikalsk database, så vil alle forvente at dette vert gjort.

Vi veit alle at det ikkje er mogleg å lage denne altomfattande databasen i dag, og sannsynlegvis vil ein heller aldri nå heilt i mål. Men det er ei plikt for oss å skjele til det som ligg i dagens og morgondagens teknologi når vi lagar løysingar. Ein kan ikkje forsvare å lage ordbøker i dag som ikkje kan inngå i ein større heilskap ved eit seinare høve. Det vil vere ein hån både mot fagmiljø, brukarar og dei som betalar for arbeidet.

4. Vegen mot vitskaplege leksikalske ressursar

Er det så i praksis mogleg å lage slike ressursar på eit slikt vis at alle vert nøgd? EDD har forsøkt, og trass i einskilde feilskjer undervegs, har vi kome eit godt stykke på veg.

Men utfordringa ligg ikkje primært i å lage gode system for spesifikke problem. EDD og Norsk Ordbok har i fellesskap skapt fleire ulike databasar som på ulikt nivå tilfredsstiller dei krava vi har sett til det vi kallar vitskaplege databasar. Dess enklare basane er, dess lettare er det å oppfylle desse krava. Databasane over setelarkiva til Norsk Ordbok er døme på banale basar i denne samanhengen. Basane er berre ein digital representasjon av gamle papirbaserte arkiv.

EDD har også utvikla databasane som Norsk Ordbank mellom anna brukar til å registrere bøyingsinformasjon for meir enn 100.000 grunnformar for dei to norske målformene, bokmål og nynorsk, men heller ikkje her ligg det særleg referanse til dømes til empiriske data.

EDD har også utvikla fleire ulike redigeringsapplikasjonar for spesifikke ordbøker. Dei fleste av desse er applikasjonar som ikkje gjer anna enn å registrere den teksten redaktøren vil ha på trykk, og på det viset framstår dei, etter våre krav til vitskap, ikkje som vitskaplege databasar. Det ligg lite eller ingen informasjon som gjer det mogleg å gå til kjeldematerialet.

Databasen for Norsk Ordbok har vi derimot forsøkt å bygge på ein slik måte at alle våre krav til vitskaplege metoder er oppfylte.

Metaordboka er også ein svært enkel konstruksjon, som enkelt forklart er ein indeks til alle dei andre leksikalske ressursane EDD og Norsk Ordbok rår over. Men sidan vi har brukt Metaordboka til å kople saman mange ulike arkiv, ordbøker og ordlister, står samlinga av mindre og enkle databasar i større grad fram som eit vitskapleg heile. Det er i dag mogleg å gå frå ein artikkel i Norsk Ordbok, via Metaordboka, til Norsk Ordbank, og få ut alle brukte bøyingsformar av ordet. Ein kan også gå frå handordbøkene, Nynorskordboka og Bokmålsordboka, via Metaordboka til Norsk Ordbok, og der få utfyllande informasjon om bruksmåtar, dialektformer, osb., og sjølvsagt vidare derifrå til å sjå på kjeldematerialet, anten i form av ordsetlar, eller i form av konkordanseliner henta frå eit korpus. Samankoplinga gjer det enkelt å navigere mellom dei ulike databasane, og dermed står dei einskilde basane fram med mykje større vitskapleg tyngde enn før.

5. Samarbeid

Frå dette forholdsvis enkle systemet for samankopling av data, ser vi at det er svært mykje å hente ved å kople saman ressursar. Vi har så langt prioritert å kople saman dei ressursane vi sjølv rår over. Vi har også planar om å legge inn eit større tal dialektordbøker i dette systemet, for ytterlegare å auke nytta av databasane.

I framtida ser vi at det kunne vore nyttig å vurdere alle dei store ressursane i Norden med omsyn på ei slik samankopling. Det er svært mange tunge og viktige databasar i Norden, og ved å kople dette saman vil sannsynlegvis leksikografar i heile området få langt meir ut av også sine eigne databasar.

Det er klart at dette vil krevje ein god del avklaring, både på den tekniske sida, men også når det gjeld spørsmål om opphavsrett, kommersielle rettar, osb. Vi ønskjer likevel å kome i gang med dette arbeidet, spesielt på den tekniske sida. Det hadde difor vore ønskjeleg at personar med teknisk kompetanse frå dei ulike fagmiljøa i Norden kunne samlast for å diskutere korleis ein skal gå fram for å utvikle metodar for utveksling av data. Eit mål må vere at ein slik samling blir gjennomført før neste konferanse om leksikografi i Norden.

Litteratur

Ore, Christian-Emil Smith 1998a: Hvordan lage fagdatabaser for humanistiske fag. I: *Fra skuff til skjerm. Om universitetenes databaser for språk og kultur.* Oslo: Universitetetsforlaget. ISBN 82-00-12670-6. 30 s.

Ore, Christian-Emil Smith 1998b: Making multidisciplinary resources. I: Lou Burnard, Marilyn Deegan and Harold Short (eds.) *The Digital Demotic, A Selection of Papers from Digital Resources in the Humanities 1997.* Publication 10, Office for Humanities Communication, King's College, London, 1998, ISBN: 1-897991-12-7.

Ore, Christian-Emil Smith 2001: Metaordboken – et elektronisk rammeverk for Norsk Ordbok? I: Martin Gellerstam et al. (eds.): *Nordiska studier i leksikografi 5. Rapport från Konferens om Lexikografi i Norden, Göteborg 26.-29. maj 1999.* Göteborg, 202-216.

Ore, Christian-Emil Smith; Tvedt, Lars Jørgen; Bjørnstad, Tone 2002: *The Meta Dictionary* ALLC/ACH 2002; 24.07.2002-28.07.2002. http://www.edd.uio.no/artikler/leksikografi/meta_dictionary.html

Text Encoding Initiative Consortium, C.M. Sperberg-McQueen and L. Burnard (red.) 2002: *TEI P4: Guidelines for Electronic Text Encoding and Interchange. Print Dictionaries*, Chapter 12, XML Version. Oxford. http://www.tei-c.org/P4X/

Web-sider (alle sider sjekka 22. september 2005):
ALLEX, http://www.dokpro.uio.no/allex/allex.html
CIDOC, http://www.willpowerinfo.myby.co.uk/cidoc/
Corpus Workbench, http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
Dokumentasjonsprosjektet, http://www.dokpro.uio.no/
Eining for digital dokumentasjon, http://www.edd.uio.no/
Museumsprosjektet, http://www.muspro.uio.no/
Norsk Ordbok, http://no2014.uio.no/
Tekstlaboratoriet, UiO, http://www.hf.uio.no/tekstlab/