

Den elektroniske utgaven av Fritznors ordbok og Menota

Christian-Emil Ore, Enhet for digital dokumentasjon, Universitetet i Oslo
16. september 2004

Enhet for digital dokumentasjon og Dokumentasjonsprosjektet

Det norske Dokumentasjonsprosjektet (Dokpro) ble gjennomført i perioden 1991–1997. I 1996 ble det lagt planer for en fase 2 av Dokumentasjonsprosjektet, men videreføringen ble dessverre avslått av det daværende Kirke-, forsknings- og utdanningsdepartementet. Prosjektorganisasjonen ble i 1998 erstattet av en liten driftsenhet, kalt Enhet for digital dokumentasjon ved HF (EDD eller Ø). EDD har siden 1999 drevet et omfattende databaseprosjekt for de norske kultur- og naturhistoriske universitetsmuseene.

Gammelnorsk materiale i Dokumentasjonsprosjektet

I Dokpro ble det konvertert en lang rekke tekster for Gammelnorsk Ordboksverk (GNO), blant andre *Diplomatarium Norvegicum* I – XXI, “nytranskriberinger” av norskspråklige diplomer fra 1349–1402 og en del norske sagatekster. I tillegg ble størstedelen av ordseddelarkivet ved GNO konvertert. Det sistnevnte materialet er nå under ferdigstilling som lemmatisert og bøyingskodet elektronisk tekst som skal gjøres tilgjengelig også gjennom Menota. Den norske delen av Menota-arbeidet kan på mange måter sees som en videreføring av ideene bak Dokpro.

Konverteringen av Fritznors ordbok

I planleggingen av fase 2 av Dokpro kom det frem et klart ønske om å få konvertert Johan Fritznors *Ordbog over det gamle norske Sprog* I–IV. Våren 1998 besluttet Bjørn Eithun ved GNO og Christian-Emil Ore ved EDD likevel å prøve å få konvertert Fritznors ordbok. Til alt hel hadde vi fra Dokpro en disponibel innskrivergruppe i Larvik. Gruppen var et arbeidsmarkedstiltak og bestod av 2–5 personer med ulike yrkeshemninger. Arbeidet med ordboken startet i september 1998 og ble avsluttet i oktober 2000. Eithun stod for opplæring og oppfølging av gruppen, EDD finansierte husleie og lokal ledelse og Ore stod for IT samt senere bearbeiding av materialet.

flugr, m. (G. -ar) **1) Flyven, Flugt,** =
flaug. *SE.* I, 212¹²; *Stj.* 270²¹; beina
flug dvs. **udbrede Vingerne til Flugt,** *SE.*
I, 80²⁰. 284²¹. **2) Flugt, Flyen, Flygten,**
= flótti 1; trauðr flugar *Hund.* I, 82;
glöggr flugar *Sig.* 1, 7. **3) hurtig Fart,**
= flug 1; vera á ferð ok flug (**jvf.**
för ok flaug *Heilag.* II, 386⁸) *Fld.*
I, 6⁷.

Figur 1: Utsnitt av MSWord-fil for side 446 i bind I

Ordboken er delvis satt i fraktur og trykket er uklart. Optisk lesing var utelukket. Alle de fire bindene ble derfor tastet inn for hånd. Innskriverne ble bedt om å gjengi ordboksteksten slik den stod med samme linje-, spalte- og sideinndeling som originalen. Side og spalteskift ble markert ved at innskriverne satte inn SGML-taggene ‘<SIDE = “nn”>’ og ‘<SPALTE>’. Det er ikke brukt halvfet skrift i ordboken. Vi kunne derfor la fraktur bli gjengitt som halvfet skrift. Teksten ble skrevet i MSWord med en trykkside pr fil. Da arbeidet var ferdig, satt vi med 3 481 filer i MSWord format. Noen få stikkprøver underveis hadde vist at innskriverne var ekstremt nøyaktige. Vi fant det derfor forsvarlig ikke å få lest en ekstra korrektur. En manuell korrektur vil være svært kostbar og neppe gjøre det ferdige produktet vesentlig bedre enn det vi har nå.

Videre bearbeiding av tekstfilene

En ting er å få skrevet av en ordbok, noe annet å vite hva man skal gjøre med teksten. I dette tilfellet var teksten spredt på 3 481 filer. Filene ble først konvertert til RTF-formatet. Deretter ble det kjørt et program på RTF-filene som satte inn formattagger for halvfet (fraktur), kursiv, hevet og senket skrift, se figur 2. Deretter ble RTF-filene konvertert til ren tekst. Tilslutt ble alle småfilene slått sammen bindvis til fire temmelig store filer.

flugr, <K>m</K>. (<F>G</F>. -ar) <F>1) Flyven, Flugt</F>, = flaug. <K>SE</K>. I, 212¹²; <K>Stj</K>. 270²¹; beina flug dvs. <F>udbrede Vingerne til Flugt</F>, <K>SE</K>. I, 80²⁰. 284²¹. <F>2) Flugt, Flyen, Flygten, </F> = flótti 1; traudr flugar <K>Hund</K>. I, 82; glöggr flugar <K>Sig</K>. 1, 7. <F>3) hurtig Fart, </F> = flug 1; vera á ferð ok flug (<F>jvf</F>. för ok flaug <K>Heilag</K>. II, 386⁸) <K>Fld</K>. I, 6⁷.

Figur 2: Teksten i figur 1 konvertert til rentekst-format med XML-koding for setteinformasjon.

På grunn av for stort arbeidspress ble ordboksfilene liggende altfor lenge. Men etter innstendig oppfordring fra Odd Einar Haugen besluttet jeg å gjøre filene såpass ferdig at de kunne legges i en database og publiseres på Internett. Jeg har i årenes løp analysert og autotagget en rekke større verker, for eksempel Oluf Ryghs *Norske Gaardnavn*, *Diplomatarium Norvegicum* og de fire første bindene av *Norsk Ordbok*. Det er ofte mulig å komme svært langt med slik maskinell tagging basert på tekstens originale format. For Fritzners ordbok ble tiden litt kort og analysen og taggingen foreløpig litt vel grov. I figur 3 er det vist et utsnitt av den taggete teksten som ligger til grunn for den eksisterende nettversjonen av ordboken.

```
<ART ID="8669"><FYL></FYL><OPP AID="8669" LID="8714" LEMMA="flugr"
GRM="m">flugr</OPP><FYL>, </FYL><GRM V="m"><K>m</K></GRM>.
(<F>G.</F> -ar) <DEF NR="1"><B>1</B> <F>Flyven, Flugt</F>, =<L>flaug.
<K>SE</K>. I, 212<SUP>12</SUP>; <K>Stj</K>. 270<SUP>21</SUP>;<UTR>
beina<L>flug dvs. <F>udbrede Vingerne til Flugt</F>, <K>SE</K>.<L>I,
```

80²⁰. 284²¹. <DEF NR="2">2 <F>Flugt, Flyen, Flygten</F>,<L> = flótti 1;<UTR> trau_r flugar <K>Hund</K>. I, 82;<UTR><L>glöggr flugar <K>Sig</K>. 1, 7. <DEF NR="3">3 <F>hurtig Fart</F>,<L> = flug 1;<UTR> vera á fer_ ok flug (jvf<L>för ok flaug <K>Heilag</K>. II, 386⁸) <K>Fld</K>.<L>I, 6⁷.</ART>

Figur 3: Teksten i figur 1 og 2 etter en foreløpig innholdsanalyse.

Noen av taggene kan virke overflødig. FYL-taggen er tatt med fra analysen av Norsk ordbok og er satt rundt tegn som i en fulltagget versjon er overflødige fordi strukturinformasjonen er representert ved de andre elementene. I eksempelet over er den ikke konsekvent satt inn. For en ny, moderne ordbok kan slik fylltekst settes inn ved hjelp av et "stylesheet" når teksten skal settes. Fritznors ordbok er et historisk dokument og dessuten ikke helt konsekvent med hensyn på bruk av slik markering. Det siste gjelder for så vidt de aller fleste moderne ordbøker også. I teksten er det også satt inn tagger som markerer hovedbetydninger og tagger som markerer brukseksempler. Kildehenvisningene er ikke markert i teksten i figur 3. Men det er etter måten enkelt å la programmet sette inn tagger for disse. Jeg vil arbeide videre med dette utover høsten.

Fritznors ordbok på Internett

Etter oppfordring ble ordboken lagt ut på Internett i juni 2004 (se <http://www.dokpro.uio.no> el. <http://www.dok.hf.uio.no/perl/search/search.cgi?appid=86&tabid=1275>) Bak nettversjonen finnes ganske enkel database med noen få tabeller:

1. En tabell med en post for hver artikkel. Her finnes teksten i en tagget og en ren tekst versjon. Den taggete versjonen brukes til fremvisning på nettet, mens den rene teksten danner grunnlaget for fullt fritekstsøk i hele ordboksteksten.
2. En tabell med oppslagsordene og ordklassene.
3. En egen ordklassetabell.
4. Koblingstabeller mellom 1, 2 og 3.

Grensesnittet lages automatisk på grunnlag av informasjon i en metadatabasetabell som er felles for alle våre databaser. Brukerne kan velge mellom ulike søkeskjemaer og resultatoppsett. Resultatet kan sorteres etter verdiene i de ulike kolonnene ved å klikke på kolonnetittelen. Jeg viser for øvrig til den lille hjelpeteksten det vises til i skjermbildet.

Standardsøkeskjemaet gir brukeren mulighet til å søke etter oppslagsord og ordklasse. En kan også velge et annet søkeskjema der det er mulig å foreta fritekstsøk i hele ordboksteksten. Det er ingen begrensninger på antall treff. Dersom en lar alle søkefelt være tomme og trykker på søkknappen eller enter-tasten, får en frem hele ordboken. Om en deretter sorterer på oppslagsord kan en altså bla frem og tilbake i hele ordboken. Den sorteringen av hele ordboken tar litt tid, dvs. i underkant av to minutter.

Grensesnittet skal forbedres med bedre håndtering av spesialtegn, lagrede søk og muligheten for å laste ned hele resultatsett som tab-separerte tabeller til brukerens maskin.

Fritzners ordbok og andre leksikalske resurser for nordisk middelalderspråk

Et av målene med arbeidet med å analysere og tagge ordboksteksten, er å lage en såkalt Metaordbok for norrønt. Metaordboken er et verktøy vi har utviklet for Norsk Ordbok 2014. Den inneholder kraftige verktøy for å systematisere leksikalsk grunnlagsmateriale.

Fullt utbygd og korrekt anvendt vil Metaordboken avspeile prosessen med å skape ordboksartikler helt fra ubehandlede samlinger, via omstruktureringer og systematiseringer til en eksplitt beskrivelse av hva de rå samlingsdataene betyr. I prinsippet skal man ha tilgang til alt materiale som er brukt i definisjonsprosessen, også det som ikke er kommet med i ordboksartikkelen. På den måten vil man komme et steg videre med å gjøre en ordbok vitenskapelig etterprøvbart.

Det eneste store norrøne ordboksprosjektet i dag er *Ordbog over det norrøne prosasprog* i København. Jeg har ikke til hensikt å forsøke å legge om dette prosjektets redigeringsprosess. Men det hadde vært en besnærende tanke om Ordbogens redaksjon anvendte det samme verktøyet som Norsk Ordbok. Tanken er bare å benytte Metaordbokens mekanikk til å lage en felles database for de ressursene vi disponerer. En slik Metaordbok vil bestå av en “artikkel” for hvert leksem som inneholder pekere til bruksbelegg og til definisjonsartikler i Fritzners ordbok. I tillegg vil metaordboksartiklene være indeksert etter de tre ulikt normaliserte rettskrivningene som finnes for norrønt dvs. rettskrivningene til Fritzner, Gammelnorsk ordboksverk og *Ordbog over det norrøne prosasprog*. I forlengelsen av dette prosjektet kunne man også tenke seg krysskoblinger til Söderwalls ordbok ved Språkdata i Göteborg.