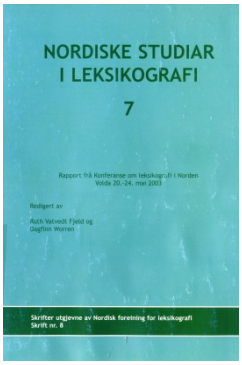
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Appendiks om digitaliseringen av Norsk Riksmålsordbok

Forfatter: Knut Lunde

Kilde: Nordiske Studiar i Leksikografi 7, 2005, s. 175-176
Rapport frå Konferanse om leksikografi i Norden, Volda 20.-24. maj 2003

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.

Knut Lunde

Appendiks om digitaliseringen av *Norsk Riksmålsordbok*

Forarbeid

Norsk Riksmålsordbok (NRO) består av fire hovedbind fra 1957, og to supplementsbind fra 1995. En revisjon måtte starte med å fremskaffe verket i elektronisk form, og ved prosjektstart fantes dette ikke for hoveddelen, mens tilleggsbindene fantes i et eldre satsformat.

Etter noen prøveskanninger, som viste en tilfredstillende kvalitet, ble NRO bind I-IV i 1999 optisk lest og OCR-behandlet på Aktietrykkeriet i Trondheim. Resultatet ble levert tidlig i 2000, og deretter maskinelt bearbeidet av Per Richard Osmundsen i ProText Systems (nå Type-it AS) med programsystemet Synergy. Systematiske feil som lot seg oppdage med logiske regler, ble rettet, og siden konvertert til en svært enkel XML-kodet tekst bestående kun av tagger for fet og kursiv, som var den eneste typografien som OCR-lesingen kunne spore. Teksten ble så konvertert til RTF, som var mest hensiktsmessig for korrektur mot bokutgaven. Etter denne korrekturen ble RTF-filene returnert til ProText, som laget en enkel XML-struktur.

Parallelt med dette forarbeidet med bind I-IV ble bind V og VI ekstrahert fra filer som kom fra en Norsk Data-maskin hos tidligere AlfaBeta i Halden. Disse ble konvertert til samme enkle XML-struktur.

I 2001 og inn i 2002 ble jobben med å gi teksten en leksikografisk merking påbegynt. Utfordringen var å gjenkjenne de forskjellige leksikografiske elementene i artiklene, og å legge til XML-koder i samsvar med den på forhånd definerte modellen for artiklene.

For å gjenkjenne de forskjellige leksikografiske elementene i en artikkel brukte ProText blant annet:

- typografi (fet, kursiv og rett)
- forskjellig tegnsetting som semikolon, punktum, komma, skråstrek, dobbel skråstrek m.m. (jf. forordet i NRO)
- faste tekster (som er lagt inn i tabeller, og som konverteringsprogrammet tester imot):
    - språkforkortelser (som ghty., gno., senlat. m.m.)
    - forfatter- og verknavn m.m.
    - navn på bibelske skrifter
    - avisnavn
    - fag- og stilforkortelser (som folkemed., kunsth., dial., folk., etc.)
- andre faste tekster som:
    - Jvf., se
    - Hertil
- forskjellige endelser, som f.eks:
    - -t, -te, -lte, -de

Disse kriteriene, av og til i kombinasjon, ble brukt for å merke artiklene leksikografisk. Langt på vei lyktes det å skille ut oppslagsord, bøyninger, uttale og etymologi, betydninger med nummerering og nyanser samt definisjoner og eksempler med henvisninger, sitat og kilder.

Totalt er det 14 moduler i ProTexts konverteringsprogram som bearbeider teksten og legger til XML-kode. Artiklene fra bindene V og VI ble til slutt lagt inn på alfabetisk riktig plass i filstrukturen til bind I- IV.

Strukturredigering

Når ProText antok at man ikke kunne komme stort lenger med maskinelle metoder uten å legge betydelig mer ressurser i denne fasen, måtte artiklene over i en manuell redigeringsfase. Her stod man overfor to hovedoppgaver. Den ene var å smelte artiklene fra bind I-IV sammen med materialet fra bind V og VI. Den andre var å rydde opp og rette strukturen etter den maskinelle kodingen. I denne fasen ble artiklene redigert samlet i filer i en XML-editor. Arbeidet ble avsluttet medio oktober 2003, og deretter ble det foretatt nok noen innstramninger og opprydninger i artikkelstrukturen.

Et meget enkelt system ble benyttet til dette arbeidet. Adept Epic ble valgt som XML-editor og XSL og MS Internet Explorer 6.0 ble brukt til artikkelvisning.

Fremtidig redigeringsverktøy

Norsk Riksmålsordbok har henvendt seg til Enhet for digital dokumentasjon ved UiO om bruk av redigeringsverktøyet som er utviklet for Norsk Ordbok. En avtale vil kunne gi NRO et moderne, skreddersydd redigeringsverktøy, men vil også kunne gi Norsk Ordbok og forskere ved UiO elektronisk tilgang til NRO. NRO er fra høsten 2004 publisert i Kunnskapsforlagets nettbaserte tjeneste www.ordnett.no.