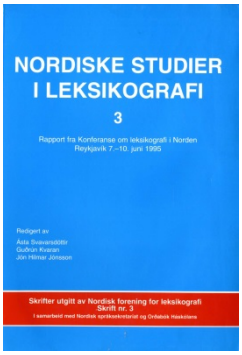
NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	Ei ordliste trekt ut av tre ordbøker/ordlister
Forfatter:	Marit Hovdenak
Kilde:	Nordiske Studier i Leksikografi 3, 1995, s. 205-211 Rapport fra Konferanse om leksikografi i Norden, Reykjavík 7.-10. juni 1995
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.

Marit Hovdenak

Ei ordliste trekt ut av tre ordbøker/ordlister

The article deals with the compilation af a spelling dictionary, *Nynorskordlista*, based on three electronically stored sources. Two of the sources are printed dictionaries. The types of data extracted from each source are described. The data were combined into the compilation format of *Nynorskordlista*. Some advantages of the method are discussed.

1 Innleiing

Dette innlegget er ein presentasjon av arbeidet med ei rettskrivingsordliste — *Nynorskordlista* — som er redigert på dataskjerm. Ordlista — som like gjerne kunne kallast rettskrivings*ordbok* — er laga på grunnlag av tre elektronisk lagra ordlister. I framstillinga ligg hovudvekta på dei meir tekniske sidene ved redigeringa.

2 Tre hovudkjelder

Nynorskordlista byggjer på desse tre hovudkjeldene:

- *Nynorskordboka*
  Definisjons- og rettskrivingsordbok, 2. utg. 1993
  90 000 oppslagsord, 755 sider

- *Bokmålsordlista*
  Rettskrivningsordliste for skole, hjem og kontor, 1992
  56 000 oppslagsord, 397 sider

- *IBM-morfologien*
  Elektronisk ordliste med bøyingsmønster
  111 000 oppslagsord

Ut frå desse hovudkjeldene — og frå mange andre aktuelle kjelder — har eg saman med Anne Torunn Engø redigert *Nynorskordlista*. Ordlista har fått om lag same omfanget som *Bokmålsordlista*. Utgjevinga skal vente til rettskrivingsvedtaka i Norsk språkråd i fireårsperioden 1992-95 er endeleg godkjende våren 1996.

Nynorskordlista er eit samarbeidsprosjekt mellom Avdeling for leksikografi ved Universitetet i Oslo, Norsk språkråd og Det Norske Samlaget. Initiativet til å utarbeide større rettskrivingsordlister for bokmål og nynorsk kom frå Språkrådet. Desse ordlistene er å rekne som offisielle rettskrivingsordlister på linje med *Retskrivningsordbogen* i Danmark og *Svenska Akademiens ordlista*.

Nordiske studier i leksikografi III, 205–211

3 Ordbokstype

Nynorskordlista blir ei større ordliste over det sentrale allmenne ordtilfanget i nynorsk. Talet på oppslagsord, bortimot 60 000, blir om lag dobbelt så stort som det vanlege skuleordlister har. Ordtilfanget byggjer særleg på *Nynorskordboka*, men opplegget for ordlista følgjer langt på veg *Bokmålsordlista*, og vi har bygd på den i stor grad.

I tillegg til grunnord og avleiingar, som har sin sjølvsagde plass i ordlista, er det med ein del samansetningar, ikkje minst for å vise samansetningsfuge. Ofte har vi vist samansetjingsmåten slik, utan å ta med døme:

religion [. . .] i sms.: **religions-**

Det er lagt vekt på å få med nye, aktuelle ord frå ymse område. Den offisielle norma i nynorsk har vore nokså atterhalden med å ta inn særleg dansk-tyske lånord, til dels også nyare engelske ord. Ordlista vår bryt ikkje med denne tradisjonen, men har opna litt for desse kategoriane av ord.

Mange namn er med som oppslagsord, særleg stadnamn, norske og utanlandske, i mange tilfelle med innbyggjarnemningar til. Ein del vanlege historiske og mytologiske namn er med, og til ein viss grad namn på statsinstitusjonar og internasjonale organisasjonar. Vanlege forkortingar og symbol er tekne med som oppslagsord.

Om kvart oppslagsord er det først og fremst teke med formelle opplysningar: rettskriving og bøying. Tydinga til orda er gjerne teken med for å skilje homografar frå kvarandre og ved meir ukjende ord. Uttaleopplysningar er med særleg ved uvanlege eller lite kjende ord.

4 Utval frå kjeldene og utforming av redigeringsformat

Dei tre hovudkjeldene er lagra elektronisk i nokså ulike format. I startfasen av arbeidet vart det valt ut aktuelle typar data frå kvar kjelde, og dei vart sette saman til eit redigeringsformat.

Utgangspunktet var *Nynorskordboka*, som vart skriven datamaskinelt i "Hjulstad-formatet", med namn etter Håvard Hjulstad, som laga edb-opplegget for ordboka. I Hjulstad-formatet har kvart oppslagsord ein **post** delt inn i **felt**, som her i dei to oppslagsorda *eple* og *eplekart* i *Nynorskordboka*:

```
=
NN001 eple
NN001a N2
TR006
..OPP #>$Ceple@ n1
..ETY (norr $Bepli@; smh med $Bapal@)
..DEF $C1@ frukt av epletre
..UTR $Bplukke, hauste, sylte, mose e- !@
..UTR $Be-t fell ikkje langt frå stamma@
..FOR han, ho liknar far sin el. mor si !
..UTR $Bmåtte bite i det sure e-t@
..FOR måtte finne seg i noko ein ikkje liker
```

..DEF $C2@ potet, jordeple
..UTR $Bsetje, hyppe, ta opp e-@
..DEF $C3@ rundvoren del el. ting
..UTR $Badamse- !@
..UTR $Baugee-@
=
=
NN001 eplekart
TR006
..OPP $C~kart@
=

På trykk blir dette:

eple n1 (norr *epli*; smh med *apal*) **1** frukt av epletre *plukke, hauste, sylte, mose e- / e-t fell ikkje langt frå stamma* han, ho liknar far sin el. mor si / *måtte bite i det sure e-t* måtte finne seg i noko ein ikkje liker **2** potet, jordeple *setje, hyppe, ta opp e-* **3** rundvoren del el. ting *adamse- / augee-* . . . **~kart**

Først i kvar post er det linjer som ikkje er med i den trykte versjonen (topplinjene, NN. . . , og TR006), følgde av felt for oppslagsord med bøying, etymologi, definisjonar og uttrykk. Til *Nynorskordlista* plukka vi ut oppslagsfeltet (. . . OPP), dessutan topplinjene, som inneheld fullt utskrivne oppslagsformer også for samansetningar, og dessutan bøyingskodar etter eit meir utbygd system enn trykkversjonen av boka.

Bokmålsordlista er redigert i eit tekstbehandlingsformat utan feltinndeling, berre med ei blank linje mellom avsnitt, som til dømes i oppslagsordet *eple* med samansetningar:

$Ceple@ $B-et@; $B-er@, $B-a@/$B-ene@ $C~kake~kart~slang@

Dette er på trykk:

eple *-et; -er, -a/-ene* **~kake**
~kart ~slang

IBM-morfologien er ei elektronisk ordsamling med full bøying for kvart ord. Nynorskversjonen, som vi brukte, har om lag 111 000 ord og berre hovudformene i offisiell rettskriving, ikkje sideformer (klammeformer). Ordsamlinga er utarbeidd av IBM Norge og kjøpt inn av Dokumentasjonsprosjektet. Frå IBM-morfologien vart det trekt ut ei oppslagsform for kvart ord, dessutan delar av bøyingsmønstra i høveleg rekkjefølgje.

Data frå dei tre kjeldene vart så samsorterte alfabetisk slik at det for kvart oppslagsord i råmanuset var samla dei utvalde opplysningane i éin artikkel. Formatet på råmanuset er ein omarbeidd variant av Hjulstad-formatet, og det skal nokså enkelt kunne konverterast til den no så allmenne SGML-standarden.

To samsorterte ordartiklar i råmanuset med data frå alle tre kjeldene kan sjå slik ut:

IBMopp: eple
IBMbøy: -et; -e, -a

```
IBMtyp: NN (802) 17526
NN001 eple
NN001a N2
..OPP #>$Ceple@ n1
NOBløp: 15475
nold
BOL $Ceple@ $B-et@; $B-er@, $B-a@/$B-ene@ $C~kake~kart~slang@
BOLløp: 9247
IBMopp: eplekart
IBMbøy: -en; -ar, -ane
IBMtyp: MN (700) 17538
NN001 eplekart
..OPP $C~kart@
NOBløp: 15487
nold
```

Først i kvar av desse to postane kjem tre linjer frå IBM-morfologien, med oppslagsform, bøying og bøyingstype i IBM-typologien. Oppslagsorda fekk løpenummer i kvar kjelde. Deretter kjem linjene frå *Nynorskordboka* (NN... osv.). Tydingsfeltet *nold* er sett inn i redigeringsformatet. Til slutt kjem eventuelt linjer frå *Bokmålsordlista* (BOL).

5 Redigering

Redigeringa gjekk teknisk sett ut på å merkje av dei orda som skulle vere med som oppslagsord i ordlista, å kontrollere og eventuelt endre oppslagsform og bøying, og å setje til uttale- og tydingsopplysningar der det var ønskjeleg. Dei to øvste linjene, altså *IBMopp*: og *IBMbøy*:, kunne ofte brukast direkte som dei var, til oppslagsform og bøying. Dei andre linjene fungerer som datagrunnlag, kontroll og referanse. Vi brukte redigeringsprogrammet **emacs** (GNU-emacs), som er kraftig og har mange funksjonar.

Dei same to artiklane som før, ferdig redigerte, ser slik ut:

```
#> #o
IBMopp: eple
nolb -et; -, -a
IBMbøy: -et; -e, -a
IBMtyp: NN (802) 17526
NN001 eple
NN001a N2
..OPP #>$Ceple@ n1
NOBløp: 15475
nold
BOL: $Ceple@ $B-et@; $B-er@, $B-a@/$B-ene@ $C~kake~kart~slang@
BOLløp: 9247
#o
nolo~kart
```

IBMopp: eplekart
nolb
IBMbøy: -en; -ar, -ane
IBMtyp: MN (700) 17538
NN001 eplekart
..OPP $C~kart@
NOBløp: 15487
nold

Artiklar som skulle med i ordlista, markerte vi med visse teiknsekvensar over postane: *#> #o* står for ordartikkel i nytt avsnitt, *#o* står for påhengd artikkel utan nytt avsnitt.

I det første dømet ovanfor kunne oppslagsforma *eple* frå IBM-materialet brukast som ho stod. Bøyinga er ikkje utforma heilt som den som er vald for *Nynorskordlista*, og difor sette vi inn eit nytt bøyingsfelt (*nolb*) over IBM-bøyinga. Vi fastsette at det alltid er det øvste bøyingsfeltet som skal gjelde. Eit eventuelt uttalefelt kan setjast inn før bøyinga. Tydingsfeltet (*nold*) står tomt her.

I den påhengde artikkelen (*eplekart*) sette vi berre etterleddet med tilde framfor. Til det førte vi inn eit nytt oppslagsfelt (*nolo*) og eit tomt bøyingsfelt, fordi samansetningar i regelen ikkje har bøying oppført. Redigeringa gjekk i det heile ut på å leggje til, ikkje stryke.

Oppslagsord som ikkje stod i nokon av hovudkjeldene, vart redigerte i tilsvarande postar som dei over.

Det vart laga utskriftsprogram for to slags korrekturar. Dei fleste felta i redigeringsformatet vart strokne i korrekturutskriftene. Den endelege utskrifta blir om lag slik (temmeleg lik *Bokmålsordlista*):

eple *-et; -, -a* ~**kart** ~**sider**

6 Manglar ved samsorteringa

På to punkt fungerte ikkje samsorteringsprogrammet: ved homografar og ved teikn utanom bokstavane *a-z*. Homografar stod i klyngjer i råmanuset, programmet skilde dei ikkje frå kvarandre. Homografseparereringa måtte difor gjerast enkeltvis på skjermen i redigeringa. Ved ikkje-alfabetiske teikn som bindestrek og aksentar og til dels ved *æ, ø* og *å* slo også samsorteringa ofte feil. Kjeldene var ikkje alfabetiserte etter heilt dei same reglane. Det hadde truleg vore teknisk mogleg å lage eit program som greidde opp med både homografar og ulik alfabetisering, men det ville krevje så mykje meir programmeringsarbeid at det spørst om det ville vere verdt det.

7 Informatikarinnsats

Denne arbeidsmåten der så mykje blir gjort elektronisk, krev ein stor programmeringsinnsats, særleg i startfasen. Til samanstøypinga av råmanuset og utarbeidinga av redigeringsformatet fekk vi edb-hjelp frå Dokumentasjonsprosjektet. Det viste seg å bli ein heil prosess å komme fram til eit godt format og ikkje minst ein økonomisk arbeidsmåte, ein prosess med ein del prøving og feiling. Informatikaren frå Dokumentasjonsprosjektet var lydhør og hjelpsam,

og fann alltid tid til å løyse problem. På vegen fann vi fram til ein del tidssparande finessar, som er avgjerande for arbeidstempoet. Det er avgjort ein fordel å vere litt fascinert av kva datateknikken kan utrette. Arbeidet var òg lagt godt til rette med datautstyr på avdelinga, og det har vore god driftsstøtte på HF-fakultetet.

Eg nemner dette fordi vi slapp mange frustrasjonar som sikkert ville komme dersom utstyr og edb-ekspertise ikkje var så lett tilgjengeleg.

8 Vurdering av opplegg og arbeidsmåte

Ein stor fordel med å ha tre hovudkjelder elektronisk lagra og samsorterte er at ein slepp stadig slå opp i desse bøkene/kjeldene og jamføre. Ein er dessutan nokså godt sikra mot at ein overser aktuelle oppslagsord i desse kjeldene. På den andre sida er det ein fare for at ein kan bli for bunden av ordutvalet i akkurat desse kjeldene.

Det blir opplagt mindre skrivearbeid med denne arbeidsmåten enn med meir tradisjonelle måtar i og med at mykje av teksta alt er inne. Ein unngår ein del trivielle skrivefeil. Redigeringa går mykje ut på å flytte skrivemarkøren i posisjon, å kopiere, slette o.l., det vil ofte seie å bruke kontrolltastar, eventuelt mus, i staden for bokstavtastane. Det verkar som desse operasjonane er fysisk meir slitande enn vanleg tekstskriving.

Manuset til ordlista krev stor plass på dataskjermen. Med ein vanleg skjerm er det berre plass til to-tre oppslagsord på skjermbiletet. Det ville vere ein fordel med ein større skjerm. I redigeringa var det nødvendig å kike i dei trykte bøkene for å vurdere kva ord som skulle veljast ut. Det var best å ta eit overblikk over trykksider for å vurdere korleis det tilsvarande utsnittet i ordlista skulle bli. Å redigere direkte på skjerm er lite oversiktleg.

Noko av meininga med å lagre ordlister og andre tekster elektronisk er at dei lett skal kunne rettast opp og forbetrast. Arbeidet med seinare utgåver av ordlista skulle såleis vere enklare enn før i tida. Feil som er funne i kjeldene mens arbeidet gjekk føre seg, er noterte nokså systematisk — for å betre kvaliteten på lagra data.

Sidan *Nynorskordlista* finst i elektronisk form, skulle ho liggje godt til rette for gjenbruk. Ordartiklane er delte inn i stort sett homogene felt, utan skriftkodar, og skulle kunne nyttast direkte om ein til dømes vil lage ei rein rettskrivingsordliste utan uttale- og tydingsfelt. Andre aktuelle gjenbruksprodukt er stavekontrollprogram og orddelingsprogram eller -ordbøker.

Ved sida av utskriftsprogram i to versjonar vart det laga enkle program for å plukke ut lister med typar av oppslagsord, til dømes ord merkte med fagområde eller stadnamn og andre namn. Redaktørane har lært å ta utskrifter og lage ymse lister sjølve, det er ein fordel ikkje minst i det avsluttande kontroll- og korrekturarbeidet.

Litteratur

Ordbøker

Bokmålsordlista 1992. Oslo: Universitetsforlaget.

Nynorskordboka 1993. 2. utg. Oslo: Det Norske Samlaget.

IBM-morfologien. Elektronisk ordliste utarbeidd av IBM Norge.

Annan litteratur

Fjeld, Ruth Vatvedt (red.) 1992: *Nordiske studier i leksikografi. Rapport fra Konferanse om leksikografi i Norden*. Oslo: NFL.

Garde, Anna/Pia Jarvad (red.) 1994: *Nordiske studier i leksikografi II. Rapport fra Konference om Leksikografi i Norden*. København: LEDA.

Jörgen Pind 1989: Computers, Typesetting, and Lexicography. I: Jörgen Pind/Eiríkur Rögnvaldsson (utg.): *Papers from the Seventh Scandinavian Conference of Computational Linguistics*. Reykjavík: Institute of Lexicography/Institute of Lingustics.