Håvard Hjulstad
Norsk termbank /
Norsk leksikografisk institutt

DATABEHANDLING AV NORSK HANDORDBOK

Det har vore arbeidd redaksjonelt med handordbøkene, ei for bokmål og ei for nynorsk, sidan 1974 som eit samarbeidsprosjekt mellom Norsk språkråd og Norsk leksikografisk institutt. Ordbøkene skal dekkje allmennspråket, og innanfor ei ramme på om lag 900—1000 sider skal dei gi ortografi, bøying, uttale, etymologi, synonym/definisjonar og døme på bruk.

Heilt frå starten var det meininga å kople datamaskina inn i arbeidet. Frå førsten var det til å velje ut ordtilfang og særleg til å samordne ordtilfanget i dei to utgåvene ein tenkte seg at datamaskina kunne yte ein innsats. Frå før var det lagra ved Prosjekt for datamaskinell språkbehandling i Bergen eit stort tilfang på bokmål og nynorsk. Ein tenkte seg at samsorterte lister av dette tilfanget kunne vere startgrunnlaget for redigeringa. Dette vart ikkje følgt opp, og redigeringa har heile tida vore reint manuell.

I 1979 kom ein i gang med dataføringa av nynorskutgåva. Det var lagt opp etter eit system som utvikla seg frå det som vart brukt til behandlinga av Norsk landbruksordbok. Innskrivingsformatet ligg nær opp til trykk; ein "simulerte" setjeri i innkodinga. Rett nok var kodesystemet meir fininndelt enn det som kjem fram i trykk. Det var etter måten enkelt å produsere ei trykt bok ut frå dette formatet på data, men vitskapleg bruk av data var heller tungvint.

I byrjinga av 1981 fann vi det føremålstenleg å leggje om dette opplegget. I datamaskinbaserte ordboksprosjekt snakkar ein no stadig meir om felt, og også Norsk handordbok har fått sine felt merkte i data.

Litt feltfilosofi

Noko er felles frå ordbok til ordbok.

gutunge|n
gut ... -unge som unge
gutunge gutungen gutungar gutungane

Det står det same, men på ulike vis. I maskinversjonen må ein få fram at det er det same, men på trykk må ein få lov til å uttrykkje det ulikt. Også andre opplysningar kan uttrykkjast ulikt om dei tyder det same.

I den grad det er råd å gi ei opplysning i ein kode eller ei anna "reinsa" form, bør denne forma finnast i maskinversjonen. Også i det andre av dei tre døma ovanfor må det vere råd å leite maskinelt på "gutunge".

Databehandling av norsk handordbok
Håvard Hjulstad
Proceedings of NODALIDA 1981, pages 171-174

Somtid er det mest føremålstenleg å analysere den redigerte "trykkdelen" og dra ut slike opplysningar. Andre gonger kan ein gå andre vegen og generere det som skal trykkjast ut frå dei koda opplysningane.

I Norsk handordbok gjer ein det på første måten. Redaktørane får "drive med sitt" utan å leggje om arbeidsforma. Men det er jo å vone at framtidige prosjekt kan endre på arbeidsforma på dette punktet. Da blir det truleg enda meir å tene på det også.

Ein post er no samansett av ein prosjektuavhengig del og ein eller fleire prosjektavhengige delar. Kvar av desse delane inneheld i regelen fleire felt. Når ein er ute etter ei opplysning, leiter ein berre i det feltet der ein kan vente å finne denne opplysninga.

Ned på jorda

Nokre postar i Norsk handordbok (nynorskutgåva) ser no slik ut:

```
NN001    kabrette
NN001a   F2
TR006
..OPP    £>$Ckabret>te@ f2
..ETY    (truleg sm o s norr tilnamn $Bkakbretta@ kanskje
         smh med $BI kabbe@ 'klump')
..DEF    ostevelling av innkokt myse
=
NN001    kabel
NN001a   M1t
TR006
..OPP    £>$Ckabel@ $B-en, -blar@
..ETY    (gj lty $Bkabel@ og fr $Bc$3able@ frå lat. $Bcapulum@ 'reip'
         av $Bcapere@ 'ta')
..DEF    $C1@ kraftig trosse av tau el. ståltråd
..DEF    $C2@ elektrisk leidning med kraftig isolering
=
NN001    kabelfarty
NN011    kabelfartøy
TR006
..OPP    $C~farty@ $C<<~fartøy>>@
..DEF    farty som legg sjøkabel
=
NN001    kabelfjernsyn
TR006
..OPP    $C~fjernsyn@
..DEF    fjernsynsoverføring til mottakaren delvis med
         hjelp av *kabel (2)
=
NN001    kabelgatt
NN002    kabelgat
TR006
..OPP    $C~gat(t)@
..DEF    rom i farty for tauverk
=
```

(Merknader til utskrifta:
Feltkodar: NN tyder "nynorsk", 001 tyder "første hovudform", 001a tyder "grammatisk kode til første hovudform", 011 tyder "første sideform", TR006 tyder "trykkdel prosjekt nr 006", "trykkdelkodane" som tek til med .. skulle vere sjølvforklarande. Trykkodar: £> tyder "nytt avsnitt i trykk", $B ..@ tyder "kursiv skrift", $C .. @ tyder "halvfeit skrift". << og >> står for hakeparentes.)

Det er fleire opplysningar som kan få plass "på toppen" enn dette. Det skal fyllast ut med bøyingskodar. Kanskje noko formalisert etymologi, og i alle høve formalisert semantikk bør ein freiste å "pine" ut av data. Formalisert syntaks skulle også gjerne vore der, men ordboka har få eller ingen opplysningar om dette.

Feltinndelinga i den prosjektavhengige delen er "lausare". Her må ein tillate variasjonar frå prosjekt til prosjekt.

Innskrivinga

Innskrivinga har vi freista å gjere så enkel som råd med å bruke eit enkelt feltkodesystem. Dei første postane i dømet ovanfor vart skrivne inn slik:

```
O kabret>te
G f2
E (truleg sm o s norr tilnamn $Bkakbretta@ kanskje
 smh med $BI kabbe@ 'klump')
D ostevelling av innkokt myse
=
O kabel
B -en, -blar
E (gj lty $Bkabel@ og fr $Bc$3able@ frå lat. $Bcapulum@ 'reip'
 av $Bcapere@ 'ta')
D1 kraftig trosse av tau el. ståltråd
D2 elektrisk leidning med kraftig isolering
-
```

Ein sparer mykje skrifttypekoding med å gjere det slik. Som ein ser er felta O, G, B slått saman i ..OPP-feltet i utskrifta på førre sida. Opplysningane er her fullstendig analyserte til "toppdelen", og den finare feltinndelinga er ikkje lenger nødvendig.

Litt framtid

Når eit prosjekt er ferdig, tek ein med seg "toppdelen" til neste prosjekt. "Trykkdelen" får leve sitt eige liv, men via ein identifikasjonsnøkkel som skal få plass i begge delane, kan ein seinare kople dei saman.

Det kan godt vere at ein flik av framtida vil sjå slik ut: