

Ivar Fønnes:

EDB I ORDBOKSPRODUKSJON:

TILRETTELEGGING AV NORSK LANDBRUKSORDBOK FOR TRYKKING.

1. Innledning.

Norsk Landbruksordbok er en definisjonsordbok bestående av ca. 17000 oppslag som innleder definerte ordartikler, og ca. 8000 oppslag med referanse til definerte ord, altså totalt ca. 25000 oppslag. Hele materialet er registrert i maskinleselig form og lagt opp som en databank. På dette grunnlag produseres det så magnetbånd for fotosetting og trykking av ordboken.

I de fleste ordartikler er det angitt synonymer på en del andre språk. Disse synonymene trekkes ut i egne lister (én for hvert språk) med de norske oppslagsord som referanse, alfabetiseres og legges opp på magnetbånd for fotosetting.

Ordboken vil bli trykket i to bind, hvorav bind 1 (hoveddelen) inneholder de norske oppslag med definisjoner og referanser, og bind 2 (registerbindet) inneholder alfabetiserte synonymlister på 7 språk (samisk, svensk, dansk, finsk, islandsk, engelsk og tysk) med henvisning til de norske oppslagsord i hoveddelen.

Databehandling har spilt en sentral rolle i produksjonsarbeidet. Det har vært et verdifullt hjelpemiddel i det avsluttende redigeringsarbeid, og har muliggjort fullstendig automatisert produksjon av registerbindet. Dessuten har vi kunnet produsere hele materialet på fullt ferdig såkalt "drivetape" som kan gå direkte til fotosetting.

I Norge har det ikke tidligere vært produsert definisjonsordbøker ved hjelp av slike metoder, og vi har ikke hatt kjennskap til lignende prosjekter som kunne danne forbilde for vårt opplegg. Materialet i Norsk Landbruksordbok er forøvrig nokså spesielt og i vår sammenheng komplisert. Det forekommer en rekke trykkvarianter i hyppig veksling, stort tegnrepertoar, formler osv. (jfr. vedlegg).

Materialet er også av betydelig størrelse, anslagsvis omkring 8 mill. tegn. Det nærmer seg altså størrelsen på f.eks. Brown Corpus. I tillegg kommer så synonymlistene som bygges opp maskinelt. Disse vil i trykk utgjøre nesten like mange sider som hovedmaterialet.

Denne framstilling tar sikte på å redegjøre for hvordan EDB kan være til nytte i produksjonen av en slik ordbok, og trekke fram noen av de mest sentrale problemer som knytter seg til EDB-arbeidet i en slik sammenheng.

2. Materialet.

Hver ordartikkel (se eksempel i vedlegg) består av et oppslagsord med eventuelle bøyingsformer og varianter for nynorsk og bokmål, eventuelt etymologi, definisjon av ordet m.v., angivelse av fag-

område, signatur for den faglig ansvarlige, og synonymer på norsk og inntil 6 andre språk. I tillegg kan det forekomme spesifikasjoner av oppslagsordet (f.eks. med adjektiv) og påhengte oppslagsord med egne definisjoner og synonymer.

Alle disse opplysningene følger tett etter hverandre, men ulike trykkvarianter gjør det likevel forholdsvis lett for leserne å holde dem fra hverandre: Oppslagsord i halvfet, bøyningsformer i kursiv, fagmerking med versaler, signaturer med kursiverte versaler, utenlandske synonymer i hakeparenteser, spesifikasjoner i sperrede kapitler, nynorsk- og bokmålsvarianter med hevet henholdsvis n og b osv.

Materialet er delt inn i ca. 40 fagområder, og før trykkeprosjektet startet forelå det meste i stensilerte hefter (maskinskrevet) med ett hefte for hvert fagområde. Oppslagsordene var ordnet alfabetisk innen hvert hefte.

Ved trykking av ordboken skulle hele materialet ordnes i ett alfabet. Dette måtte nødvendigvis bli et nokså omfattende sorteringsarbeid. Dessuten måtte det foretas en del redaksjonelle endringer som følge av den nye sorteringen. Eksempelvis måtte oppslagsord som fantes i flere hefter, bygges sammen i én og samme ordartikkel.

Redaksjonen ønsket også å bygge inn en del opplysninger som var kommet til etter at heftene ble skrevet, og foreta en del justeringer. F.eks. tok man sikte på å endre genusmarkeringen for alle substantiver.

Endelig skulle de utenlandske synonymer trekkes ut med referanser og alfabetiseres innen hvert språk.

3. Formål og fordeler ved bruk av EDB.

Blant flere mulige opplegg for trykking av ordboken valgte man fotosetting via EDB. Fotosats gir samme kvalitet som blytsats og praktisk talt fritt valg med hensyn til tegn- og type-repertoar.

Bruk av EDB i produksjonen ble valgt av flere grunner:

a) Lavere kostnader enn ved manuelt opplegg. Dette var delvis basert på billig maskintid ved Universitetet i Oslo. Men til tross for at prosjektet har måttet kjøpe en del maskintid utenfor universitetet, har vurderingen vist seg å være riktig.

b) Biprodukt i form av en databank, tilgjengelig for forskningsformål og en fordel ved eventuelle senere utgaver av boken.

En slik databank vil også kunne få betydning utover det som var planlagt. Ved EF-kommisjonen i Brussel arbeides det med en databank for landbruksterminologi, og det er interesse for en kopi av vårt materiale. Fra vår side er det ønskelig å bygge inn synonymer på flere språk, f.eks. fransk. Utveksling vil trolig finne sted etter at trykkeprosjektet er avsluttet.

c) Reduksjon av redaksjonens avsluttende arbeid i betydelig grad, og muligheter for systematisk kontroll av materialet.

Alfabetisering av materialet samt utplukking og alfabetisering av synonymer er meget betydelige arbeidsoppgaver som nå er blitt utført automatisk.

Finalalfabetisk liste over oppslagsord muliggjør systematisk kontroll av morfemmarkeringer, betoningsaksenter og skrivemåte. Lister over alle ulike fagmerkinger og signaturer muliggjør kontroll av disse osv.

Totalt sett vurderte man det slik at EDB ville gi betydelige gevinster i form av lavere kostnader, systematiske kontrollmuligheter og en databank. Dette har vist seg å være en korrekt vurdering. I tillegg bør nevnes den erfaring prosjektet har gitt i arbeidet med å produsere trykklare data på et slikt materiale, ikke minst fordi materialet både er stort i omfang og komplisert.

4. Sentrale problemområder i EDB-opplegget.

4.1. Utgangspunktet.

Materialet til Norsk Landbruksordbok var samlet inn gjennom mange år og lagt til rette for trykking på tradisjonell måte. Spørsmålet om EDB hadde aldri vært inne i bildet, og opplegget var naturligvis da heller ikke på noen måte tilpasset slike metoder. På den annen side var det en selvsagt forutsetning at bruk av EDB ikke skulle påvirke ordbokens utseende - fotosetting via EDB skulle gi samme resultat som manuell blytsats.

Dette noe ugunstige utgangspunkt, sett fra EDB-siden, representerte imidlertid ikke noe særlig betydelig problem. Markeringer av forskjellige typer av opplysninger i ordartiklene var gjort med henblikk på at det skulle være lett å finne fram for leserne. Det var lagt opp til å anvende ulike skrifttyper som halvfet og kursiv i tillegg til vanlig skrift, og dessuten kapiteler og versaler samt parenteser, hakeparenteser, skråstreker osv. Dette opplegget passet egentlig veldig bra for databehandling. Ved å legge inn de markeringer som var nødvendige for å skille mellom ulike trykkvarianter, fikk vi samtidig inn de opplysninger maskinen trengte for å produsere listeprodukter for redaksjonen, og også hoveddelen av de markeringer som ble ansett nødvendige for å utnytte materialet maskinelt i en databank. Noen få tilleggsmarkeringer ble lagt inn med henblikk på maskinell utnyttelse av materialet, og disse ble naturligvis fjernet under opplegget av et trykklart magnetbånd.

Et større problem var egentlig det forhold at man ikke hadde full oversikt over hva som kunne forekomme av tegn og kombinasjoner i materialet. For eksempel: Vi visste at det forekom både greske bokstaver og matematiske formler, men kunne en gresk bokstav forekomme som eksponent i en formel (altså som hevet tegn i trykk)? I tillegg kom at det foregikk redaksjonsarbeid parallelt med registreringen slik at det kunne oppstå nye varianter underveis.

Det var altså ikke til å unngå at det dukket opp nye problemer underveis i prosjektet, og løsning av disse måtte da innpasses i opplegget på best mulig måte. For programmeringen er dette alltid en ugunstig situasjon. Likevel er det klart at det ikke dukket opp ting som representerte noe alvorlig problem i forhold til det opplegg som var valgt.

4.2. Datarepresentasjon.

Vi skulle altså representere et materiale med et tegnrepertoar som langt overstiger det man vanligvis har tilgjengelig i en datamaskin og dessuten markeringer for ulike skrifttyper og andre trykkvarianter. (Ser man bort fra forskjellen mellom vanlig skrift, kursiv og halvfet er repertoaret på ca. 220 tegn. Tar man med kursiv- og halvfet-variantene stiger tallet til ca. 420). Dette måtte markeres med en eller annen form for funksjonskoder. Vi valgte å bruke en funksjonsmarkering (tegnet %) etterfulgt av et tall eller en bokstav som angir hvilken funksjon det dreier seg om, og la funksjonen gjelde inntil den blir opphevet (med tegnet \$). Flere funksjoner kan forekomme inne i hverandre, og et \$-tegn opphever alltid sist angitte funksjon.

På denne måte kan vi representere nær sagt alle tenkelige varianter av tegn og tegnkombinasjoner. Vi bruker slike funksjoner for å markere kursiv og halvfet, hevede og senkede tegn, gammelnorsk, samisk og gresk alfabet, aksenter, brøkstrek, kvadratrot osv. På den annen side koster det noe ekstra arbeid å skrive slike funksjoner under dataregistreringen, og det vil også kreve ekstra påpasselighet å holde rede på hvor man til enhver tid befinner seg, særlig hvis man har flere funksjoner inne i hverandre. For å lette registreringsarbeidet ble det innført forenklede varianter av enkelte av de mest høyfrekvente funksjoner, f.eks. at * (asterisk) medfører at neste tegn skal være hevet (meget hyppig ved n og b for nynorsk og bokmål), og at & markerer at neste tegn er en aksent som skal over foregående bokstav.

Til tross for denne forenkling er det klart at registreringsarbeidet var forholdsvis krevende, og det var utvilsomt en betydelig fordel at registratoren kjente materialet og oppbyggingen av ordartiklene gjennom arbeid i redaksjonen.

4.3. Presentasjon av data for korrektur.

Med alle de funksjonskoder som måtte brukes, til dels meget hyppig, er det klart at materialet ikke var særlig lett å lese. Det var derfor en viktig oppgave å finne fram til en bedre presentasjonsform med henblikk på korrekturarbeidet.

Her hadde vi et velegnet grunnlag i de stensilerte hefter hvor materialet forelå maskinskrevet. I heftene var trykkvariantene markert med ulike typer understrekinger, understreking med + - tegn punktum, likhetstegn, asterisk, vanlig understreking osv. Dette systemet kunne vi anvende ved å skrive ut selve materialet på annenhver linje, og la maskinen konvertere en del av funksjonskodene til understrekinger på de mellomliggende linjer.

Dette systemet viste seg meget vellykket. Materialet kunne nå skrives ut i en etter forholdene oversiktlig form, og dessuten i en form som redaksjonen var vant til fra før gjennom de stensilerte heftene.

4.4. Alfabetisering av materialet etter oppslagsord.

Alfabetisk sortering av materialet kunne ikke gjennomføres direkte på oppslagsordene slik de forelå i materialet. For det første inneholder ordene en del informasjon som måtte bort (punkturering for morfemgrenser, betoningsaksenter og annen ikke-alfabetisk informasjon som f.eks. vanlige aksenter og tall).

For det andre måtte en del informasjon konverteres. Bokstaver som ä, ö, ü (svensk eller tysk) og andre ikke-norske bokstaver, måtte konverteres til norske ekvivalenter som ga dem riktig plass i alfabetet. Romertall (betydningsnummer) foran oppslagsord måtte flyttes bak ordet, det samme gjaldt bindestrek som stod foran ordet. Oppslag på mer enn ett ord måtte markeres spesielt osv.

Det måtte altså etableres egne sorteringsfelter for de enkelte oppslagsord. Regelverket for etablering av sorteringsfelt ble relativt komplisert. Dette skyldes at det forekommer diverse ikke-alfabetiske tegn i oppslagsordene som ofte representerer signifikant informasjon for sorteringsrekkefølgen. En ekstra komplikasjon var at noen få ord måtte behandles særskilt på tvers av regelverket.

4.5. Klargjøring av materialet for fotosetting.

Klargjøringen for fotosetting omfattet i hovedsak følgende arbeidsoppgaver:

- Fjerning av informasjon som var lagt inn med henblikk på databehandling.
- Ny linjeinndeling.
- Bearbeidelse av materialet etter trykkeriets spesifikasjoner og opplegg av korrekte koder på magnetbånd.

4.5.1. Linjeinndeling.

I og med at innholdet i hver ordartikkel skrives fortløpende, må nødvendigvis linjeinndelingen i den trykte versjon bli forskjellig fra den man brukte ved dataregistreringen. Dette ville ikke representere noe betydelig problem i et vanlig tekstmateriale, men med alle de spesielle markeringer som forekommer i denne ordboken, viste det seg forholdsvis komplisert å utarbeide kriterier og programmer for linjedeling. Hvilke tegn skulle følge ordet foran, hvilke skulle følge neste ord (dvs. ned på neste linje) hvilke skulle behandles som selvstendige enheter, hvilke tegn kunne ikke innlede en linje, hvilke kunne ikke avslutte en linje osv. Så kommer spørsmålet om deling av formler, at fagmerking og signaturer ikke kan deles, og dessuten vanlig orddeling som kompliseres ved at en rekke ord inneholder ikke-alfabetiske tegn, deler av ordet kan stå i parentes osv.

A) Oppbygging av en linje.

Rent teknisk har man følgende utgangspunkt: Tegnene har varierende bredde som angis i såkalte relative enheter (RE). F.eks. har bokstaven i en bredde på 4RE og bokstaven m hele 13RE i vanlig skrift. I kursiv er de tilsvarende tall 4,5 RE og 12,5RE (tallene gjelder for skriften Times). Totalbredden på spalten (= linjelengden) er også angitt i relative enheter, i vårt tilfelle 486RE. Programmet løper altså gjennom materialet tegn for tegn,

summerer opp breddeverdier og ser hvor mye det blir plass til innenfor de 486RE.

Man kan imidlertid ikke bare avslutte linjen etter siste hele ord før grensen er nådd. Alle linjer unntatt den siste i hver ordartikkel skal ha rett høyremarg, og jokeren i arbeidet med å få dette til er ordmellomrommene (space). Disse kan variere betydelig i bredde (i vårt tilfelle fra 4 til 14RE), og dette gir forholdsvis gode muligheter til å få delt linjen på et naturlig sted og likevel oppnå rett høyremarg. Programmet må altså normalt finne fram til ett eller to steder hvor det er naturlig å dele linjen (mellom ord el.lign.) og undersøke om breddeintervallet på ordmellomrom tillater deling der. Hvis ikke må det foretas orddeling.

B) Orddeling.

Orddeling er et velkjent problem innen automatisert tekstbehandling. I vårt tilfelle ble forholdet ytterligere komplisert ved at materialet inneholder ord på 6 andre språk som f.eks. tillater en del konsonantkombinasjoner som ikke finnes i norsk. Likeledes måtte vi ta hensyn til ikke-alfabetiske tegn samt parenteser inne i ord.

I vår sammenheng syntes det lite hensiktsmessig å legge mye arbeid i å komme fram til et perfekt orddelingsprogram som kunne gi korrekt orddeling i alle tilfeller. Vi vurderte det som mindre arbeidskrevende å bruke et enkelt program som ga riktig deling i de fleste tilfeller, og så rette opp manuelt de delinger som var feil. Det program vi har brukt er utviklet for norsk tekst og bygger på enkle prinsipper.

Vårt opplegg går da i korthet ut på at ord som må deles blir avkledd all ikke-alfabetisk informasjon, parenteser holdes utenfor, og så overlates ordet til orddelingsprogrammet. Alle orddelinger som foretas skrives ut i en egen oversiktlig liste. Redaksjonen leser korrektur på denne listen og anmerker feil. Feilene kan rettes av oss som vanlig korrektur eller av trykkeriet under settingen. Vi har kommet til at det siste vil være det enkleste.

Resultatet av kjøring på ca. 20% av materialet antyder følgende statistikk for orddelinger:

Totalmaterialet er på ca. 100.000 trykte linjer.
Orddeling foretas i ca. 5.000 linjer, dvs. ca. 5% av linjene.
Feil orddeling forekommer i ca. 800 linjer, dvs. ca. 16% av orddelingene er feil, eller: Feil orddeling forekommer i ca. 0,8% av linjene.

Vi anser dette for å være fullt tilfredsstillende - det vil ikke koste mye arbeid å gjennomføre slik kontroll og korrektur.

De typer av feil som forekommer, er først og fremst plasseringen av s i sammensatte ord, f.eks.

årsvekst blir til år-svekst
og feil av typen

endring blir til en-dring.

Noen feil skyldes også spesielle konsonantkombinasjoner i andre språk, f.eks. tysk.

4.5.2. Tilrettelegging av magnetbånd for fotosetting ("drivetape").

Etter at linjeinndelingen var klar skulle så materialet organiseres etter trykkeriets spesifikasjoner og legges opp på magnetbånd.

Tegnene som den aktuelle fotosetteren har til rådighet er organisert på såkalte fonter, med inntil 112 tegn på hver. De vanligste tegn er lagt på 3 parallelle fonter, en for vanlig skrift, en for kursiv og en for halvfet. Dessuten har vi til rådighet 2 fonter med mer spesielle tegn, hvorav den ene er lagt opp med spesielt henblikk på dette prosjektet.

Angivelsen av hvilken font et tegn skal hentes fra skjer ved funksjonskoder, noe i likhet med det prinsipp vi har brukt i databanken. Når riktig font er angitt, følger så koden for det eller de tegn som skal hentes fra denne fonten, deretter ny fontangivelse osv. Systemet har så pass mange felles trekk med vårt funksjonskodeopplegg at det gikk forholdsvis greitt å tilrettelegge materialet på denne måten.

Magnetbåndet ("drivetapen") skrives i såkalt TTS-kode som er en 6-bits kode og dermed bare har plass til 64 ulike tegn. Dette gjør det nødvendig med nokså mye skift (f.eks. for store og små bokstaver) og funksjonskoder. Et problem i denne forbindelse har vært å kontrollere innholdet av dette magnetbåndet i forbindelse med uttesting av programmet. Kontroll på grunnlag av "dump" fra båndet er både spesielt tidkrevende og forholdsvis utsatt for feil. Den virkelige kontroll har derfor måtte vente til det aktuelle prøvemateriale var kjørt gjennom trykkeriets fotosetter (i Stockholm).

Det er utdrag fra en slik trykkprøve som er gjengitt i vedlegget.

