Ole Norling-Christensen

EDB OG DEN DANSKE ORDBOG

Elektronisk databehandling (edb) kan bruges, og bliver brugt, på alle stadier af den leksikografiske proces: Til studiet af ordbogens kilder (tekstkorpora, excerptsamlinger, andre ordbøger); til indskrivning og redigering i mere eller mindre strukturerede formater; ved revisioner; og ved selve den typografiske fremstilling af den tekst, der skal trykkes. Desuden er en ordbog ikke længere altid en *bog*. Ordbøger på disketter og på CD-ROM, til brug på computere, kan købes både i Danmark og i Nederland. I det følgende vil anvendelsen af edb ved forberedelserne af Den Danske Ordbog, og specielt dens tekstkorpus, blive belyst.

Den Danske Ordbog

Arbejdet med denne nye seksbinds ordbog begyndte den 1. september 1991. Ordbogen udarbejdes af *Det Danske Sprog- og Litteraturselskab*, som i sin tid også stod for den 28 binds *Ordbog over det Danske Sprog* (*ODS*, 1918-56). Redaktionen ledes Ebba Hjorth, Iver Kjær og mig. Projektet finansieres halvt af staten (Kulturministeriet) og halvt af Carlsbergfondet. Opgaven er at fremstille den bedst mulige ordbog indenfor et budget på 30 millioner kroner (ca. 9 millioner gylden) i løbet af otte år: 18 måneders forberedelser, 5 års manuskriptredigering, og 18 måneder til korrekturlæsning og projektafslutning.

Ordbogens vigtigste kilder er: 1) et tekstkorpus på 40 millioner løbende ord; 2) visse centrale eksisterende ordbøger; 3) materiale som Dansk Sprognævn har indsamlet siden det blev oprettet i 1955, dels "seddelsamlingen" dvs. ca. en million excerpter, dels et register over alle sammensatte ord i

seddelsamlingen, ordnet efter sammensætningernes sidste led; samt endelig, naturligvis, 4) redaktørernes sproglige kompetence. *Korpuset* er en afbalanceret samling af prøver på alle slags tekster, både tale- og skriftsprog, både privat og offentligt sprog, og både almensprog og alment fagsprog, fra perioden 1983-92. Blandt ordbøgerne er de vigtigste Retskrivningsordbogen (RO, 1986), som er den officielle danske norm for stavning og bøjning, samt de to største tosprogsordbøger: Dansk-engelsk (V&B[3], 1990) og Dansk-fransk (B&H[4], 1991). De foreligger alle tre som elektroniske data.

På få undtagelser nær skal *Den Danske Ordbog* være deskriptiv, altså beskrive sproget som det *er*, ikke som det ifølge én eller anden norm burde være. Undtagelserne er stavning og bøjning; her skal ordbogen følge den officielle norm, og det er især her, Retskrivningsordbogen kommer ind i billedet. Førsteudgaverne af de store tosprogsordbøger byggede begge på Ordbog over det Danske Sprog 1-28 (ODS, 1918-56), som dækker det danske (skrift)sprog fra 1700 til ca. 1915/55. De benytter i det væsentlige de samme betydningsopdelinger som ODS; men i løbet af tre (V&B[3]) henholdsvis fire (B&H[4]) revisioner har de naturligvis moderniseret ordforrådet og bestanden af kollokationer og idiomer. De er de fornemste repræsentanter for en levende dansk leksikografisk tradition, som Den Danske Ordbog skal fortsætte.

Computere bruges på alle stadier af arbejdet. Under den indeværende forberedelsesperiode (september 1991 - marts 1993) udvælges tekstprøver til ordbogens korpus, og de indlægges i computersystemet ved scanning, ved indskrivning, eller ved konvertering af alle slags tekstbehandlings- og fotosatsformater. Mulighederne for automatisk (syntaktisk og) morfologisk opmærkning (tagging) af korpusteksterne undersøges; men endelig beslutning herom er ikke truffet. Dog opdeles skriftsprogsteksterne i afsnit, som igen opdeles i sætninger, og som et biprodukt af denne opdelingsproces er proprier og forkortelser blevet mærket som sådanne. Til hver tekstprøve knyttes, mere eller mindre automatisk, oplysninger om forfatter(e), teksttype osv. Samtidig opbygger vi ordlister, som gradvis udbygges til "artikelskeletter", der i struktureret form rummer så mange oplysninger som det har været muligt at udtrække fra de nævnte trykte ordbøger. Før og i løbet af redaktionsperioden (april 1993 - marts 1998) fremsøges korpusbelæg på de enkelte opslagsord.

De analyseres, struktureres og knyttes til "skeletterne". Disse halvfabrikata bliver det så redaktørernes opgave at omdanne til færdige ordbogsartikler. Til redaktionsarbejdet bruges det SGML-baserede system GestorLEX, som er udviklet af det danske softwarehus TEXTware A/S i nært samarbejde med mig og mine kolleger i DANLEX-gruppen. Systemet opfylder praktisk taget alle de krav, som opstilles i (DANLEX 1987: 239-251). Til søgning i korpuset og til analyse af søgeresultaterne bruges specialprogrammel der kan fungere sammen med GestorLEX. Det er under udvikling og vil være færdigt i november 1992.

Maskinlæsbar / maskinbrugbar

Mange tekster kan i vore dage fås i såkaldt maskinlæsbar form, dvs. som datafiler fra sætterier eller fra forfatternes tekstbehandlingsudstyr. Der er dog også tekster, som må gøres maskinlæsbare ved scanning af en trykt version eller ved indtastning, førend de kan indlægges i korpus. Uanset hvordan teksterne er blevet inddateret, må de efterfølgende omdannes til ét ensartet og velbeskrevet format, der egner sig til de søgninger og analyser, som korpuset skal bruges til. At definere et sådant format er ingen trivel sag. Ikke blot skal der defineres et alfabet (character set / code page), der må også træffes en lang række beslutninger om, *hvilke træk* ved teksten man ønsker at repræsentere i den maskinlæsbare version. Træk, som er irrelevante for korpusarbejdet bør udelades, mens de relevante træk skal repræsenteres i dataene på ensartet og entydig måde.

Skal der for eksempel være særlige koder for avisens lugt eller kvaliteten af dens papir? - næppe. Papirets farve? - det kunne have en særlig betydning. Bogstavernes størrelse? - forskelle i størrelse signalerer ofte forskellige slags tekst; men hvilke forskelle, der er tale om, varierer fra tekst til tekst. En indlysende konsekvens af spørgsmål som disse er, at kodningen skal være *generisk*, dvs. afspejle *hvilken slags* tekst, der er tale om; den skal ikke afspejle, hvordan forfatteren eller forlaget valgte at præsentere de forskellige slags tekst. Altså: måske en kode for "erhvervssider" (i avis), men ikke for

"lyserødt papir"; måske en kode for "overskrift", men ikke for "store fede bogstaver". The Text Encoding Initiative (TEI 1990; TEI 1991) har for en lang række tekstarter defineret en sand overflod af generiske koder, som tager højde for de fleste af de træk, som humaniske forskere kan få brug for. I praksis bliver det dog nødvendigt at udvælge et meget begrænset antal træk og kun lade disse være repræsenteret i korpus.

Repræsentation af talt sprog

Ifølge planen skal Den Danske Ordbog *dække* det skrevne sprog, men *inddrage* det talte. Årsagen til denne ubalance er indlysende: talesproget i alle dets variationer er langt vanskeligere at afgrænse og at repræsentere i et korpus. Vi har ikke tilstrækkelige resurser (hverken tid eller penge) til selv at indsamle og udskrive ret meget talesprog. I stedet har vi prøvet at finde alle slags allerede eksisterende udskrifter, og efterhånden er der dukket ret meget materiale op. Tit har vi også kunnet få de oprindelige båndoptagelser eller lydspor, så vi har kunnet tjekke udskrifternes kvalitet og om nødvendigt rette dem til. Udskrifterne er lavet af forskellige mennesker og til forskellige formål: forskning inden for sprog, psykologi, sociologi, eller simpelhen dokumentation af, hvad der er sagt i radio og fjernsyn, eller i Folketinget og Københavns Borgerrepræsentation. Udskrifterne bruger derfor mange forskellige konventioner for notering af især træk som pauser, sætningsbrud, latter og uforståelige passager, samt for udskriverens sidebemærkninger. Også her er det vigtigt med klare regler for, *hvilke træk* ved teksten man ønsker at repræsentere i den maskinlæsbare version af talesproget.

Et standard format for korpusenheder

Den internationale standard SGML (ISO 8879, 1986) for generisk beskrivelse

af tekststrukturer og for opmærkning af teksterne i overensstemmelse med beskrivelsen, bruges af Den Danske Ordbog til at beskrive og strukturere ikke blot ordbogartiklerne, men også korpuset. En glimrende introduktion til SGML for humanister findes i (TEI 1990: 9-32).

Til brug for korpuset har vi i SGML-formalismen defineret en såkaldt dokumenttype ved navn *Korpusenhed*. Den rummer en formular til registrering af de nødvendige (ikke-sproglige) oplysninger *om* teksten, samt muligheden for på entydig måde at opmærke de (sproglige) træk ved selve teksten, som vi har besluttet skal være repræsenteret. Hver Korpusenhed består af en Header, der rummer oplysninger om tekstens art og hvor den stammer fra, samt selve Teksten.

Headeren

Ved opstillingen af headeren og afgørelsen af, hvilke oplysningstyper der skulle indgå, fandt vi megen inspiration i (Atkins 1991). Headeren til hver enkelt tekstprøve består af to hovedafsnit, nemlig oplysninger om kilden (kildeangivelse) og oplysninger om selve tekstprøven (tekstbeskrivelse). *Kildeangivelsen* består af en entydig identifikation (tekstgruppe + tekstnr), angivelse af eventuelle restriktioner mht. brug af teksten (specielt ved private og/eller fortrolige tekster), oplysninger om Sprogbruger, dvs. den eller dem som har frembragt teksten (forfattere hhv. talende), samt om tekstens titel, forlægger, datering, og lokalisering (fx sidetal). Der er én *Sprogbruger*-beskrivelse for hver person som har andel i teksten, og specielt ved talesprog er det oftest mere end én. Her beskrives personens rolle (fx interwiever eller intervievet), køn, uddannelse, beskæftigelse, fødeår og -sted, samt sprogvariant (rigssprog eller regionalsprog). Under overskriften *Tekstbeskrivelse* gøres der rede for, om sproget er alment eller fagligt (sprogtype), om det er skrevet eller talt (udtryksmedium), om det er produceret af de få for de mange (reception) eller af nogle af de mange ikke-professionelle (produktion), aldersrelationen mellem afsender og modtager (voksen-voksen, voksen-ung, voksen-barn, ung-voksen, .. , barn-barn), medie (fx bog, avis, tv), genre, emne,

samt antallet af løbende ord i prøven (omfang).

Headerens fuldstændige struktur og indhold kan beskrives som nedenfor vist. Et spørgsmålstegn efter en oplysningstype betyder, at den er fakultativ, dvs. den medtages kun, når det er relevant og dens indhold er kendt. Plusset (+) efter Sprogbruger betyder at der kan være flere af dem.

Header
Kildeangivelse
Tekstgruppe*entydig identifikation af en gruppe (beslægtede) tekster*
- **Tekstnr***løbenummer inden for tekstgruppen*
- **Restriktion?**
  - **Restriktion_a** *teksten skal anonymiseres: "ja"/"nej"*

Restriktion_b *teksten må ikke bruges til andre formål end ordbogen*
  - **Udløbsår** *for restriktion b*
- **Sprogbruger+**
  - **Rolle?** *især ved flere sprogbrugere, fx "lærer", "elev"*
  - **Identifikation?***entydig tretegnskode, hvortil replikker kan referere*
  - **Efternavn?** *hvis det kendes*
  - **Fornavn?** *hvis det kendes*
  - ***Køn** *"m"/"k"/"u[kendt]"*
  - **Uddannelse?** *hvis det kendes*
  - **Erhverv?** *hvis det kendes*
  - ***Fødselsår** *et heltal mellem 1880 og 1990*
    - **Sikker?** *"?", hvis året ikke kendes med sikkerhed*
  - **Fødested?** *hvis det kendes*
  - ***Sprogvariant**
- **Teksttitel?** *hvis der er en titel*
- **Værktitel?** *Navn på fx antologi, avis, ugeblad, o.l., hvis relevant*
- **Forlag?** *normalt kun ved bøger og radio-/tv-stationer*
- **Datering**
  - **Dag?** *hvis den kendes*
  - **Måned?** *hvis den kendes*
  - **År** *et heltal mellem 1983 og 1992*
  - **Sikker?** *"?", hvis året ikke kendes med sikkerhed*
- **Lokalisering?** *fx sektion/side/spalte i avis; (bind og) side ved*

bøger

Tekstbeskrivelse

***Sprogtype**	*"almensprog"/"(alment)fagsprog"*
***Udtryksmedium**	*"skrift","tale", eller et par mellemformer*
***Synsvinkel**	*"reception"/"produktion"*
***Aldersrelation?**	*"barn-barn"/"barn-ung"/"barn-voksen"/../"voksen-voksen"*
***Medium**	*ét fra en liste på 12, bl.a. bog, avis, radio, film*
***Genre?**	*én fra en liste på 124 delvis medieafhængige genrer, fx roman, brev*
***Emne?**	*ét fra en liste af 64, fx biologi, litteratur, politik, fysik*
Omfang	*antal løbende ord i tekstprøven*

De oplysningstyper, som er mærket med asterisk (*) ovenfor, er standardiserede deskriptorer, som har en særlig funktion ved søgninger og analyser. For hver af disse deskriptorer er der defineret en værdimængde, dvs. liste over tilladte værdier. Forskellige teksttyper, og tilsvarende subkorpora, kan defineres ved hjælp af kombinationer af deskriptorerne, fx "kvinder født før 1940, som taler til børn" eller "avistekster der handler om politik". Desuden udnyttes deskriptorerne ved undersøgelse af sproglige træks fordeling over teksttyper.

Teksten

Struktureringen af selve teksten afhænger af, om den indeholder skriftsprog eller (nedskrevet) talesprog. Skriftsprog opdeles i afsnit, som igen opdeles i sætninger. Sætninger er normalt umærket tekst. Indlejret heri kan dog forekomme særligt markerede dele, som betegnes Fremhæv, Note, Propr og Abbr. *Fremhæv* dækker over enhver slags fremhævelser i originalteksten, det være sig understregning, fed, kursiv, spatiering, eller større eller på anden måde afvigende skrift; *Note* omfatter såvel fod- som slutnoter; de indsættes på det sted i teksten, hvor de hører til, altså normalt der, hvor notehenvisningen

står. *Propr* og *Abbr* er henholdsvis egennavne og forkortelse.

Talesprog inddeles normalt ikke i afsnit og sætninger. Derimod vil teksten oftest være inddelt i replikker. Som nævnt, er de fleste talesprogstekster samtaler eller interviews, hvor flere personer er involveret, og headeren rummer derfor to eller flere Sprogbrugere. Hver af disse indeholder under *Identifikation* en karakteristisk tretegns kode, hvortil der kan refereres fra den enkelte replik. I øvrigt rummer en *Replik* først og fremmest umærket tekst, eventuelt iblandet angivelser som {tøven}, dvs. tøvelyde som 'øh', 'mmm', {pause}, {uf} som repræsenterer en uforståelig passage, {latter}, samt dele, der er mærket Kommentar eller Usikker. *Kommentar* er udskriverens "regibemærkninger" som ikke er en del af talen; *Usikker* er en passage, som udskriveren ikke var sikker på, men har prøvet at rekonstruere.

Datamatiske værktøjer

Så mange som muligt af header-oplysninger indføjes fuld- eller halv-automatisk, og det samme gælder til en vis grad opmærkningen af selve teksten (i en del tilfælde bruges her dog tekstbehandlingsmakroer). Dette indebærer, at der for hver gruppe af tekster af samme oprindelse eller type skrives et særligt konverteringsprogram, som ikke blot ændrer et givent tekst-behandlingsformat til korpusets standard, men i visse tilfælde også genkender og behandler den pågældende forfatters særlige måder at markere de træk, vi er interesserede i.

De vigtigste programmer der bruges til bearbejdelse af ordbogens maskin-læsbare kilder, er: den kontekstfrie chartparser DIPA, skrevet i program-meringssproget C af Peter Molbæk Hansen, lektor i datalingvistik ved Kø-benhavns Universitet; det generelle tekstkonverteringssystem DICONV, som jeg selv har skrevet i programmeringssproget Turbo Pascal (de objektorien-terede versoner 5.5 og 6.0); samt Paradox Engine fra Borland International Inc. Desuden bruges et almindeligt tekstbehandlingsprogram (WordPerfect 5.1) og databasesystemet Paradox. I forskellige kombinationer anvendes disse værktøjer ikke blot til forarbejdning af korpusenheder, men også til

fremstilling af ordlister og til den gradvise udvidelse af disse til "artikelskeletter".

DIPA (*Di*ctionary *Par*ser) blev oprindelig udviklet til brug for mit arbejde med strukturanalyse og SGML-mærkning af af ordbogsdata (Norling-Christensen 1992). Ved Den Danske Ordbog bruges den især til bearbejdelse af de ordbøger, som udgør en del af vores kildemateriale, men også fx til at analysere og opmærke headeroplysninger.

DICONV (*Di*ctionary *Conv*erter) blev oprindelig udviklet til at konvertere typografiske filer til tekstbehandlingsfiler og omvendt, og til at lave alle slags automatiske ændringer og rettelser i ordbogsdataene. Sidenhen er den blevet udvidet med faciliteter til behandling af SGML's træstrukturer og har fundet anvendelse som præ- og postprocessor til DIPA. Ved Den Danske Ordbog bruges den især til at bearbejde tekstprøver til korpuset.

Paradox Engine er et programmeringsværktøj, som kan bruges sammen med såvel Pascal som C. Det giver programmøren adgang til alle slags basale databasefunktioner, såsom oprettelse af tabeller med et eller flere indexer, søgninger, skrivning og læsning af databaseposter, osv. Filformaterne er de samme som i databasesystemet Paradox og visse andre af Borland's produkter. Det indebærer, at de samme samme data kan bruges og bearbejdes både af Paradox og af éns egne programmer.

Til den datamatstøttede fremstilling af headere til korpusenhederne bruger vi en skræddersyet Paradox-applikation, som på én skærm viser data fra tre midlertidige tabeller: en sprogbruger-tabel, en tabel med resten af headeroplysningerne, samt endelig selve korpusteksten. Denne skærmformular bruges til indtastning af de headeroplysninger, som ikke har kunnet indlægges automatisk. De tre tabeller fremstilles ved hjælp af Turbo Pascal og Paradox Engine. For hver gruppe tekster indlægges så mange headeroplysninger som muligt, før tabellerne gøres færdige af en redaktør. Oftest vil de automatisk indlagte oplysninger være sådanne som er fælles for en lang række tekster. De

kan fx alle komme fra samme kilde eller dække samme fagområde, genre, eller sprogtype. I nogle tilfælde kan dog også oplysninger, der er specifikke for den enkelte tekst, overføres. Det gælder især tekster, som allerede har været beskrevet og klassificeret af andre, såsom tekster fra forskellige mindre korpusprojekter og fra et dagblads interne tekstsøgningssystem, hvis data vi har modtaget på magnetbånd. Når en gruppe headere er gjort færdige af redaktionen, tømmes de midlertidige tabeller igen, idet deres indhold eksporteres til SGML-mærkede filer, der siden kan importeres til korpus-systemet. Headeroplysningerne (men ikke teksten) overføres desuden til en varig database, som løbende giver os overblik over, hvilke tekster vi har, og hvorledes de fordeler sig på de forskellige typer.

Perspektiver

Formålet med det korpus, som her er beskrevet, er udarbejdelsen af en ordbog over nutidigt dansk almensprog. Følgelig er egentligt fagsprog, det sprog som produceres af fagfolk for fagfolk, ikke repræsenteret. Korpus dækker perioden 1983-92, og nyere tekster vil ikke blive tilføjet, i hvert fald ikke af Den Danske Ordbogs medarbejderstab. Med vilje har vi indskrænket os til at forsyne teksterne med de headeroplysninger og anden opmærkning, som vi anser for nyttige til ordbogsarbejdet. Forskere på andre områder kunne sikkert have ønsket sig flere headeroplysninger og andre slags opmærkninger. Alligevel forventer vi, at der bliver stor efterspørgsel efter Den Danske Ordbogs korpus fra andre der beskæftiger sig med udforskningen af dansk. Det er nemlig langt det største danske tekstkorpus, og det eneste der rummer så mange forskellige teksttyper. Og de oplysninger, som korpusenhederne er forsynet med, forligger i en konsekvent og veldokumenteret form. Bortset fra eventuelle konkurrerende projekter, er det derfor meningen at andre forskere skal have adgang til materialet, og det er vort håb, at der vil blive fundet en måde hvorpå de metoder, værktøjer og principper, vi har udviklet, kan udnyttes og videreudvikles i fremtidigt korpusarbejde.

Litteratur

Atkins 1991: Atkins, Sue, Jeremy Clear & Nicholas Ostler: Corpus
Design Criteria. 8th November 1991. To appear in *Literary and Linguistic Computing*.
B&H[4] 1991: Blinkenberg, Andreas & Poul Høybye: Dictionnaire
Danois-Français/Dansk-fransk Ordbog. 4. udg. ved Jens Rasmussen & al. Vol. 1-2. København 1991.
DANLEX 1987: The DANLEX Group (Ebba Hjorth, Jane R.
Jacobsen, Bodil Nistrup Madsen, Ole Norling-Christensen, Hanne Ruus): Descriptive Tools for Electronic Processing of Dictionary Data. Studies in Computational Lexicography. Lexicographica Series Maior 20. Tübingen 1987.
ISO 8879 1986: International Organization for Standardization:
Information processing - Standard General Markup Language (SGML). [Genève]: ISO, 1986.
Norling-Christensen 1992: Struktureret redigering af Ordbøger. In Ruth
Vatvedt Fjeld (ed.): *Nordiske Studier i Leksikografi. Rapport fra Konferanse on Leksikografi i Norden 28.-31. mai 1991.* Oslo 1992: 447-454.
ODS 1918-56: Ordbog over det Danske Sprog. Udgivet af Det Danske Sprog- og Litteraturselskab. Vol 1-28. København 1918-56.
RO 1986: Dansk Sprognævn: Retskrivningsordbogen. København 1986.
TEI 1990: Burnard, Lou, & C.M. Sperberg-McQueen (ed.s): Guidelines
For the Encoding and Interchange of Machine-Readable Texts. Document Number: TEI P1. Text Encoding Initiative, Chicago, Oxford. ACH, ACL, ALLC. Draft: Version 1.1, October 1990.
TEI 1991: Burnard, Lou, & C.M. Sperberg-McQueen: Living with the
Guidelines. TEI EDW21: An Introduction to TEI Tagging. The first TEI European workshop 1-2 July 1991. Oxford University Computing Service 24 Jun 1991.
V&B[3] 1990: Vinterberg, Hermann & C.A. Bodelsen: Dansk-engelsk
Ordbog. 3. udg. ved Viggo Hjørnager Pedersen. København 1990.