Hanne Ruus
Institut for nordisk filologi
Københavns Universitet

ORDBØGER FOR FREMTIDEN

I midten af det tyvende århundrede havde et ordbogsarbejde typisk følgende faser:

1. Opstilling af retningslinjer for udvælgelse af materiale.
2. Indsamling af materiale.
3. Fastlæggelse af redaktionsprincipper. Udarbejdelse af redaktionsregler.
4. Redaktion af ordbogsartiklerne.
5. Trykning og distribution af ordbogen.

Alle dele af processen blev udført manuelt. En stor del af fase 2, 4 og 5 bestod af af- eller renskrivningsarbejde og korrekturlæsning.

I begyndelsen af det enogtyvende århundrede vil automatiseringen have ført til, at et ordbogsarbejde typisk har følgende faser:

1. Detaljspecifikation af ordbogens indhold og opstilling.
2. Materiale til ordbogen fremtages automatisk fra andre maskinlæsbare ordbøger og fra ord- og tekstbanker.
3. (Semi-)automatisk redaktion af ordbogsartikler.
4. Trykning eller distribution ad andre databærende medier som informationsnetværker og teletekst.

I denne version af arbejdsgangen vil fastlæggelse af redaktionsprincipper og redaktionsregler indgå i arbejdet med detaljspecifikationen af ordbogen.

Fase 2, fremskaffelse af materiale, vil kunne foregå automatisk, idet udvælgelseskriterierne vil fremgå af detaljspecifikationen.

Ordbøger for fremtiden
Hanne Ruus
Proceedings of NODALIDA 1981, pages 149-153

Fase 3 vil i det ideelle tilfælde bestå i en kontrollerende gennemlæsning af ordbogsartiklerne, som er fremstillet automatisk ved hjælp af generelle programmer, hvis specielle anvendelse er udledt af detaljspecifikationen.

Fase 4 foregår helt automatisk. Af trykningsmetoder kan man tænke på laserprintning, mikrofiche eller fotosætning. Ved distribution ad andre databærende medier vil det primært være et spørgsmål om at vejlede ordbogsbrugerne i, hvilke ordrer de skal skrive for at finde de nye oplysninger i eksisterende databaser. Sådanne vejledninger vil også kunne fremstilles automatisk og lægges ind som en del af systemerne.

Her i firserne er vi nået et godt stykke vej mod automatiseret fremstilling af ordbøger:

Maskinellet til datamatiseret ordbogsarbejde findes stort set i dag. Lagring af store datamængder er f.eks. ikke længere noget problem. Det mindst menneskevenlige på maskinelsiden er inddateringsmedierne, hvor f.eks. diakritiske tegn som oftest kræver specialbehandling.

På programmelsiden bør det ikke vare længe, før man kan købe en generel programpakke, der er egnet til udarbejdelse af ordbøger. Ingredienserne til en sådan pakke findes allerede i tekstbehandlingssystemer, i dokumentations- og informationssystemer og i forskelligt databaseprogrammel.

Man kan pege på flere andre faktorer, der befordrer automatiseringen af den almensproglige leksikografi:

1. Automatiseringen af trykkeprocessen.
Så godt som alle bøger sættes i dag automatisk og deres tekst gøres derfor maskinlæsbar i fremstillingsprocessen. Denne trykketekniske udvikling kan man få gavn af også i næsten afsluttet ordbogsarbejde f.eks. ved at få gjort teksten maskinlæsbar, når kladden til det endelige manuskript foreligger, som det sker ved Nye Ord i Dansk (Riber Petersen 1981).

2. Opbygningen af større samlinger af maskinlæsbare tekster. I Skandinavien kan man pege på Logoteket ved Språkdata i Göteborg, Norsk Tekstarkiv, DANwORDs samling af tekstprøver i København. Sådanne tekstsamlinger kan danne basis for ordbogsprojekter, som det f. eks. er planlagt i Göteborg med projektet Leksikalisk databas (Allén, Gavare, Ralph 1981).

3. Maskinlæsbare ordbøger.
Alle ordbøger, der udgives nu, eksisterer i maskinlæsbar form jf.1. Hvis man kan komme til rette med eventuelle copyrightproblemer, er der her et rigt materiale for andre ordbøger. Desuden findes der ordbøger, der er født maskinlæsbare. Det gælder alle ordbøger, der anvendes inden for projekter om automatisk analyse af naturlige sprog f.eks. kunstig

intelligens og maskinoversættelse. Sådanne ordbøger kan indeholde anselige datamængder f.eks. indeholder den tyske analyseordbog til SUSY ved SFB 100 i Saarbrücken omkring 300 000 leksikalske indgange (Maas 1980).

I den almensproglige leksikografi er frekvensordbøgerne nærmest automatiseringen. I udarbejdelsen af Nusvensk Frekvensordbok er der anvendt edb i alle faser af arbejdet, i DANw ORDsprojektet er tekstprøveindsamling og indkodning manuel, mens fremstilling af frekvenslister foregår automatisk.

Nærmest den fuldstændig automatiserede ordbogsfremstilling er man for tiden i fagsprogsleksikografien, hvor f.eks. TEAM, Siemens' fagsproglige database, kun behøver seks uger til at fremstille en ordbog. Til en flersproget ordbog over edb-termer brugte de således fem timers maskintid til at udtage materiale, en halv times maskintid til at redigere det og halvanden times maskintid til at fotosætte ordbogen (Sager and McNaught 1980).

Automatiseringen af ordbogsfremstillingen vil ændre det leksikografiske arbejde. F.eks. vil effektiv søgning kræve, at man begrænser antallet af synonyme betegnelser og at forskellige typer af oplysninger holdes klart adskilt.

Den mest indgribende ændring i ordbogsarbejdet vil nok være, at den detaljerede planlægning flyttes til den indledende fase. Denne planlægning kan blive meget omfattende, især når det drejer sig om leksikalske data, som tænkes lagret maskinelt til direkte benyttelse for oversættere, terminologer og undervisere. Det ser man af forarbejderne til DANTERM (Frandsen og Nistrup Madsen 1980, Nistrup Madsen 1981) og til British Linguistic Data Bank (McNaught 1981).

Hvis man ændrer sine planlægningsvaner og datamatiserer sit ordbogsarbejde, får man til gengæld en række muligheder, som savnes ved manuel ordbogsfremstilling jf. en række bidrag i denne publikation.

Datamater er uovertrufne til at sortere, søge og ændre - hundrede procent konsekvent. Man har altså mulighed for at undersøge klassificeringer og typer af oplysningers distribution i ordbogen på alle tænkelige måder med deraf følgende gevinst i form af konsekvens i behandlingen af samme type ord forskellige steder i alfabetet. Hvis man har planlagt sin ordbogsartikel passende struktureret, giver ændringsfaciliteterne mulighed for at ændre samtlige forekomster af en bestemt værdi på en bestemt plads i artiklerne lige til dagen før, man sender ordbogen til trykning.

Der bliver også mulighed for at foretage eksperimenter, som på længere sigt vil kunne ændre alle forestillinger om, hvordan ordbøger er indrettet og ser ud.

Man kan afprøve ordning af ordene efter deres betydningsmæssige sammenhæng f.eks. i et flerdimensionalt netværk som nævnt af Ralph (1979). Behov for forskning i denne retning kan udledes af Sager and McNaught (1980). Deres forespørgsler til potentielle brugere af en British Linguistic Data Bank viser, at oversættere gerne vil have overbegreb, synonymer og antonymer på kildesproget foruden målsprogsækvivalenten, når de slår et ord op.

I ordbøger, der som de fagsproglige databaser er tilgængelige via elektroniske medier, kan man anvende den ordning af ordene, der er mest meningsfuld for brugerne f.eks. en sammenkædning efter betydningsfællesskab, idet det overlades til programmerne at finde rundt i datamængden efter de kriterier, som kan udledes af brugerens spørgsmål.

Hvis teletekst ad åre bliver lige så udbredt, som telefonen er det i dag, er det nærliggende at gøre også almensproglige ordbøger tilgængelige via elektroniske medier. I så fald kan man også i disse ordbøger udnytte mulighederne for at koble ordene sammen efter betydningsfællesskaber i ordstoffet frem for efter formelle egenskaber som samme begyndelsesbogstav. Her vil man få hårdt brug for resultater af forsøg med at kæde almensprogets ord sammen i thesauruslignende strukturer.

Henvisninger.

Allén, Sture: Nusvensk Frekvensordbok, baserad på tidningstekst, 1-4, 1970-1980.

Allén, Sture, Gavare, Rolf and Ralph Bo 1981: Språkdata Research Report 1980.
Compiling nr. 10, feb. 1981.

Frandsen, Lene and Nistrup Madsen, Bodil 1980: The Setting up and Operation Of a Danish Terminological Data Bank (The DANTERM Project), i Human Translation - Machine Translation ed. by Suzanne Hanon and Viggo Hjørnager Pedersen = NOK 39, Romansk Institut, Odense Universitet, s.121-131.

Maas, Heinz-Dieter 1980: Zur Entwicklung des Übersetzungssystems SUSY und seiner einzelnen Komponenten, i Maschinelle Übersetzung, Lexikographie und Analyse, Akten des 2. Internationalen Kolloquiums Saarbrücken, November 1979, herausgegeben von Hans Eggers = Linguistische Arbeiten, Neue Folge, Heft 3,1 Universität des Saarlandes, Sonderforschungsbereich 100, s.7-16.

Maegaard, Bente og Ruus, Hanne 1978: DANwORD, Hyppighedsundersøgelser i moderne dansk: Baggrund og materiale, i Danske Studier 1978, s.42-70.

Maegaard, Bente og Ruus, Hanne: Hyppige Ord i Danske Børnebøger, Gyldendal, nov. 1981

Maegaard, Bente og Ruus, Hanne: Hyppige Ord i Danske Romaner, Gyldendal, nov. 1981.

McNaught, John 1981: Terminological Data Banks: a model for a British Linguistic Data Bank (LDB), i Aslib Proceedings 33 (7/8), July/August, s.297-308.

Nistrup Madsen, Bodil 1981: Nye veje inden for fagsproglig leksikografi, i SPRINT 2, Sproginstitutternes tidsskrift Handelshøjskolen i København, s.18-23.

Ralph, Bo 1979: Leksikologi som datalingvistik, i Nordiske Datalingvistikdage i København 9.-lo. oktober 1979, Foredrag udgivet af Bente Maegaard, s.161-170.

Riber Petersen, Pia 1981: Ordbøger og edb, i SAML 8, Udgivet af Københavns Universitets Institut for anvendt og matematisk lingvistik, s.179-191.

Sager, J.C. and McNaught, J. 1980: Feasibility Study of the Establishment of a Terminological Databank in the U.K., British Library R. & D. Report Nr. 5642.
Selective Survey of Terminological Databanks in Western Europe, British Library R. & D. Report Nr. 5643.
Model Specification of a Linguistic Databank for the U.K., British Library R. & D. Report Nr. 5644.