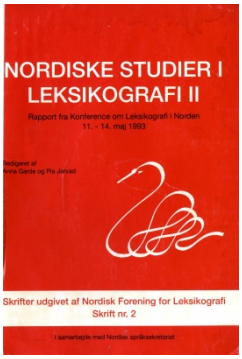
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Den Danske Ordbogs tekstkorpus og spORDhunde

Forfatter: Kjeld Kristensen

Kilde: Nordiske Studier i Leksikografi 2, 1993, s. 138-142
Rapport fra Konference om leksikografi i Norden, 11.-14. maj 1993

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.

Den Danske Ordbogs tekstkorpus og spORDhunde

Kjeld Kristensen

En spORDhund er et væsen, der har snuden i niveau med sprogets græsrødder eller lidt over, og med vibrerende næse og ører er på jagt efter ord, altsammen med henblik på det overordnede mål, Den Danske Ordbog. Den Danske Ordbog har kontakt med 600 spORDhunde, personer, der registrerer nye måder at bruge sproget på og indsender resultaterne til os på redaktionen. Jeg vender tilbage til spORDhundene, men til dem har vi netop udsendt det nyhedsbrev, I har fået et eksemplar af. På bagsiden kan man læse en punktvis præsentation af Den Danske Ordbog, og den kan måske være nyttig at se på som optakt til mit og de to følgende foredrag.

Den Danske Ordbog

- er en ordbog over dansk sprog i perioden fra ca. 1950 til i dag, med hovedvægten på de sidste ti år
- skal rumme over 100.000 opslagsord
- skal indeholde oplysninger om ordenes stavning, ordklasse, bøjning, udtale, betydning, konstruktion (forbindelser med andre ord) og etymologi (ordenes oprindelse og historie), og autentiske citater skal belyse ordenes brug
- skal vise sproget, som det er, men også vejlede brugeren
- skal være et udbredt og populært værk, tilgængeligt for både den professionelle og den utrænede bruger
- udarbejdes på grundlag af en elektronisk tekstsamling på 40 millioner løbende ord, andre ordbøger og leksika, Dansk Sprognævns seddelkartotek og mange tusinde ordsedler indsendt af flere hundrede spORDhunde
- redigeres af de tre ledende redaktører mag.art. Ebba Hjorth, cand.mag. Kjeld Kristensen og mag.scient. Ole Norling-Christensen samt en halv snes redaktører og konsulenter
- udgives af Det danske Sprog- og Litteraturselskab og finansieres af Kulturministeriet og Carlsbergfondet i fællesskab
- vil koste ca. 2000 kr. og udkommer i 6 bind på forlaget Gyldendal i 1998-1999

Mit indlæg skal især handle om ordbogens elektroniske tekstkorpus. Et elektronisk tekstkorpus kan defineres som en samling elektronisk læsbare tekster, sammensat ud fra nogle bestemte kriterier med henblik på et bestemt formål. Formålet er i dette tilfælde givet ved, at korpus skal gøre det muligt for os at gennemføre de lingvistiske undersøgelser og analyser, der sætter os i stand til at løse den opgave, vi har fået stillet, nemlig at redigere ordbogen. At korpus så også kan og bør bruges til mange, mange sprogvidenskabelige, litteraturvidenskabelige og andre undersøgelser, er så en anden sag. Men en konsekvens af det formål, som korpus skal sammensættes efter, er, at korpus på en eller anden måde skal afspejle moderne dansk sprog i perioden 1983-92. Korpus skal være repræsentativt for det tekstunivers, der hedder dansk 83-92 - eller skal det?

For det første er repræsentativitet, i hvert fald kvantitativ repræsentativitet, et idealt krav. For det andet vil jeg hævde, at vi ved opbygningen af vores korpus hverken kan eller skal stræbe efter at opfylde dette ideale krav. Man kan nemlig stille sig selv en række spørgsmål, fx: Tales der flest ord eller skrives der flest ord en tilfældig dag i Danmark? Der tales langt flere ord, end der skrives. Skal der så ikke være mest talesprog i vores korpus (forudsat vi også ønsker at beskrive talesproget i ordbogen - hvad vi gør)? Nej, dels vil det naturligvis være alt, alt for ressourcekrævende at optage og udskrive alt det talesprog (især da langt det meste produceres face-to-face i hverdagens almindelige situationer), dels ville man derved få underbelyst skriftsproget, der jo oftest udbredes videre end talesproget; det er især de professionelle skribenters, bl.a. journalisters og forfatteres kommunikation med de mange i aviser, blade og bøger. Og hvordan skal vi repræsentere professionelt sprog i henseende til medium, genre og emne: skal vi vægte efter oplagstal og seer- og lyttertal? Nej, repræsentativitet er umulig at opnå, og den er heller ikke ønskelig. Det, det gælder om, er at etablere en afbalanceret samling eksempler på dansk sprog brugt i tale og skrift, som almensprog og som fagligt sprog, spredt over et stort antal medier og genrer og emner, og produceret af danskere af begge køn, i forskellige aldre og med forskellig uddannelse og erhverv. Det skal, så vidt ressourcerne rækker, være et eksemplarisk, afbalanceret korpus. Det er bredden, alsidigheden, der tæller. Og skulle man så ved brugen af korpus opdage, at sammensætningen er lidt skæv på et punkt eller to, kan man jo løbende udskifte nogle af teksterne med andre - eller supplere.

På to punkter har vi bestræbt os på at give vores korpus en særlig drejning, som ligger i forlængelse af den bredde, der skal præge ordbogen. Den skal være bred i anlægget, dvs. vise sproget, som det er (en deskriptiv ordbog), og ordbogen skal være et udbredt og populært værk. Derfor skal det ikke bare være det professionelle skriftsprog, der beskrives i ordbogen; Den Danske Ordbog skal ikke kun være en litterært baseret ordbog som Ordbog over det danske Sprog, domineret af citater fra tidens førende penne. Vi bygger så vidt muligt også på almindelige menneskers dagligdags sprog, ikke mindst deres talesprog. Vi lægger vægt på at få meget produktion med i vores korpus, altså sprog, som folk selv frembringer, i modsætning til det sprog, som folk læser eller hører, (reception). Og vi lægger vægt på at få meget talesprog med. Vi samler breve, dagbøger, lejlighedssange, læserbreve, opgaver og danske stile som eksempler på produktion, og vi

samler interviews og gruppesamtaler fra adskillige sprogvidenskabelige og andre projekter. Desuden en mængde udskrifter af talesprog i radio og tv og af taler i Folketinget og Københavns Borgerrepræsentation. Vi etablerer en samling talesprogstekster, der siger sparto til det meste af, hvad der findes af den slags i hele verden: ca. 5 mio. løbende ord. Alligevel vil professionelle skribenters sprog blive dominerende i korpus, deres tekster er lettest og billigst at få fat på, men alligevel...

Det praktiske arbejde med at indsamle tekster og opbygge korpus foregår ved, at vi tager kontakt med avisredaktioner, bladredaktioner, forlag og andre instanser, vi regner med har noget at tilbyde. Vi har overalt mødt en kolossal imødekommenhed og velvilje uden økonomiske bagtanker. Vi får så teksterne på diskette eller tape, i et eller andet format, konverterer fra dette format til DOS, retter teksterne til i WordPerfect efter vores standarder og lægger dem ind i vores database, PARADOX. Hver tekst forsynes med en såkaldt header med oplysninger om teksten. Jeg vil kort gennemgå, hvilke oplysninger sådan en header indeholder.

Header

Kildeangivelse	
Tekstgruppe	entydig identifikation af en gruppe (beslægtede) tekster
Tekstnr	løbenummer inden for tekstgruppen
Restriktion?	
Restriktion_a	tekstens navne skal anonymiseres: "ja"/"nej"
Restriktion_b	teksten må ikke bruges til andet end ordbogen: ja"/"nej"
Udløbsår	for restriktion b
Sprogbruger+	
Rolle?	især ved talesprog, fx "interviewer og "interviewperson"
Identifikation?	entydig tretegnskode, hvortil replikker i talesprogstekster kan referere
Efternavn?	hvis det kendes
Fornavn?	hvis det kendes
***Køn**	"m"/"k"/"u[kendt]"
Uddannelse?	hvis den kendes
Erhverv?	hvis det kendes
***Fødselsår?**	et heltal mellem 1880 og 1990, hvis året kendes
Sikker?	"?", hvis året ikke kendes med sikkerhed
Fødested?	hvis det kendes
***Sprogvariant**	"rigssprog"/"regionalsprog"
Teksttitel?	hvis der er en titel
Værktitel?	navn på fx antologi, avis, blad, 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 (begge inkl.)
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 med 12 medier, bl.a. bog, avis, blad, radio
***Genre?**	én fra en liste med 124 delvis medieafhængige genrer, fx roman, brev, interview, nyhedsudsendelse
***Emne?**	ét fra en liste med 64 emner, fx biologi, litteratur, politik, fysik
Omfang	antal løbende ord i teksten (beregnes maskinelt)

Indrykningerne viser den hierarkiske struktur. Kildeangivelsen indeholder nogle facts om teksten, som umiddelbart foreligger, eller som kan slås op i egnede håndbøger. Tekstbeskrivelsen er mere en ekstern sproglig beskrivelse af teksten mht. en række parametre. * betyder, at det, man skal skrive i det pågældende felt, er hentet fra et lukket univers; man skal altså vælge én ud af en begrænset mængde bestemte værdier. + angiver, at feltet er multipelt (der kan være flere sprogbrugere), ? at feltet er fakultativt.

Meningen med headeroplysningerne er, at man maskinelt opbevarer de oplysninger om teksten, man senere får brug for, fx som kildeoplysninger ved citater, eller som man skal bruge ved de lingvistiske analyser af ordenes brug og betydning, fx en undersøgelse af, om et ord i den og den betydning især bruges af børn og unge, især i talesprog, osv. Man kan også foretage selektive søgninger i korpus, fx af mænds sprogbrug i læserbreve, eller af det faglige sprog i tekster om biologi. Det er bl.a. her, man har glæde af felterne med det lukkede univers af værdier.

Til sidst lidt om spORDhunde-initiativet. Det har et engelsk forbillede i Word Watcher-kampagnen. Men vi har altså valgt et sjovere og mere kreativt ord. Vi havde en artikel i brugsforeningsbevægelsens blad "Samvirke" i september 92, med en præsentation af ordbogen og en opfordring til læserne om at henvende sig til os, hvis de var interesserede i at finde ord til os. Der kom 600 henvendelser pr. telefonsvarer og brev. Vi udsendte informationsmateriale og ordsedler til alle, og der er indtil nu kommet 3-4000 udfyldte ordsedler retur fra over 200 spORDhunde. Hensigten med kampagnen er at få knyttet

en masse sproginteresserede danskere til Den Danske Ordbog - det er jo fremtidige kunder, men det viser sig også, at det materiale, vi får ind, er virkelig værdifuldt. Vi har i nyhedsbrevet skrevet en lille artikel med de første resultater: nye ord, etablerede ord og udtryk med ny betydning, nye slangudtryk. Det er kun et lille og lidt tilfældigt udvalg; der er som sagt kommet flere tusinde sedler ind. Oplysningerne fra sedlerne lægges ind i en database til senere brug. Og kampagnen fortsætter.