Danske kerneord

Anmeldt af Svend Bruhns

Hanne Ruus: ***Danske kerneord : centrale dele af den danske leksikalske norm.*** **Museum Tusculanums Forlag, 1995. Bd.1-2 (226 + 313 s.).**

Hanne Ruus har skrevet en doktordisputats om »Centrale dele af den danske leksikalske norm«, dvs en beskrivelse og vurdering af en metode til på sprogligt-statistisk grundlag at finde de basale ord i dansk skriftsprog, kerneordene. Afhandlingen rummer et helt lexikon over de fundne hyppigste danske ord. Desuden diskuterer forfatteren semantiske relationer og ordnetværker. Der er meget i denne afhandling som vil være nyttigt for bibliotekarer som beskæftiger sig med indexering og klassifikation. Bogens egentlige text er på mindre end 150 sider, det øvrige - herunder hele bind 2 - er lister over kerneordene. Nogle af kapitlerne er naturligvis ret teoretiske, men bogen er velskrevet, og begreber og terminologi forklares på stedet. Forfatterens »testlæsere anbefaler et sindigt læsetempo« men siger også korrekt at »alligevel kommer man hurtigt frem til det væsentlige«.

Hanne Ruus er magister i nordisk filologi, ansat ved Københavns Universitet, og har i en del år beskæftiget sig med hyppighedsundersøgelser af danske texter. Materialet stammer fra det store DANwORD-projekt som Hanne Ruus lavede i samarbejde med nuværende professor Bente Maegaard fra KU. DANwORD-materialet er sammensat af de mest læste texter udgivet 1970-74 (dvs hyppigst udgivet/udgivet i størst oplag) fra fem textarter: Aviser, fagblade, ugeblade, romaner og børnebøger. Hele dette materiale - korpus som man siger - består af 1.250.000 løbende ord statistisk udvalgt fra 1000 textprøver fra hver af de fem textarter. Lister over de hyppigste ordformer er publiceret i 4 bind *Hyppige ord i danske børnebøger* osv, Gyldendal, 1981-86. Bøgerne rummer bare de nøgne ord, ordnet både efter hyppighed og alfabetisk, men korpus findes naturligvis i en database (åbenbart en relationsdatabase) og her findes ordene i kontext.

Disputatsen går først og fremmest ud på at under-

søge hvordan man igen kan finde *crème de la crème* af et sådant korpus og præsenterer faktisk resultatet, altså de 1117 danske kerneord.

Det første skridt var at gennemgå de 14.948 forskellige ordformer som fandtes i DANwORD-materialet og lemmatisere dem. Lemmatisere vil sige at sammenføre alle et ords bøjningsformer til en grundform, »lemmaet«, der svarer til den form man bruger som opslagsform i ordbøger. Hvis nogle oplagte former, især grundformen, manglede i materialet måtte den indføjes. Listen der kom ud af det var altså delvis teoretisk. Her var flere problemer, især ved homografer (altså forskellige ord der skrives ens). Af flere grunde valgte Hanne Ruus en kun delvis automatisk procedure, især fordi helautomatik ganske vist ville have givet en mere fuldstændig ordbog, men også have medført teoretisk mulige former som næppe kunne registreres i sprogbrugen. Disse blev filtreret fra af det menneskelige filter.

Det næste problem var at finde kriterier for »kerneordskhed«! Det første kriterium er naturligvis at lemmaet findes meget hyppigt, og det blev udregnet teoretisk. Det andet kriterium er at lemmaet faktisk forekommer i en bestemt textart, og hertil blev valgt ugeblade, hvilket skyldes at ordforrådet i ugebladene havde de færreste specielle ord af de textarter som er med i DANwORD. Tærskelen for kerneordskhed blev så at forekomsterne af et lemmas forskellige ordformer skulle være 20 eller derover i ugebladsordlisten. November er en trist måned efter min og åbenbart også Ugebladenes smag, thi den er ikke med blandt kerneordene. Det er ikke svært at finde andre »inkonsekvenser«, for grundlaget er jo simpel optælling i bestemte texter, men stort set er det nyttige lister. Kerneordene findes dels i lister efter ordklasse, og dels i een stor alfabetisk liste (=bd.2), hvori man også kan se antal forekomster i de øvrige textarter (Avis osv). Her er også medtaget lemmaer som teoretisk kunne have været med, men som ikke havde 20 forekomster i ugebladene. Jeg er ikke fuldstændigt overbevist om at disse statistisk fundne kerneord virkelig er de nødvendigste ord i dansk; statistikken må suppleres af sund fornuft & filologi.

Et sådant arbejde kan have interesse på mange områder i bdi-sammenhæng. Her vil jeg blot nævne fænomenet »stemming«; dette engelske ord er afledt af ordet *stem* 'ordstamme' og det er det samme som lemmatisering.

»Stemming« er ret almindeligt på Internettet Det skulle således i InfoSeek give samme resultat ligegyldigt om man søger på agreed eller agrees eller agreeing for alle disse ordformer »stemmes« til agree. Men det er ikke en trunkering, for vi skulle ikke få de irrelevante termer med. En førsteklasses stemmingsfacilitet af især emneord er derfor at foretrække for trunkering. Det kræver imidlertid at systemet har en art ordbog over lemmaer.

Den anden del af bogen handler om ordnetværker og de tre semantiske relationer: antonymi; over-underordningsrelationer; del-helhedsrelationer. Især denne del af bogen bygger på resultater fra filosofiske studier af helt grundlæggende relationer (bl.a.især fra Peter Zinkernagel og Lakoff). Hver af de tre relationsarter behandles teoretisk, og man skal være klar i hjernen når man læser afsnittet, men det kan lade sig gøre, bl.a. fordi der er gode exempler. Især afsnittet om antonymi, altså modsætninger, er instruktivt. Jeg kunne nu godt have tænkt mig at se en behandling af en relation som *årsag-virkning*. »Årsag« er iøvrigt slet ikke kerneord, men "virkning" er.

Hanne Ruus præsenterer derefter de semantiske relationer mellem kerneordene som hun har kunnet registrere, bl.a. ved hjælp af Harry Andersen: *Dansk Begrebsordbog* (1945). Et værk som ikke mindst vi informationssøgere i højeste grad savner en ny udgave af. Jeg kan ikke lade være med at tænke at en ny udgave af *Dansk Begrebsordbog* med udgangspunkt i Hanne Ruus' undersøgelser burde have form som en hypertext, så man kunne klikke på ordene og deres relationer og derved få sådanne ordnetværker frem som Hanne Ruus omtaler, men kun præsenterer i skitseform.

Hanne Ruus gør selv på s.120 opmærksom på at terminologer [som ofte er teknisk uddannede.SvB] og semantiske lexikografer i stor udstrækning ignorerer hinanden [de citerer ikke hinanden, som vi citationister ville sige], men for os bdi-folk er det også påfaldende at Hanne Ruus ikke citerer nogen bdi-klassifikationsforsker. Zipf - den store ordstatistiker - er heller ikke nævnt, eller de mange andre informationsforskere som er kendt på Biblioteksskolen, i hvert fald på Overbygningsuddannelsen. Kun lingvister og filosoffer som Lakoff. Det er nok ensidigt op til bdi-forskerne hvis der skal skabes et fælles forskningsområde.

PS Det er godt at der er registre i bogen, ordregister, som der gerne er i sprogvidenskabelige bøger, og sagregister. Men sagregistret er helt til grin, et typisk exempel på bevidstløs anvendelse af registerfunktionen i et textbehandlingsprogram. Man skal fx slå op på *Anbringelse...* for at finde henvisning til siden hvor det omtales hvordan man anbringer manglende lemmaer, men der er ikke noget opslagsord [Lemmaer, manglende] eller omvendt.

PPS. *Informationsordbogen* af Friis-Hansen et al. (Dansk Standard) gir gode forklaringer på relevante sprogvidenskabelige begreber.