NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Leksikografisk tradition og fornyelse: tre revolutioner på 100 år?

Forfatter: Lars Trap-Jensen

Kilde: Nordiske Studier i Leksikografi 12, 2013, s. 42-68
Rapport fra Konferanse om leksikografi i Norden, Oslo 13.-16. august 2013

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.

Leksikografisk tradition og fornyelse: tre revolutioner på 100 år?

Lars Trap-Jensen

Viewed over the last 20 years, the development in lexicography has been overwhelming: many products have changed completely and so has the way of producing lexicographic products. The Internet, computer technology and publication forms are all different today from what they were just a few years ago, and like everyone else lexicographers have to deal with the situation. In many ways it is exciting times and many have probably asked themselves: is it really a paradigm shift we are going through, or is the whole thing just hot air? In this article I focus on three potential "revolutions" and assess in what way they have changed the conditions for practical lexicographic work and the way we think about dictionaries.

1. Indledning

Hvis man betragter udviklingen på ordbogsmarkedet de seneste tyve år, bliver man hurtigt forpustet. De fleste produkter har ændret sig grundlæggende, og måden vi laver ordbøger på, har ændret sig. Internettet, computerteknologien, publiceringsformerne, stort set alt har ændret sig på få år, og som alle andre

må leksikografer og forlag forsøge at hænge på udviklingen så godt de kan. Det er på mange måder en spændende tid, men mange har sikkert også indimellem spurgt sig selv: Er det et reelt paradigmeskift vi oplever, eller er det hele bare varm luft? I denne artikel fokuserer jeg på tre perioder inden for de sidste 100 år som alle kan kaldes potentielle "revolutioner" inden for leksikografien. Jeg ser på hvilke spor de har sat sig, og på hvordan de har ændret det praktiske arbejde og måden vi tænker om ordbøger på.

2. Den første revolution: det deskriptive paradigme

Hvis man går ca. 100 år tilbage og betragter ordbogsproduktionen omkring år 1900, er det ikke svært at få øje på mange forskelle. Det er omkring dette tidspunkt de nationale modersmålsordbøger er skudt frem og har fundet sin form med kartotekskasser fulde af excerpter og en hær af flittige leksikografer der producerer det ene bind efter det andet, sådan som vi kender det fra OED i England, SAOB i Sverige og ODS i Danmark. De er i dag kendte og klassiske værker hvis akronymer kendes langt ud over leksikografernes snævre kreds. Men selvom projekterne er gamle og hjælpemidlerne kan forekomme simple, adskiller metoden sig ikke principielt fra den vi kender og bruger i dag: beskrivelse af sproget og sprogbrugen på empirisk grundlag – korpusbaseret om man vil. Dengang bestod korpus af excerperede sprogprøver der blev opbevaret i kartotekskasser og analyseret for hvert ord og betydning. I Danmark

udgjorde denne praksis et virkeligt nybrud, et paradigmeskift i klassisk, kuhnsk forstand: ODS gjorde op med 1800-tallets fremherskende tradition og insisterede på at være deskriptiv. Man kan kalde det for *den deskriptive revolution.*

Traditionen før ODS var præget af det man har kaldt *akademiprincippet* med henvisning til den praksis som var gældende for Det Franske Akademis ordbog. I overensstemmelse med dette princip var ordbøgerne før ODS *præskriptive*, de var båret af et pædagogisk hensyn og et ønske om at opdrage befolkningen til det gode og smukke sprog med gode eksempler og derved opsætte en rettesnor for det. Jacob Langebek, en af pionererne bag Videnskabernes Selskabs Ordbog, mente at kun de «gode, rene, overalt brugelige og upaatvivlelige Danske» ord burde medtages, mens der omvendt ikke var plads til:

> … Alle grove, plumpe og liderlige Oord og Talemaader, som stride imod Ærbarheden … thi de behøves ikke at vides for dem, der ikke skiøtte derom, og de som have lyst til at vide dem, lære dem nok foruden des. (Langebek ca. 1738)

Ordbøgerne skal altså være selektive og kun medtage de gode ord, hvorimod de "grove, plumpe og liderlige" ikke har nogen plads i ordbogen.

Den samme tanke finder vi hos Molbech:

> Selv den hyppigste Brug af et nydannet Ord, især i Talesproget, giver det ingen Auctoritet, og beviser intet for dets Brugbarhed i den rene tale og gode Stiil, eller for dets Optagelighed i en Ordbog, saalænge det støder et for Sprogbrugen dannet Øre og den finere Sprogsands. (Molbech 1833:XIII, genoptrykt i Molbech 1859:VIII)

Ordbøgerne skulle kun indeholde «gode» ord, «sprogets skönneste blomster». «Det skal så at sige være en ære for et ord at blive optaget i ordbogen, ligesom det er en ære for et kunstværk at få plads i statens kunstsamlinger»– sådan udlægger ODS' grundlægger, Verner Dahlerup, Molbechs princip (Dahlerup 1907:68). Selv skriver Molbech at hans ordbog skal være «en Tolk for det rene, det dannede Skriftsprogs rigtige Brug i vor nærværende Alder».

Et sidste eksempel på akademiprincippet stammer fra den første danske slangordbog, *Ordbog over Gadesproget og saakaldt daglig Tale.* Forfatteren er V. Kristiansen, et pseudonym for professor Viggo Fausbøll, professor i indisk-østerlandsk filologi ved Københavns Universitet. Han skriver i forordet til ordbogen at formålet med ordbogen ikke er oplysende, men tværtimod at advare mod det vulgære sprog:

> Dette vulgære sprog … truer i den sidste tid med at trænge ind i familierne ... Ved her at have samlet noget af hvad der hører herhen, har jeg, næst efter at have haft et rent sprogligt formål, tillige i vort land villet henlede opmærksomheden på faren og søge at vække modstanden mod samme, og jeg antager, at når folk en gang have fået öjet åbnet for den utilbørlige overskridelse af grændsen, ville alle dannede mennesker være enige om at bandlyse gadesproget fra ethvert godt selskab og overlade det til gadedrenge og de Grundtvigianere, i hvis smag det falder. (Kristiansen 1866:V)

I førsteudgaven af ordbogen er de mest vulgære ord derfor sat med græske bogstaver så den brede befolkning ikke skulle blive fordærvet af ord de bestemt ikke behøvede at kende.

En anden ordbog man kunne nævne i samme forbindelse, er

afløseren for Molbechs ordbog, der udkom omtrent samtidig med at ODS gik i gang, nemlig Dahl og Hammers *Dansk Ordbog for Folket,* der bl.a. er kendt for at indeholde mange puristiske afløsningsord (fx *haandig* for *manuel*, *lemfaldsyge* for *spedalskhed* og *akselblad* for *skulderblad*), igen ud fra et ønske om at opdrage befolkningen og skåne den for uønsket ordstof, hvad enten det var fremmedord, vulgært sprog, dialekter eller andre former for lavsprog.

Det er denne præskriptive tradition ODS gør op med**.** Dahlerup lagde vægt på at ODS skulle være et videnskabeligt og praktisk hjælpemiddel til forståelse af sproget, og så nytter det ikke at udelade ord i sproget fordi man mener de er dårlige eller skadelige:

> jeg kan ikke først og fremmest spørge: "bör det og det ord bruges?", men: "bruges det, eller har det været brugt?"; hvis dette er tilfældet, optager jeg ordet, for så vidt hensynet til bogens omfang tillader det. (Dahlerup 1907:71)

Citatet stammer fra en artikel i Danske Studier, hvori Dahlerup redegør for sine overvejelser om den nye, store ordbog. Dahlerup var ikke den eneste der mente at den faktiske sprogbrug var nøglen til semantisk beskrivelse. Synspunktet var fremme i tiden med junggrammatikernes optagethed af samtidssproget som centralt for forståelsen af de undtagelsesløse lydlove og blev knæsat inden for filosofi og sprogvidenskab i løbet af det 20. århundrede. Princippet om empirisk analyse som grundlaget for en god beskrivelse følger vi stadig i dag, 100 år efter. Men det empiriske beskrivelsesgrundlag har ændret sig, ganske meget endda.

3. Den anden revolution: korpusrevolutionen

Korpusrevolutionen er det næste store spring der fundamentalt ændrede måden at lave ordbøger på. Teoretisk-metodisk er excerpering af tekster og belæg på kartotekskort ganske vist ikke grundlæggende forskelligt fra korpuskonkordanser og betydningsannotering – man kan sige at det mere er en forskel mellem analog og digital metode. Men den tekniske udvikling der fandt sted fra 1960'erne og fremefter, betød at en stor del af arbejdet kunne automatiseres med computerens hjælp. Det medførte en betragtelig udvidelse af beskrivelsesgrundlaget samtidig med at arbejdet kunne udføres for færre resurser.

> The proper way to describe a word is to identify the grammatical constructions in which it participates and to characterize all of the obligatory and optional types of companions (complements, modifiers, adjuncts, etc.) which the word can have in such constructions, in so far as the occurrence of such accompanying elements is dependent in some way on the meaning of the word being described. (Fillmore 1995)

Potentialet i brugen af korpus er enormt hvis man tager det alvorligt – som Fillmore her udtrykker det – at det hører med til beskrivelsen af et ord at undersøge samtlige forbindelser ordet indgår med andre ord for at se om de på den ene eller anden måde betinger det pågældende ords betydning. Fillmore har selv gjort det i sit FrameNet-projekt, og også inden for leksikografien har vi set projekter der er optaget af tankegangen, mest direkte i den engelske DANTE-database der eksplicit bygger på tankerne fra FrameNet (jf. Atkins 2010), men også i en række

andre projekter er der en tendens til at fokus forskydes fra *lemmaet* som leksikografisk grundenhed til en mindre enhed bestående af en betydning og dens tilhørende lemmaform. Sådan er det i DANTE, og det er også sådan en udvikling som Det Danske Sprog- og Litteraturselskabs projekter har gennemgået, og som gør det muligt at knytte Den Danske Ordbogs (DDO) ord og betydninger til enhederne i afledte projekter som DanNet og Den Danske Begrebsordbog (jf. Lorentzen, Nimb & Troelsgård denne udgivelse).

Inden for leksikografien tog korpusrevolutionen for alvor fat omkring 1980 med COBUILD-projektet i Birmingham. I første omgang var den største fordel at eksempelmaterialet blev meget større i forhold til den hidtidige fremgangsmåde med excerpering og kartotekskasser. Hvis man alene ser på datamængden målt i antal løbende ord, er korpusserne steget med noget der ligner en tidobling for hver 10-15 år: I 1960'erne og 70'erne var de første korpusser på omkring 1 mio. ord. Norden var her godt repræsenteret blandt de allerførste med Sture Alléns arbejde i 1960'erne og Press65. COBUILD's korpus var i 1985 på ca. 18 mio. ord og rummer i dag omkring 650 mio. ord under navnet Bank of English. DDO's korpus fra midten af 1990'erne var på 40 mio., mens British National Corpus nåede 100 mio. omtrent samtidig med DDO's i midten af 1990'erne. Til de største korpusser vi har i dag, hører det tyske COSMAS (eller DeReKo: Deutsches Referenzkorpus) fra IDS i Mannheim med over 5 mia. løbende ord og Googles korpus over tekster der er blevet scannet i forbindelse med projektet Google Books: over 500 mio. bøger fra år 1500 til i dag for en række af de større sprog. Om korpusset indeholder 155 mia. ord eller 175 eller 200, er ikke afgørende. Kvantitativt er det tæt på hvad de fleste bare regner for uendelig stort. Norden kan ikke helt være med størrelsesmæssigt, men Språkbanken i Göteborg har dog

mere end 1 mia. ord i deres korpussamlinger, og også det norske aviskorpus indeholder mere end 1 mia. ord. Danmark ligger i sammenligning hermed noget lavere. Hos Det Danske Sprog- og Litteraturselskab forsøger vi at indsamle lidt bredere og har vel i alt tekster med omkring ½ mia. ord, men ikke alle tekster er offentligt tilgængelige. Det vil de efter planen blive for en del teksters vedkommende i løbet af 2014.

Man kan spore en parallel udvikling i brugen af korpus til ordbogsarbejde svarende til væksten i de tilgængelige korpusser. I de tidlige år var den helt store gevinst at man hurtigt fik adgang til et langt større eksempelmateriale end hvad der havde været muligt ved hjælp af kartotekskort. For leksikografen bestod arbejdet i at kigge konkordanser igennem og ordne forekomsterne i homografer og betydninger, svarende til arbejdet med kartotekskort. Men enhver der har arbejdet med konkordanser, ved også at selv i mindre korpusser bliver opgaven en stor mundfuld når det drejer sig om at undersøge sprogets relativt almindelige ord fordi de mange konkordanslinjer hurtigt bliver uoverskuelige.

Det var derfor et fremskridt da man begyndte at få annoterede korpusser. Ved at begrænse søgningen til fx verbalforekomster af en homograf eller kun udvalgte bøjningsformer af et ord bliver konkordansen renset for uønskede forekomster og arbejdet dermed lettere og hurtigere for leksikografen at udføre.

Omkring årtusindskiftet kom de syntaksopmærkede korpusser, hvad der yderligere gjorde det muligt at finde prægnante mønstre i teksterne, fx i form af såkaldte Word Sketches (Kilgarriff m.fl. 2004) eller andre former for leksikalske profiler. Og samtidig med at korpusserne voksede voldsomt i volumen, blev det i stigende grad nødvendigt at gøre noget ved den overvældende informationsmængde der fulgte med. Tendensen går derfor i retning af mere og mere præprocessering af materialet

ved hjælp af teknikker som forhåndsanalyserer teksterne efter forskellige parametre som gør det muligt for redaktørerne at finde netop de forekomster de har brug for på det aktuelle sted i redigeringsprocessen. Lad os se på nogle af de muligheder som udforskes.

At opdele korpusforekomster i homografer og betydninger står centralt i enhver leksikografs daglige arbejde, og selvom der endnu ikke – mig bekendt – er udviklet en operativ teknik til automatisk sortering af forekomsterne, skal det alligevel nævnes først da konkordansopstillingen fra første færd havde dette som formål. Dog har betydningsannotering indtil videre været noget som redaktørerne måtte foretage manuelt. I korpuslingvistikkens nuværende fase kan leksikalske profiler hjælpe med til at afsløre nogle betydninger, idet der er en sammenhæng mellem semantisk beskrivelse og fx valensmønstre eller fagtilknytning. Begge dele kan ofte påvises automatisk ved hjælp af korpuslingvistiske metoder.

Brug af korpus som redskab i lemmaselektionen er en anden indlysende mulighed. Korpusfrekvens er et af de parametre der indgår når man skal afgøre hvilke ord der skal med i ens ordbog.

Statistiske metoder som Mutual Information, T-score m.fl. er metoder der er egnede til at påvise hvordan ord tiltrækker hinanden. Alt hvad falder inden for det fraseologiske område, kan man derfor få god hjælp til med korpuslingvistiske teknikker. Valensmønstre og leksikalske profiler bør nævnes i samme forbindelse da de findes med samme teknik som de fraseologiske kombinationer, blot med den forskel at det ikke er ordenes direkte tiltrækning af hinanden der måles, men udfyldningen af syntaktiske kategorier og ledfunktioner i det opmærkede korpus.

En forholdsvis ny og derfor mindre kendt anvendelse er at

bruge korpus til at overvåge sproglig udvikling, mest oplagt til at finde neologismer ved hjælp af et dynamisk monitorkorpus der analyseres diakront med henblik på at afsløre om der dukker ord eller samforekomster af ord op i de nyere tekster som er fraværende i et referencekorpus som der sammenlignes med (se fx Halskov 2010). Hvis det er tilfældet, kan det være tegn på en sproglig nydannelse. Metoden er især oplagt til at finde helt nye ord, mens det ikke er helt så enkelt at finde nye betydninger af eksisterende ord og flerordsforbindelser af eksisterende ord. Den samme metode kan bruges mere generelt: Ved at sammenligne et såkaldt *fokuskorpus* med et *referencekorpus* (jf. Cook m.fl. 2013) kan overhyppige forekomster i fokuskorpusset påvises og analyseres. Hvis der viser sig en ny kombination af eksisterende ord eller et ændret syntaktisk mønster, kan det være tegn på et nyt udtryk eller en ny betydning i sproget.

På samme måde kan sammenligning af et fagsprogligt fokuskorpus med et almensprogligt referencekorpus eksempelvis bruges til først at karakterisere fagsproget, dernæst identificere en bestemt type fagtekst blandt andre tekster og endelig påvise en udvikling af fagsproget. Hvis fx ord som *virus, orm* eller *sky* dukker op med en højere frekvens end normalt i tekster der handler om IT, er det et tegn på at der er tale om nye betydninger af de ord. Retningen kan også være den modsatte, fra det faglige til det almene, når man finder udtryk fra sportsverdenen i almensproglige tekster: *sænke paraderne, stå på mål for en sag* eller *skyde til hjørne* osv.

Mere generelt kan teknikken endvidere bruges som leksikografisk redskab i arbejdet med at karakterisere sprogbrugen ved hjælp af sprogbrugsmarkører. Hvis sammenligningen viser at et bestemt ord eller udtryk er overrepræsenteret i tekster fra et bestemt domæne, er det sandsynligt at der er tale om en faglig betydning. Hvis korpusteksterne er forsynet med metaoplys-

ninger om det, kan sammenligningen ligeledes vise om bestemte sprogbrugere eller teksttyper er overrepræsenterede.

Gode sprogbrugseksempler er et andet område hvor der har været bestræbelser på at præprocessere korpusmaterialet således at de bedst egnede citater kommer til at stå først i konkordansen. Som bekendt er det resursekrævende at finde frem til de mest velegnede citater, så her er et område hvor der kan spares tid og penge ved at præsentere materialet hensigtsmæssigt. Noget af det der kendetegner et godt citat, er at det består af en hel sætning, hverken for lang eller for kort (vel omkring 10-20 ord, måske 25), det bør ikke indeholde proprier (fordi navngivne personer eller andet navnestof kræver særlig viden); den bør ikke indeholde deiktiske udtryk eller pronominer med reference uden for sætningen; og det bør ikke indeholde svære eller sjældne ord, men derimod gerne en typisk kollokation (se også Lorentzen 1999). Den slags kriterier kan fastlægges på forhånd sådan at leksikografen som det første bliver præsenteret for de eksempler der opfylder kriterierne, og som regel vil der vise sig at være citater der egner sig til at komme i ordbogen. Denne måde at arbejde på kan fx bruges sammen med korpussøgeværktøjet SketchEngine, og et tilsvarende system er udviklet ved Berlin-Brandenburgische Akademie der Wissenschaften for tysk (Didakowski, Lemnitzer & 2012).

Dette er nogle af de muligheder der findes, men mulighederne for at udnytte korpus er dermed ikke udtømt. Et mål har længe været at kunne opdele konkordanserne i semantisk meningsfulde dele sådan at korpussøgeprogrammet præsenterer leksikografen for et automatisk forslag til betydningsopdeling. Mig bekendt er der dog ikke udviklet en lovende metode til semantisk tagging endnu, men der eksperimenteres med forskellige fremgangsmåder. Det Danske Sprog- og Litteraturselskab deltager eksempelvis i et treårigt forskningsprojekt i

samarbejde med Center for Sprogteknologi, Københavns Universitet, der har til formål dels at udvikle en maskinlæringsmetode, dels at udføre semantisk annotering af vores resurser, bl.a. ved hjælp af DanNet-resursen.

Sammenfattende kan korpusrevolutionen ikke betegnes som et paradigmeskift i kuhnsk forstand. Den videreudvikler den deskriptive tradition som blev grundlagt med de klassiske, store nationalordbøger omkring år 1900. Det var her idealet om ordbogen som et spejl af sproget blev fastlagt, men de store korpusser gjorde det muligt at komme gradvis tættere på idealet. Beskrivelsesfeltet blev udvidet fra at fokusere på det «gode» sprog hos de klassiske forfattere til at omfatte flere og flere andre tekstgenrer indtil vi med fx BNC og også DDO's korpus forsøgte at nærme os noget der lignede sproget i sin mangfoldighed, indfanget i et enkelt korpus. I 1990'erne mente vi at repræsentativitet var mindst lige så vigtigt som volumen, men den tanke er flere steder ved at blive forladt, formentlig mere af praktiske end teoretiske grunde. Balancerede korpusser er dyrere at udvikle, og det er kendetegnende for de store korpusser med milliarder af ord at de ikke er særlig godt balancerede. De består enten overvejende af avistekster (COSMAS, de store norske og svenske korpusser) eller er indhøstet fra nettet (fx Oxfords English Corpus, 2 mia.). Tilgængelighed er nøgleordet (se også Jakubíček m.fl. 2013).

Hvis man anskuer udviklingen udefra, er korpusrevolutionen formentlig ikke noget ret mange uden for den leksikografiske verden overhovedet har bemærket, for den har næppe ført til radikalt anderledes ordbøger. Den var til gengæld interessant for leksikograferne fordi den gav os et forbedret beskrivelsesgrundlag og dermed mulighed for at lave bedre ordbøger.

De voksende korpora har ført til ændrede arbejdsbetingelser og en anden rollefordeling for os der arbejder som redaktører. I

gamle dage stod redaktøren for næsten alt arbejdet med at redigere artiklen; det eneste præprocesserede materiale der fandtes, var kassen med excerpter, resten var redaktørens ansvar. Sådan er det ikke længere. I korpusleksikografiens tidlige dage var der stadig mange kreative opgaver for redaktøren: Korpus var et hjælperedskab, mens det var redaktørens rolle at gennemlæse konkordanserne og sortere og udvælge fra dem. Men heller ikke det er muligt når materialet bliver for stort. Redaktøren bliver i stigende grad præsenteret for en række halvfabrikata i form af forslag til oplysninger som computeren på forhånd har analyseret sig frem til: staveformer, bøjningsformer, valensmønstre, kollokationer, idiomer, morfologiske og syntaktiske begænsninger, måske citater og sprogbrugsmarkører for at nævne de vigtigste. Redaktørens rolle bliver at kontrollere, validere og vælge blandt det materiale som kommer ud af præprocesseringen. Der sker en ændring i redaktørens arbejdsbetingelser og -funktioner og dermed også i de kvalifikationer der kræves af en dygtig redaktør. Nogle færdigheder er ikke længere så vigtige, mens andre kommer til.

4. Den digitale revolution

Den tredje store begivenhed som jeg vil nævne, er den udvikling som vi befinder os midt i. Nogle kalder det for den elektroniske eller den digitale revolution: den udvikling der har ændret ordbogen fra papirprodukt til elektronisk produkt. Udviklingen er sket gradvis over de sidste 30 år. SAOB var blandt pionererne i denne udvikling da digitaliseringen begyndte i 1983 med

OSA-projektet. I 1990'erne fik vi cd-rom'er og PDA'er, men bortset fra at disse produkter tilbød bedre søgefaciliteter, var det grundlæggende de samme produkter. I løbet af 2000-tallet er næsten enhver ordbog med respekt for sig selv flyttet over fra papir til en elektronisk version, og især er onlineordbogen blevet populær. Navnlig er udviklingen dog gået virkelig stærkt de sidste 5-7 år, ikke mindst som følge af udbredelsen af smarttelefoner og tabletcomputere. Uanset hvordan årsagssammenhængen måtte være, vokser der en generation op som kommunikerer, læser og lærer langt det meste ved hjælp af en computer – især en lille, mobil én af slagsen. Det er derfor ikke overraskende at udviklingen også har påvirket ordbogsmarkedet.

Hvor korpusrevolutionen især havde stor effekt internt i det leksikografiske miljø, er den digitale revolution i høj grad noget der bemærkes af brugerne, ja, noget der påvirker alle. I papirordbøgernes tid var der en ret tæt forbindelse mellem indholdet i databasen og det færdige produkt. Fra redigeringen af DDO husker jeg således flere eksempler på at vi måtte vælge elementer i databasen efter hvordan de kunne gengives typografisk og ikke efter deres logiske indhold. Det trykte værk var hovedproduktet, databasen en mellemstation på vejen.

I dag er vi mere opmærksom på at databasen er det centrale element i arbejdet, og at basestrukturen skal være så tilpas veltilrettelagt og fleksibel at det kan lade sig gøre at publicere til forskellige medier og på forskellige platforme. Der er tale om en kæde af data, men i én retning: Redaktøren fylder indhold i databasen under redigeringen, mens databasens opbygning bestemmer hvad og hvordan der kan publiceres. Den eneste tovejskommunikation der foregår i processen, er redaktionens fælles møder hvor de træffer beslutninger om det redaktionelle arbejde. Om det kan gøres anderledes, tages op i afsnit 5.1 Interaktion og brugerinvolvering.

4.1. Overgang til digitale ordbøger og nye muligheder

De digitale medier giver nye muligheder for bedre ordbøger: lyd, billeder, videoer og anden form for multimedieanvendelse er noget de fleste redaktører gerne vil have med i ordbøgerne. Egentlig er det kun fantasien og pengepungen der sætter grænser, og det samme gælder muligheden for at linke internt mellem relevante dele af ordbogen eller eksternt til yderligere oplysninger på andre hjemmesider. Mange projekter befinder sig dog et andet sted fordi de er i en overgangsfase mellem papir og digital form. Det vil sige at ordbogsdata fra den trykte udgave først skal konverteres til en ny og mere hensigtsmæssig struktur før mediets muligheder kan udnyttes fuldt ud. Mange ordbøger bærer stadig, men i varierende grad præg af deres fortid som papirordbøger. Et godt eksempel er princippet om pladsøkonomi og den kondenserede stil som var karakteristisk for mange papirordbøger: Det gjaldt om at få mest mulig information ind på så lidt plads som muligt. Det princip er ophævet på skærmen og afløst af andre: Undgå at bruge forkortelser og symboler, men gør teksten mere læsevenlig; lad være med at sende brugeren på rundtur med henvisning efter henvisning, men giv klar besked på stedet.

Der består en risiko for at brugerne lider informationsdøden hvis de ikke kan finde den ønskede oplysning fordi den drukner mellem alle de mange oplysninger der nu er blevet plads til. Ti citater er ikke nødvendigvis bedre end to hvis de ekstra citater ikke tilfører nyt om ordet, og ti synonymer er ikke bedre end to hvis ekstramaterialet er langt fra den pågældende betydning.

Der er også en sammenhæng mellem pladsøkonomi og lemmaselektion. Hvis en ordbog har en fast ydre grænse i form af et bestemt antal sider og bind, bliver det vigtigt hvilke ord der

beskrives på den tildelte plads. Under redigeringen af DDO var det korpusbaserede derfor vigtigt, og hvert eneste ord der blev optaget i ordbogen, var et resultat af en nøje fastlagt procedure hvor korpusudbredelse var vigtigst. Men reelt set ved vi ikke om korpusprincippet gavner brugerne. Det er i overensstemmelse med den videnskabelige, deskriptive tradition, og bag den ligger en antagelse om at hvis man beskriver de 50.000 eller 100.000 mest udbredte ord, har man også nogenlunde dækket behovet for hvad folk har brug for at slå op. Som en kollega engang bemærkede: «Når man som DDO går korpusbaseret frem, er der grænser for hvor galt det kan gå». Det er en god, pragmatisk indstilling, og formentlig er den også rigtig.

Men i stedet kunne man også interessere sig for hvad brugerne efterspørger. *Nisseghetto, glokø, mimrepenge, kampsvede, twitterflirte, pizzahak, kvalmsk* er eksempler på ord som brugere af DDO har indsendt til den nyordsbase som Det Danske Sprog- og Litteraturselskab driver sammen med Dansk Sprognævn, og man kan spørge om det snarere er den slags ord der skal med i ordbogen. Efter korpusprincippet skal de næppe, for nogle er for lavfrekvente, og andre knytter sig til emner der er for snævert knyttet til et dagsaktuelt emne, og som derfor kan forventes at forsvinde uden at have etableret sig i sproget. Omvendt må man spørge: Gør det noget når de ikke tager pladsen fra et andet ord? Nej, det gør de ikke, og argumentet mod at tage dem med er derfor mere praktisk end teoretisk. Det tager tid at oprette og redigere artikler, og derfor bør redaktionen have principper for hvordan man vælger det næste ord. På DDO's redaktion spørger vi indimellem os selv om vi skal blive ved med strengt at følge det korpusbaserede princip. For at besvare det spørgsmål ordentligt er der nogle oplysninger vi har brug for: Er der ord i DDO som aldrig bliver slået op? Hvis ja, tilhører de da frekvensmæssigt en bestemt del af korpus? Er de ord som brugerne

søger forgæves efter, frekvente, lavfrekvente eller sjældne ord? Nogle ordbogssites overlader det mere eller mindre til brugerne at styre lemmaselektionen, i Danmark fx ordbogen.com. Hvis en søgning ikke fører til en artikel, bliver artiklen oprettet og skrevet (Bergenholtz & Nordahl 2012:221). Det fører flere spørgsmål med sig: Er det mere sandsynligt at et ord der er blevet søgt efter én gang, bliver søgt efter igen frem for et andet ord med en bestemt korpusfrekvens. Gør det evt. en forskel om der er blevet søgt efter ordet to, tre eller flere gange?

Man har altså brug for at vide mere om sammenhængen mellem det brugerne selv indberetter, det de slår op, og udbredelsen af ord i korpus hvis man skal sige noget fornuftigt om disse principper som metode til lemmaselektion. Det er noget vi har taget fat på at undersøge, men resultaterne af undersøgelsen når jeg ikke at komme ind på i denne artikel.

4.2. Brugertilpasning

Da næppe to brugere har samme forudsætninger og behov, og da det teknisk er muligt at præsentere samme indhold i forskellige visninger, er det nærliggende at tilbyde en brugertilpasset version af databasen for den enkelte bruger. I løbet af de seneste år er forskellige muligheder blevet foreslået og afprøvet. I DDO tilbyder vi en kort og en lang visning, andre ordbøger skelner mellem forskellige brugerorienterede situationer som man har set det i ordbøger udgivet af Center for Leksikografi ved Aarhus Universitet. Senest har Centret dog forsøgt sig med at udgive dele af basen som selvstændige produkter i tillæg til den samlede netordbog (Den Danske Netordbog): en særlig skrive-

ordbog, en synonymordbog, en betydningsordbog og en grammatik- og skriveordbog. Selv har jeg også ved flere lejligheder argumenteret for brugertilpasning (jf. Trap-Jensen 2010), men er i dag mere skeptisk end jeg har været. Det skyldes at vores egne erfaringer ikke er positive, og indtrykket støttes af kolleger der har forsøgt sig med det samme (se Verlinde & Binon 2010:1150). Forudsætningen for at brugertilpasningen kan fungere, er at brugerne gør en aktiv indsats ved at specificere hvad de har brug for, og det sker desværre ikke. Problemet er at brugerne enten ikke opdager muligheden, eller at de ikke kan finde ud af at analysere deres behov og vælge rationelt. Hvad grunden end er, viser vores brugerundersøgelser at de ikke benytter sig af muligheden. I den nuværende udformning tror jeg derfor ikke på en fremtid for brugertilpasning – selvom jeg stadig anser det for at være en indlysende rigtig fremgangsmåde.

Noget af det man kan forestille sig ville virke mere effektivt, er en mere personlig brugertilpasning baseret på den enkelte brugers søge- og navigeringshistorik. Man kender det fra e-handelssider hvor et firma «anbefaler» andre produkter fra deres sortiment som kunne være interessant for kunden. Det kan virke grænseoverskridende i starten, men jeg må konstatere at de ofte er ganske gode til at foreslå relevante produkter, herunder ting som man ikke kendte til i forvejen. Forslagene er blevet til ved at koble søge- og købshistorik med et antal kundeprofiler. Den samme tankegang kan formentlig godt overføres til en ordbogssituation hvor ordbogsvisningen tilpasser sig en bestemt brugerprofil på baggrund af hvad der tidligere er blevet søgt og navigeret efter. Uden at være ekspert på området forestiller jeg mig at det er nemmere at indføre funktionen for betalingsordbøger med en loginfunktion, men måske er en IP-adresse og en cookie tilstrækkeligt. Man kan også tænke sig at den personlige ordbog er tilknyttet den enkelte bruger og

udvikler sig enten som følge af brugerens adfærd eller i kombination med brugerens egne aktive specifikationer.

5. Fremtidens ordbøger

Hermed er vi fremme ved det der endnu ikke er realiseret, men hører fremtiden til. I den sammenhæng er det naturligt at spørge om ordbogen overhovedet behøver at være synlig som en selvstændig resurse som i dag. Det er der ikke nogen tvingende grund til, og personlig synes jeg ikke at tanken er skræmmende. Det er vel nærmest parallelt med situationen for papirordbogen i dag: Det er muligt at papirordbogen bliver her lidt endnu som et nicheprodukt for dem som foretrækker det, og sådan vil vi måske også have det med ordbogssites i fremtiden. Man kan konsultere en onlineordbog hvis man har lyst til at læse flere artikler i træk eller gå på opdagelse i ordbogen for fornøjelsens skyld, men ellers dukker ordbogen bare op dér hvor man har brug for den, i den situation som udløste behovet for at slå et ord op. Hvis man læser en tekst og møder et ord man ikke kender, er det unægtelig nemmere at få direkte besked ved fx at klikke på ordet og få forklaringen i et pop-op-vindue frem for at skulle gå hen et helt andet sted og slå ordet op dér – helst i den betydning ordet har i den aktuelle kontekst, men den løsning hører en endnu fjernere fremtid til.

5.1. Interaktion og brugerinvolvering

Et andet kendetegn ved moderne brugere er at de færdes på sociale medier, og ligesom dér vil de gerne deltage og bidrage, men har i øvrigt ikke stor tålmodighed. Flere ordbogssites forsøger at komme brugerne i møde på dette punkt og gør meget ud af de interaktive muligheder. Det er en måde at brande ordbogen på og dermed generere trafik på siden. Det sker også i et vist omfang på ordnet.dk. Brugerne kan fx indsende forslag til nye ordbogsartikler, og de kan få svar på spørgsmål om sitet og sproglige spørgsmål. En del af dem publiceres i rubrikken Sprogligt hvis de vurderes at have mere almindelig interesse. Endelig findes der på siden et lille anagramspil, Krasser, der benytter ordene fra DDO. Dette er blot eksempler på nogle af mulighederne. Andre ordbøger har en mere udbygget sproghjælp, laver særlige nyhedsbreve eller blogindlæg. Interaktionen kan også være overvejende i modsat retning så brugerne bliver mere direkte involverede. De kan fx få mulighed for at redigere hele artikler eller udvalgte elementer af dem, fx udtale, ækvivalenter eller uploade multimediefiler.

Wiktionary er vistnok det eneste større projekt i det nordiske område som anvender brugerredigering i stor stil. I Danmark vokser Wiktionary kun langsomt og har ikke samme succes og opmærksomhed som søsterprojektet Wikipedia. I Sverige er Wiktionary væsentlig større: 54.293 opslagsord i grundform oplyser de på deres hjemmeside (pr. 31. oktober 2013).

Når det gælder kvalitetsbedømmelsen af brugergenereret indhold, kan det være problematisk at slutte fra internationale tendenser til vores område. Udover kulturelt bestemte faktorer og traditionen i det enkelte land betyder befolkningens størrelse helt oplagt noget, og hvad der er en succes i eksempelvis det

engelsksprogede område, er det ikke nødvendigvis hvis det overføres til de i den målestok små lande som de nordiske. Antallet af aktive brugere/redaktører betyder formentlig noget for kvaliteten af arbejdet. En generel erfaring fra Wikipedia synes at være at kvaliteten af artiklerne vokser jo længere historik de har. Det betyder nemlig at der er foretaget flere ændringer og har været flere personer til at rette faktuelle fejl eller tvivlsomme formuleringer. Observationen forekommer overbevisende, men jeg kender dog ikke til konkrete undersøgelser der har bekræftet den.

Hvis det er svært at overføre erfaringer fra store sprog med mange brugere, kan man til en vis grad sige det samme om finansieringen. Blandt kolleger fra det engelsktalende område er det nærmest blevet en sandhed at brugerne ikke vil betale for ordbøger, og at gratis ordbøger med reklamefinansiering er den eneste vej frem – i kombination med betalingsapps til mobile enheder. Jeg er dog ikke sikker på at den sandhed også automatisk gælder situationen i Norden. Det gør en forskel om der er over en milliard potentielle øjne på reklamen, eller om de skal måles i tusinder. Det er ikke nogen naturlov at brugere ikke vil betale, men måske kommer det an på hvem man spørger. I Danmark har forlagsbranchen luftet de samme synspunkter og i øvrigt reageret på den vigende betalingsvilje ved at skære ned på redaktionerne. Men måske har de bare ikke været dygtige nok? Måske er synspunktet for letkøbt, men som udenforstående kan jeg konstatere at mens Politikens Forlag har nedlagt ordbogsafdelingen, og Gyldendal meget længe har ligget underdrejet med en minimal redaktion, har et it-firma, ordbogen.com, formået at sætte sig solidt på det kommercielle ordbogsmarked i Danmark, og forretningen går tilsyneladende godt.

Den økonomiske afmatning som har ramt Europa generelt, har sat sit præg på de offentlige kilders og fondenes finansie-

ringsvilje, dog med visse undtagelser. Ikke mindst er den norske økonomi generelt i bedre form end det øvrige Nordens. Selvom økonomisk krise vanskeliggør finansiering, er det omvendt også paradoksalt, for digitale ordbøger er på mange måder et udtryk for effektivisering. Med en løbende digital vedligeholdelse undgår man at gentage unødigt dobbeltarbejde når ordbøgerne forældes og skal erstattes af helt nye projekter, måske allerede efter en generation sådan som man har oplevet det med de trykte ordbøger. For brugerne er det en klar fordel at indholdet altid er opdateret og aktuelt; for bevillingsgiverne er det en fordel at de får mere ordbog for de samme penge. Men forskydningen fra afsluttet projekt til løbende drift kræver en mental erkendelse og omstilling hos bevillingsgivere, politikere og embedsmænd som måske endnu ikke har indfundet sig. Det betyder givetvis noget at der er mindre prestige forbundet med løbende drift end ved at søsætte nye og spændende projekter.

Når man skal vurdere den udvikling som vi befinder os midt i, er det uklogt at udtale sig alt for håndfast om hvor vi er på vej hen. Jeg tror dog at vi denne gang, i modsætning til hvad der skete som følge af korpusrevolutionen, kommer til at opleve at ordbøgerne ændrer sig meget. I den forstand er der mere revolution over det der er sket de sidste 5-6 år, men det er overvejende en teknisk omvæltning. En sandsynlig udvikling kunne være at ordbøger i den form vi kender dem, med tiden udvikler sig til at blive et nicheprodukt som de fleste mennesker måske slet ikke kommer til at møde i deres hverdag. Til gengæld vil de være integreret i mange flere forskellige stykker software rundt og dukke op når brugeren har brug for forklaringer, hjælp til skrivning eller oversættelse. Det betyder at der bliver større afstand mellem det indhold som leksikografen redigerer, og den visning som slutbrugeren kommer til at se. For leksikograferne kan det betyde at de ikke får noget med visningen at gøre, men alene

skal koncentrere sig om indholdet. Selvfølgelig kan man vælge at bevare en ordbogshjemmeside som leksikografer og ligesindede kan konsultere for fornøjelsens skyld – men altså som et udpræget nicheprodukt.

Til gengæld skal indholdet kunne udveksles med en lang række eksterne rekvirenter: med it-firmaer, med producenter af aviser, e-bogslæsere og hjemmesider samt andre udviklere og serviceleverandører som har brug for at kunne tilgå indholdet og kommunikere med de serversystemer hvor ordbogsindholdet ligger. Udvekslingen af data må derfor i højere grad forventes at komme til at ske ved hjælp af webservices og API'er til de eksterne rekvirenter. På den måde undgår dataindehaveren at udlevere fysiske data, mens rekvirenten får adgang til data, typisk via internettet, og kan i øvrigt tilpasse ordbogsindholdet efter deres formål, dvs. hvordan slutbrugeren ser indholdet, afhænger af de tilknyttede stylesheets.

Endelig bliver leksikografiske data i stigende grad interessante for andre end dem der laver ordbøger. Der er en stigende efterspørgsel efter leksikografiske data til andre typer af sprogteknologiske produkter, fx synonymlister, fuldformsleksika og domæneinventarlister. Det er ting som indtil nu mest er blevet opfattet som spin-off af det leksikografiske arbejde med ordbøgerne, men det er sandsynligt at det bliver en mere central del af den leksikografiske aktivitet fremover. Der er med andre ord ikke tegn på at efterspørgslen efter leksikografiske data bliver mindre, og leksikografer kan derfor se fremtiden i møde med fortrøstning.

Litteratur

Ordbøger

Dahl, B.T. & H. Hammer (1907–14): *Dansk ordbog for Folket I–II.* København og Kristiania: Gyldendalske Boghandel, Nordisk Forlag.
DanNet, <http://wordnet.dk>
DDO = *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab. <http://ordnet.dk/ddo> (oktober 2013).
Den Danske Netordbog, <http://www.ordbogen.com>.
Kristiansen, V. (1866): *Bidrag til en Ordbog over Gadesproget og saakaldt Daglig Tale.* Kjøbenhavn: Boghandler H. Hagerups Forlag.
Molbech, Christian (1859): *Molbechs ordbog 1-2* (2. udg.). Kiøbenhavn: Gyldendalske Boghandlings Forlag.
ODS = *Ordbog over det danske Sprog* (1918–56): København: Det Danske Sprog- og Litteraturselskab og Gyldendal. <http://ordnet.dk/ods> (oktober 2013).
OED = *Oxford English Dictionary*, <www.oed.com>.
SAOB (1898–) = *Ordbok öv*er *svenska språket utgiven av Svenska Akademien* (Svenska Akademiens ordbok). Lund: Gleerups förlag.
Videnskabernes Selskabs Ordbog 1–8 (1793–1905). København.

Anden litteratur

Atkins, Beryl T.S. (2010): The DANTE Database: Its Contribu-

tion to English Lexical Research, and in Particular to Complementing the FrameNet Data. I: G.-M. de Schryver (ed.): *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hank*. Kampala: Menha Publishers, 267-297.

Bergenholtz, Henning & Bjarni Norddahl (2012): Ordbogsartikler, som ingen læser. I: *LexicoNordica 19*, 207-223.

Cook, Paul, Jey Han Lau, Michael Rundell, Diana McCarthy & Timothy Baldwin (2013): A lexicographic appraisal of an automatic approach for detecting new word senses. I: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.): *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, 49-65.

Dahlerup, Verner (1907): Principer for ordbogsarbejde. I: *Danske Studier 1907*, 65-78.

Didakowski, Jörg, Lothar Lemnitzer & Alexander Geyken (2012): Automatic example sentence extraction for a contemporary German dictionary. I: Ruth Vatvedt Fjeld & Julie Matilde Torjusen (eds.): *Proceedings of the 15th EURALEX international Congress 7–11 August, 2012, Oslo*. Oslo: University of Oslo, 343-349.

Fillmore, Charles J. (1995): The Hard Road From Verbs To Nouns. I: M. Chen & O. Tzeng (eds.): *In honor of William S-Y. Wang*. Taipei, Taiwan: Pyramid press, 105-129.

Halskov, Jacob (2010): Halvautomatisk udvælgelse af lemmakandidater til en nyordsordbog. I: *LexicoNordica 17*, 73-98.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý & Vít Suchomel (2013): The TenTen Corpus Family. I: *Lancaster, 7th International Corpus Linguistics*

Conference CL 2013, 125-127.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz & David Tugwell (2004): The Sketch Engine. I: Geoffrey Williams & Sandra Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*. Lorient: Université de Bretagne-Sud, 105-115.

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell & Pavel Rychlý (2008): GDEX: Automatically Finding Good Dictionary Examples in a Corpus. I: Elisenda Bernal & Janet DeCesaris (eds.): *Proceedings of the XIII EURALEX International Congress.* Barcelona: Universitat Pompeu Fabra, 425-433.

Langebek, Jacob (ca. 1738): Plan for Rostgaards ordbog. Manuskript. Her citeret fra: *Det Kongelige Danske Videnskabernes Selskab 1742-1942. Samlinger til Selskabets Historie*, v. Asger Lomholt, III, 1960, 219, med henvisning til *Nye Danske Magazin V. 4*, 1827, 271-76.

Lorentzen, Henrik, Sanni Nimb & Thomas Troelsgård <denne udgivelse>: Fra Begreb til Ord.

Lorentzen, Henrik (1999): Jagten på det gode citat. Om vanskelighederne ved at finde egnede ordbogseksempler i et korpus. I: Martin Gellerstam, Kristinn Jóhannesson, Bo Ralph & Lena Rogström (udg.): *Nordiska studier i lexikografi 5. Rapport från Konference om lexikografi i Norden, Göteborg 26–29 maj 1999*. Göteborg : Nordiska föreningen för lexikografi, 202-216.

Rydstedt, Rudolf (1988): Creating a Lexical Database from a Dictionary. I: *Studies in Computer-Aided Lexicology*. Göteborg, 228-267.

Trap-Jensen, Lars (2010): One, Two, Many: Customization and User Profiles in Internet Dictionaries. I: Anne Dykstra & Tanneke Schoonheim (eds.): *Proceedings of the XIV*

Euralex International Congress (Leeuwarden, 6-10 July 2010), Ljouwert: Fryske Akademy, 1133-1141.

Verlinde, Serge & Jean Binon (2010): Monitoring Dictionary Use in the Electronic Age. I: Anne Dykstra & Tanneke Schoonheim (eds.): *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6-10 July 2010)*, Ljouwert: Fryske Akademy, 1144-1151.

Lars Trap-Jensen
ledende redaktør
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
ltj@dsl.dk