Et net af ord – *ordnet.dk*

»Kongen havde besluttet, førend man skred til det Yderste, at forsøge alle lemfældige Midler«. Sådan skriver forfatteren Carsten Hauch i romanen »Guldmageren« fra 1836. Hans brug af ordet 'lemfældig' kan nok få en og anden til at rynke panden og gribe til ordbogen efter en forklaring. Ifølge Ordbog over det danske Sprog betyder ordet:

> omhyggelig for ikke at anrette (større) ødelæggelser, skade, gaa for haardt frem mod en olgn.; som tager forsigtigt, blidt ell. lempeligt, mildt paa noget«

Forklaringen vil sikkert undre en del moderne brugere. I en ny ordbog som Den Danske Ordbog markeres denne betydning da også som sjælden, og i stedet gives som den første betydning: »ikke omhyggelig nok; skødesløs, sjusket«.

Det må ikke forstås sådan at den ene ordbog dermed har ret, mens den anden tager fejl. Tværtimod. For sproget ændrer sig som bekendt med tiden: Nye ord kommer til, gamle går af brug, og eksisterende ord skifter betydning. Hvilken ordbog man skal benytte, afhænger i høj grad af den tekst hvori det pågældende ord optræder, og hvornår den er skrevet. Derfor er begge de nævnte ordbogsværker nødvendige. Og derfor arbejdes der nu på at give adgang til artiklerne i begge værker i et integreret elektronisk opslagsværk på internettet, *ordnet.dk*.

Betegnelsen net er naturligvis ingen tilfældigt valgt metafor for World Wide Web – internettet. Det består som bekendt af et enormt antal hjemmesider knyttet sammen af et virvar af forgreninger og indbyrdes forbindelser – links. Og ligesom ederkoppens net har det ikke nogen på forhånd fastsat størrelse; det er fleksibelt, lader sig strække og kan udvides med stadig flere tråde når der er behov for det.

På tilsvarende måde kan metaforen bruges om sprogets byggeklodser: Også sproget består af et enormt antal enheder, nemlig ord og udtryk, som står i forbindelse med hinanden – grammatisk og leksikalsk. Så når man vil beskrive det danske sprog som et *ordnet* system af enheder og forbindelser i et opslagsværk på internettet, ja, så har man – et *ord-net*.

Der er flere forskelle på at publicere elektronisk og til papirmediet. Én meget vigtig forskel er at plads så godt som ingen rolle spiller i elektroniske publikationer. Hvis det fx drejer sig om en onlinetjeneste på internettet, skal udbyderen måske have lidt flere bytes på sin server og også være omhyggelig med at sikre at søgehastigheden ikke forringes hvis den underliggende base er meget stor, men det er som regel et problem der er til at løse både teknisk og økonomisk. For brugeren, derimod, gør det ingen forskel om basen indeholder tusind poster eller to millioner: Enten finder man det man søgte, eller også gør man ikke.

En triviel konstatering, måske, men én der har store konsekvenser for hvordan man laver ordbøger og andre opslagsværker til nettet. Ordbogsredaktører er fra tidernes morgen blevet oplært til at økonomisere med den sparsomme plads og har altid nøje måttet afveje detaljeringsgraden i beskrivelsen med hensynet til hvor meget oplysningerne fylder. Men sådan er det ikke længere når man publicerer elektronisk. Her er altid plads til én til: flere opslagsord, flere citater, flere betydninger – ja, mere af hvad som helst.

Når man dernæst betænker at mediet gør det muligt at afsøge flere baser samtidig, sådan som det sker i webportaler eller vha. søgemaskiner, er det forståeligt at man med mellemrum har kunnet høre forskellige – visionære eller vidtløftige, afhængigt af øjnene der ser – forslag om at oprette ét stort, fælles site hvorfra man kan søge og få viden om hvad som helst der er ordnet systematisk i artikelform på dansk. Hvis man fx vil slå ordet 'kartoffel' op, skal man kunne få en ordbogsforklaring på ordets betydning og brug i sproget; en leksikonartikel beskriver kartoflens biologiske egenskaber, dens kultur- og dyrkningshistoriske udvikling og giver oplysninger om sygdomme, produktion og anvendelsesmuligheder, mens man i en kogebog kan finde opskrifter på retter med kartofler. For andre ord kunne det være relevant med artikler fra manualer, lovsamlinger, atlas, *you name it* – det er altså intet mindre end den virtuelle grønspættebog man forestiller sig.

Tanken er fascinerende og måske slet ikke så langt ude som det umiddelbart kan lyde. Netop fordi netpublicering har den styrke at et værk kan vokse gradvis ligesom spindelvævet der vokser i takt med at stadig flere tråde spindes. Man kan godt begynde med lidt og så fylde mere på siden hen. I Det Danske Sprog- og Litteraturselskab (DSL) har vi gjort os nogle lignende tanker, i et mere beskedent perspektiv ganske vist, men alligevel. DSL har i tidens løb været involveret i et stort antal udgivelser, dels inden for den klassiske danske litteratur, dels i beskrivelser af dansk sprog i form af ordbøger. I de senere år har DSL også beskæftiget sig med elektronisk udgivelse, specielt med udarbejdelsen af Korpus 2000, en tekstsamling hvor man kan slå ord og udtryk op og som resultat få ordnede lister med eksempler taget fra autentiske sprogprøver samt foretage forskellige sproglige undersøgelser.

DSL ønsker at gøre flere af sine udgivelser tilgængelige for offentligheden på internettet, og netop sådan at der er forbindelse mellem dem. Man skal kunne slå et ord op og få en forklaring på ordets betydning og brug

sådan som det er i dag, eller som det var på et bestemt historisk tidspunkt hvis det er en ældre tekst man sidder med. Fra ordbogsartiklen skal man kunne bevæge sig videre til en korpusdel hvis man ønsker at foretage yderligere undersøgelser eller finde flere belæg på ordet i tekster fra den periode man er interesseret i. Og omvendt hvis man som litterært interesseret er i færd med at læse en tekstudgivelse, skal man kunne klikke sig til en ordbogsartikel på de ord i teksten som man gerne vil have forklaret. Den endelige vision er at have en base med ordbogsartikler der beskriver hele sproget fra runedansk til nutiden, med korpusværktøj og teksteksempler fra hele den sproghistoriske periode og med mulighed for at bevæge sig frit mellem de enkelte dele.

Men som sagt, man kan sagtens begynde med noget mindre end det fuldt udbyggede system. Takket være midler fra Carlsbergfondet og Kulturministeriet og med støtte fra Det Elektroniske Forskningsbibliotek er DSL nu gået i gang med at spinde de første tråde ved at udvikle et elektronisk opslagsværk som i første omgang skal samle tre af selskabets udgivelser og gøre dem tilgængelige fra én hjemmeside. Det man i første omgang skal få adgang til når man klikker sig ind på siden www.ordnet.dk, er data fra Den Danske Ordbog og Korpus 2000 samt en digital version af Ordbog over det danske Sprog (inklusive supplement). Hermed vil brugeren få adgang til en fyldig beskrivelse af ordforrådet i dansk fra omkring 1700 til i dag, samtidig med at det bliver muligt at få supplerende oplysninger og foretage egne sproglige undersøgelser af det moderne sprog. Værket skal desuden give mulighed for at foretage andre typer af undersøgelser i ordbogsmaterialet end det der kendes fra papirordbøgerne. I det følgende præsenteres projektets enkelte dele og de søgemuligheder man kan tænke sig i det elektroniske opslagsværk.

Ordbog over det danske Sprog i elektronisk form

Det femte og sidste supplementsbind til Ordbog over det danske Sprog (ODS) er netop færdiggjort og udsendes i begyndelsen af 2005. Dermed er sidste punktum sat for et monument i dansk ordbogshistorie, men heldigvis slutter historien ikke her. Med ordnetprojektet bliver der også en plads til ODS i den digitale tidsalder, og en vigtig del af projektet bliver at digitalisere det gamle papirværk, flette det sammen med supplementsbindene til ét samlet værk og gøre det søgbart via ordnettet. Dermed går et længe næret ønske i opfyldelse, idet DSL gennem snart mange år har haft planer om at få digitaliseret vores store nationale ordbog på samme måde som det allerede er sket for ODS' søsterordbøger inden for det germanske sprogområde, *Oxford English Dictionary* i England, *Svenska Akademiens Ordbok* i Sverige, *Het Woordenboek der Nederlandsche Taal* i Holland og brødrene Grimms *Deutsches Wörterbuch* i Tyskland. Det har selvfølgelig den fordel at DSL kan lære af de dyrekøbte erfaringer som pionererne på området har indhøstet, og anvende en gennemprøvet model. For eksempel har der ikke været den store tvivl om at indtastning langt er at foretrække frem for OCR-scanning. Vi har valgt den model som blev fulgt af den senest digitaliserede af de ovennævnte ordbøger, den tyske *Deutsches Wörterbuch*, og samarbejder med den institution ved universitetet i Trier som var ansvarlig for det tyske projekt. Selve indtastningen udføres af et firma i Kina som har specialiseret sig i netop denne type opgaver. Ud over at kinesisk arbejdskraft er billigere end dansk, har det vist sig at det faktisk er en fordel at indtasterne ikke selv behersker det sprog de skal kopiere. En vigtig pointe er desuden at kineserne indtaster manuskriptet i to uafhængige versioner, hvorefter de to versioner sammenlignes automatisk ved en efterbehandling i Trier. Denne metode sikrer en så lav fejlprocent (der estimeres 99,999 % korrekthed, eller 1 fejl pr. 100.000 typeenheder) at den manuelle korrekturlæsning kan undværes, og det er især denne proces der er resursekrævende hvis man vælger scanneløsningen.

Efter den første fase, rådigitaliseringen, overtager DSL de typografisk opmærkede filer og kan påbegynde den indholdsmæssige del af projektet: opgaven med at fortolke de typografiske oplysninger og omsætte dem til artikelstrukturens delelementer. Det er nødvendigt for at kunne lagre ordbogsdataene i en digital struktur der muliggør tværgående søgninger i ordbogens oplysninger, af typen »hvilke ord i dansk er indlånt fra arabisk?« eller »find alle de citater der har H. C. Andersen som kilde«. Den anden store opgave for DSL bliver at flette det oprindelige ODS-manuskript sammen med ODS-supplementet så alle de senere tilføjelser og ændringer indgår på rette sted i strukturen. Supplementet selv er naturligvis blevet udarbejdet elektronisk og skal altså ikke først digitaliseres, men sammenfletningen udgør ikke desto mindre en betydelig udfordring som man ikke skal undervurdere.

Den Danske Ordbog og Korpus 2000

Korpus 2000 findes som en eksisterende hjemmeside; Den Danske Ordbog er blevet udarbejdet elektronisk, og derfor er der for disse to projekters vedkommende ikke det samme behov for en omfattende forbehandling som i ODS' tilfælde. Men det betyder ikke at der er tale om uforanderlige størrelser. Tværtimod er det som nævnt internettets natur og store styrke at det er dynamisk og hele tiden ændrer sig. Sådan bør det også være med ordbogs- og korpusdata. Ordnettet skal løbende forsynes med nye artikler og tekster så det vedligeholdes og holder sig ajour med den sproglige udvikling. Og når plads ikke spiller nogen nævneværdig rolle i elektronisk publicering, er det ikke nødvendigvis alene den nyeste del af ordstoffet der skal suppleres. En ordbog der beskriver almenordforrådet, indeholder i sagens natur en stor mængde ord som mo-

dersmålstalende udmærket kender i forvejen og derfor ikke vil have brug for at slå op. Ordbogen kan derfor med nogen ret bebrejdes at den stopper netop der hvor de fleste har brug for at kunne slå op, nemlig ved de ord man ikke kender, fx fordi de er sjældne eller tilhører et specielt fagområde. Det kan man råde bod på i et elektronisk værk hvor pladsen ikke er afgørende. Dog er det vigtigt at artiklen så forsynes med en markør der viser at ordet er sjældent.

For korpusdelen gælder det ligeledes at det er vigtigt at supplere med tekster løbende og ikke nødvendigvis alene nye tekster, men også gerne bagud i tid for at tilvejebringe et effektivt redskab til at undersøge sprogets udvikling over tid. Selvom det ikke indgår i projektets første fase, er det et ønske der prioriteres højt på længere sigt. Et oplagt materiale ligger allerede tilgængeligt i Arkiv for Dansk Litteratur (www.adl.dk), en samling skønlitterære tekster af klassiske, danske forfattere, udarbejdet i et samarbejde mellem DSL og Det Kongelige Bibliotek, og velegnet som kilde til yderligere belæg for en overvejende litterært baseret ordbog som ODS.

Hjemmeside og søgemuligheder

Det er ikke blot sproget og ordene der hele tiden ændrer sig. Det gør også teknikken og dermed de muligheder den betinger. I *ordnet.dk* vil vi bestræbe os på at udnytte mediets muligheder så man ikke blot får adgang til at se den trykte bogs sider på en computerskærm. Det har allerede været nævnt at det digitale format giver mulighed for tværgående søgninger i artiklernes oplysningstyper. Den slags avancerede søgninger skal det være muligt at udføre ved hjælp af ordnettet, og det samme gælder naturligvis fritekstsøgning og søgning med jokertegn og forskellige logiske operatorer.

Sammenkoblingen med korpustekster skal også udnyttes, ikke blot som en generel facilitet hvor man med et klik kan skifte mellem ordbogs- og korpussøgning, men gerne i mere integreret form så man kan få yderligere oplysning netop på det sted i en artikel man befinder sig. Det kan være forespørgsler som disse: »Hvordan er fordelingen mellem alternative bøjningsformer af et bestemt ord?« (fx når man befinder sig i bøjningsdelen til artiklen kursus – bør man skrive *kurset*, *kursusset* eller *kursuset*?), »hvilke ord bruges typisk som subjekt og objekt for 'sluge'?" (når man i artiklen for verbet 'sluge' læser konstruktionsoplysningen NGN sluger NGT) eller »giv mig flere citateksempler med dette ord end der er nævnt i artiklen«.

Resultatet af disse forespørgsler skal ikke redigeres af ordbogens redaktører, men fremkomme som resultat af en automatisk undersøgelse der foretages på stedet. Resultatets kvalitet hænger derfor sammen med hvilke typer undersøgelser det er muligt at udføre. De nævnte spørgsmål stiller nemlig forskellige krav til hvordan de bagvedliggende korpustekster er

opmærkede. For at få et pålideligt svar på fordelingen af bøjningsformer kræves det at teksterne er blevet opmærket morfologisk (eller *tagget*, som det hedder i fagsproget) så systemet genkender forskellige bøjningsformer som hørende til den samme grundform. Spørgsmålet om sætningsled kræver at teksterne også er blevet opmærket syntaktisk (eller *parset*, som det også kaldes). Endelig kræver det sidste spørgsmål at korpus er betydningsopmærket – hvis man vel at mærke er interesseret i eksempler på en helt bestemt betydning af et ord med flere betydninger. DSL's korpora er i deres nuværende form opmærket både morfologisk og syntaktisk, men ikke betydningsmæssigt; så vidt er sprogteknologien endnu ikke kommet.

En anden facilitet der står højt på ønskelisten, er at få ordene udtalt af en stemme. Det er en af de helt oplagte fordele ved mediet og vil være et stort fremskridt for de mange mennesker som har svært ved at afkode lydskriftens specielle tegn nøjagtigt. Men også det er til dels et spørgsmål om den rette teknologi på det rette tidspunkt.

Endelig satser vi på også at kunne tilbyde forskellige former for begrebsorienterede søgninger. Det er jo sådan at der i forvejen, direkte eller indirekte, er indeholdt mange oplysninger om ordenes indbyrdes forbindelser i en ordbog. Der oplyses om synonymer, antonymer og nærtbeslægtede ord. Mange definitioner er udformet sådan at de udpeger det nærmeste overbegreb og specificerer det med særlige kendetegn: En kasket er en hovedbeklædning med skygge foran og flad eller hvælvet puld. Der findes andre former for hovedbeklædning som brugeren kunne være interesseret i at se. Og de findes jo i ordbogen i forvejen; ordene er blot arrangeret sådan at man ikke kan finde dem hvis man ikke kender dem på forhånd. Betydningsordbøger er opbygget *semasiologisk*: Vi kender ordet, men ønsker oplysning om dets betydning. Den type søgning vi her taler om, er den modsatte: Vi kender betydningen (hovedbeklædning) og ønsker at få oplyst de forskellige betegnelser for den. Sådan en ordbog kaldes *onomasiologisk*.

Normalt finder man den slags oplysninger i en tesaurus, en begrebsordbog eller i et krydsogtværsleksikon, men da oplysningerne som sagt allerede er indeholdt i ordbasen, er det nærliggende at gøre dem søgbare ved hjælp af en særlig søgefunktion. Det kræver ganske vist en ikke ubetydelig redaktionel bearbejdning at beskrive det net af betydningsforbindelser der er mellem sprogets ord, i en form så det bliver elektronisk søgbart. Derfor indleder DSL i 2005 et samarbejde med Center for Sprogteknologi ved Københavns Universitet om at skabe et sådant sprogteknologisk betydningsnet (projektet DanNet, finaniseret af Statens Humanistiske Forskningsråd) på grundlag af Den Danske Ordbogs artikler. Resultatet vil – foruden anvendelsen til begrebsorienterede søgninger i ordnettet – også kunne benyttes mere generelt

inden for forskellige former for intelligent informationshåndtering i sprogteknologiske it-systemer, fx til indholdsresuméer af tekster eller til begrebs- og emnesøgninger med søgemaskiner i stedet for søgning alene på tekststrenge.

Tidsramme

De her nævnte muligheder skal tages med de forbehold der knytter sig til et projekt i en indledende fase. De er udtryk for hvad vi anser for både vigtigt og realistisk at gennemføre, men vi kender naturligvis ikke fuldt ud omfanget af de udfordringer der vil opstå undervejs. Efter planen vil de første tråde til ordnettet allerede kunne ses til foråret 2005, hvor DSL offentliggør en ny og forbedret version af Korpus 2000. Og når de bærende tråde for alvor er spundet i løbet af et par år, måske allerede før, vil selve ordnettet kunne lanceres i en version hvor der vil være fuld adgang til samtlige ODS-artikler, i første omgang dog sandsynligvis med reduceret søgefunktionalitet og uden at sammenfletningen med supplementsartiklerne er gennemført. Derefter vil stadig flere tråde gradvis komme til indtil ordnettet med udgangen af 2009 skal være udbygget og funktionsdygtigt. For en god ordens skyld skal jeg understrege at tjenesten stilles gratis til rådighed i projektperioden. Om det også kan ske efter 2009, ligger ikke fast på nuværende tidspunkt.

Lars Trap-Jensen (født 1960)
ledende redaktør
Det Danske Sprog- og Litteraturselskab