IV

ORDBØGERNE OG INTERNETTET

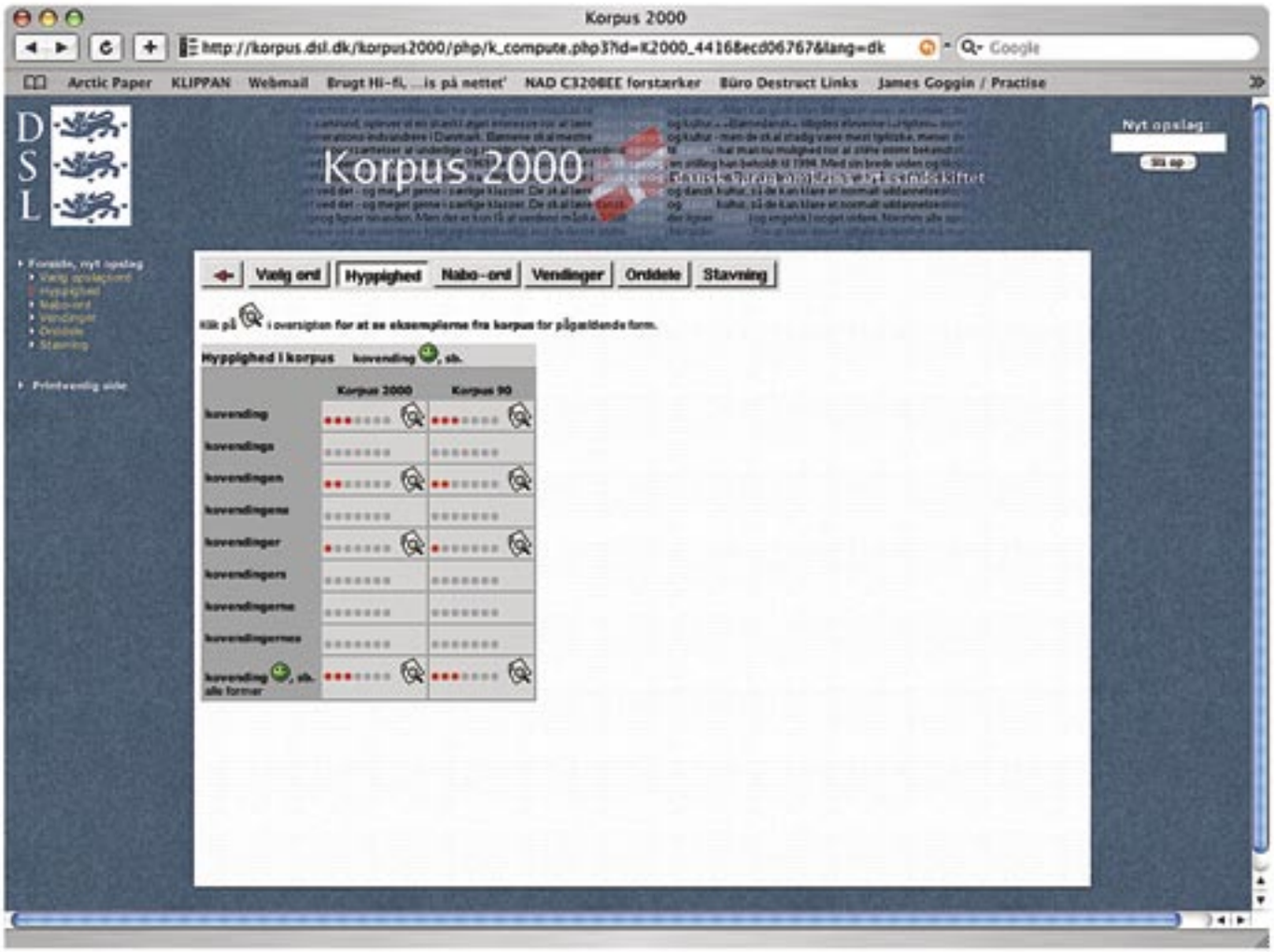
– BETRAGTNINGER OVER ORDNET.DK

AF CAND.MAG., MPHIL LARS TRAP-JENSEN
LEDENDE REDAKTØR,
DET DANSKE SPROG- OG LITTERATURSELSKAB

På mange måder er ordbogen og de digitale medier som skabt for hinanden. Computerens styrke ligger i at den er i stand til at foretage et meget stort antal beregninger i løbet af ganske kort tid – forudsat at data foreligger i et format som computeren genkender og kan bruge som udgangspunkt for sine operationer. Ordbogen er på sin side kendetegnet ved at indeholde et meget stort antal enheder som er ordnet systematisk, både i forhold til ordbogens andre artikler (alfabetisk) og internt i de forskellige specialiserede oplysninger om udtale, bøjning, sproghistorie, betydning og brug – netop den form for konsekvent systematik som computeren er velegnet til at bearbejde.

Dertil kommer den åbenlyse fordel at plads så godt som ingen rolle spiller ved elektronisk publicering. Et storværk som Den Store Danske Encyklopædi kan mageligt ligge på en enkelt cd-rom; onlineværker kræver blot at udbyderen har lidt flere bytes på sin server. Kapaciteten er ikke noget reelt problem i dag hvor en gigabyte – der svarer til flere gange Encyklopædiens datamængde – kan erhverves for en slik. For brugeren, derimod, gør det ingen forskel om værket indeholder to tusind eller to millioner artikler: Enten finder man det man søger, eller også gør man ikke. Og derfor er det heller ikke svært at indse at fremtidens elektroniske ordbøger kommer til at se anderledes ud end de papirordbøger vi kender i dag. Den strenge pladsøkonomi som kendetegner traditionelle ordbøger, er med ét slag overflødiggjort af de digitale medier. På skærmen er der altid plads til én til: Flere opslagsord, flere betydninger, flere eksempler – for slet ikke at tale om de elektroniske mediers mulighed for at bringe både billeder og lyd.

Danmark er i sammenligning med vore nærmeste naboer i Skandinavien, Tyskland og England sent ude med at tilbyde de store nationale opslagsværker i elektronisk form. Så meget mere glædeligt er det at det har været muligt at fortsætte en lang tradition for samarbejde mellem Det Danske Sprog- og Litteraturselskab (herefter DSL) og Carlsbergfondet og Kulturministeriet til også at omfatte elektroniske opslagsværker. Takket være en bevilling fra de to sidstnævnte, og med støtte fra Det Elektroniske Forskningsbibliotek, har DSL iværksat projektet *ordnet.dk*, et sprogligt opslagsværk som skal give samtidig adgang til de store nationalordbøger *Ordbog over det danske Sprog* samt de netop afsluttede *Supplement til Ordbog over det danske Sprog* og *Den Danske Ordbog*. Ordbøgerne kobles sammen med den allerede eksisterende hjemmeside *Korpus 2000*, en stor samling tekster der er opmærket så man kan



udføre forskellige sproglige undersøgelser i materialet. Projektet gennemføres over en seksårig periode, og de første synlige resultater er nu begyndt at vise sig når man klikker sig ind på www.ordnet.dk.

Den grundlæggende idé med *ordnet.dk* er at hjemmesiden skal være noget andet og mere end elektroniske versioner af papirudgaverne. Det betyder at ordbogsdataene skal konverteres til et format som åbner for andre former for søgninger af den slags som computeren er så god til. Dette er også grunden til at der foruden ordbogsværkerne desuden skal være mulighed for at lave undersøgelser i tekster fra samme periode som den artiklerne beskriver. Præcis hvilke muligheder der realiseres, ligger i skrivende stund ikke hundrede procent fast – sådan er vilkårene når man laver pionerarbejde. Ikke desto mindre vil jeg i det følgende forsøge at løfte sløret for nogle af de planer DSL arbejder med.

FIGUR 1.
Eksempel på opslag i Korpus 2000.

»ORDBOG OVER DET DANSKE SPROG« PÅ NETTET

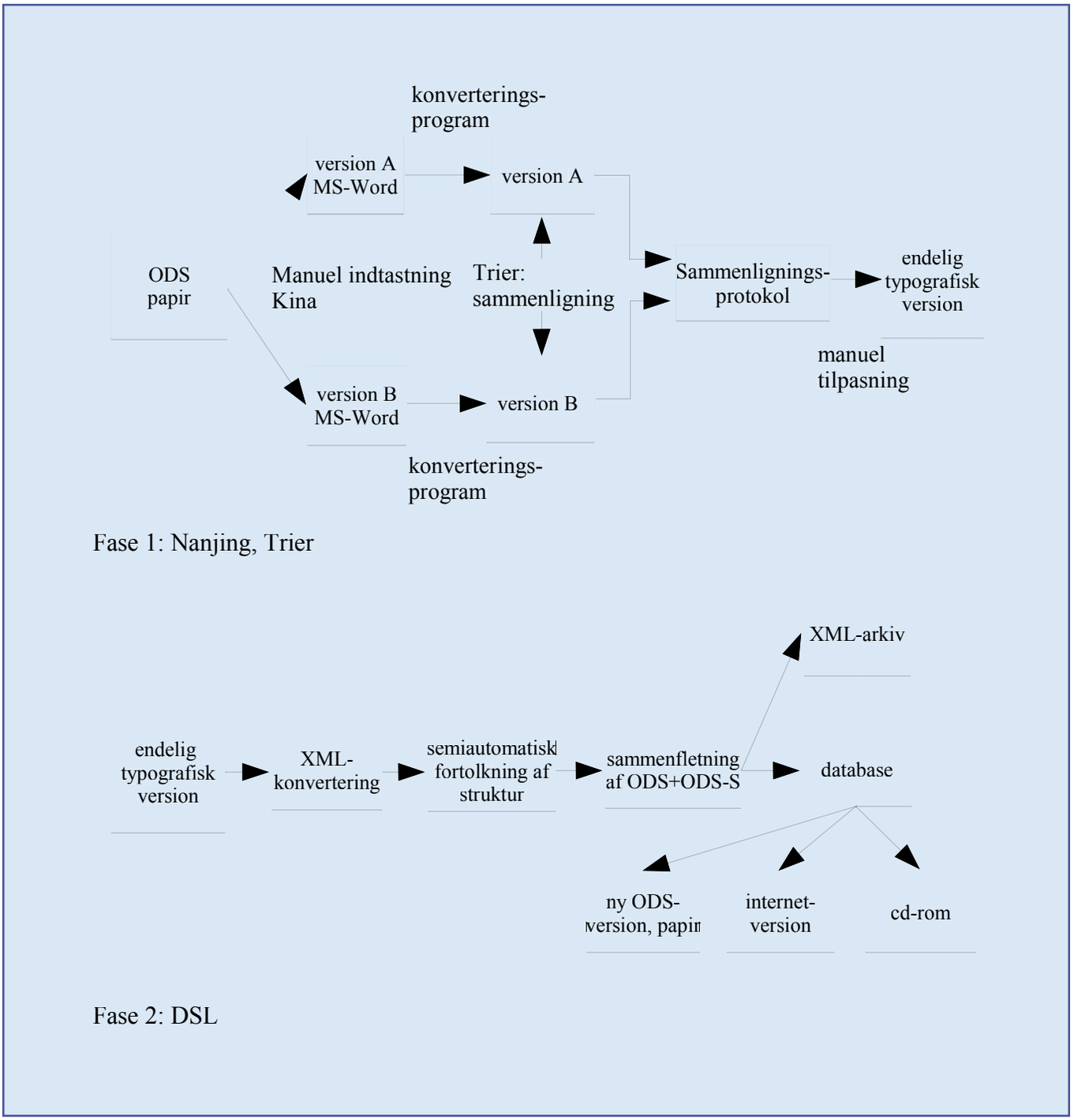
Det første synlige resultat af projekt *ordnet.dk* kom i november 2005 da *Ordbog over det danske Sprog* (herefter ODS) for første gang kunne ses på nettet. Forud for lanceringen var gået en digitaliseringsproces som blev udført i samarbejde med Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften ved Universitetet i Trier, den afdeling der også har stået for digitaliseringen af den store tyske nationalordbog, *Deutsches Wörterbuch* af brødrene Jacob og Wilhelm Grimm. Vi valgte altså at følge samme model som det tyske projekt, og processens forskellige faser er fremstillet skematisk i FIGUR 2. Først blev hele manuskriptet indtastet i to uafhængige versioner hos et firma i Nanjing, Kina, ud fra de detaljerede anvisninger vi havde udarbejdet og omsat i en indtastningsmanual. De to versioner blev herefter sammenlignet ved en automatisk proces i Trier, og uoverensstemmelserne opsamlet i en protokol som blev gennemgået dels semiautomatisk, dels manuelt. Herefter blev den færdige version af rådata overgivet til DSL til videre bearbejdning. Den videre bearbejdning består dels i at omsætte filerne til et standardformat (XML og databaseformat), dels i at fortolke de typografiske oplysninger strukturelt så ordbogsartiklernes forskellige oplysningstyper kan opmærkes og derved gøres søgbare. Den version der blev lanceret i november, er ikke helt fri for tekniske fejl og har desuden en ganske grov opmærkning. Det er derfor kun muligt at søge og få vist artikler i deres helhed. Ud over minimering af de tekniske fejl vil de senere versioner gradvis blive finere opmærket så det bliver muligt at foretage mere specifikke opslag. Det kan give svar på spørgsmål som: "Hvilke ord har vi indlånt fra arabisk i det 19. århundrede?", "Hvilke adjektiver optræder i citater af Herman Bang og J.P. Jacobsen?" eller "Hvordan afspejler artikler og citatmateriale synet på jøderne i mellemkrigstiden?". Foruden opgaven med den strukturelle fortolkning af manuskriptet tilbagestår også den ikke ubetydelige opgave at sammenflette det oprindelige værk med *Supplementet til ODS*, hvis femte og sidste bind udkom i efteråret 2005.

At det overhovedet har kunnet lade sig gøre at have en elektronisk funktionsdygtig udgave af ODS klar på under to år, skyldes i høj grad at vi har kunnet overtage en gennemprøvet model fra et lignende projekt. Hvis man i stedet betragter de første forsøg der blev gjort, skræmmer sporene noget mere. I Sverige begyndte man allerede i 1983 at digitalisere *Svenska Akademiens Ordbok*, og den løsning redaktionen valgte, var optisk scanning. Der skulle imidlertid gå næsten 15 år før processen var gennemført, og den der besøger hjemmesiden i dag, vil hurtigt indse at resultatet langtfra er perfekt. I England begyndte man at digitalisere *Oxford English Dictionary* i 1984, og her valgte man den løsning at lade værket indtaste. Det tog væsentlig kortere tid, i alt fem år, men til gengæld krævede det 120 indtastere, 50 korrekturlæsere og 13,5 mio. dollars før det digitale manuskript var etableret. Vi er derfor taknemmelige over at kunne nyde godt af disse projekters dyrekøbte erfaringer. Det har gjort det muligt at gennemføre digitaliseringen af ODS både væsentlig hurtigere og for blot en brøkdel af de beløb der har været brugt af vores svenske og engelske kolleger.

Bortset fra den faglige og tekniske gevinst der er ved at have ODS i elektronisk form, glæder vi os også over at ODS nu kan nå ud til et nyt, stort publikum. Selvom ODS siden 1918 i alt er udkommet i nær ved 10.000 eksemplarer (*Svenska Akademiens Ordbog* og *Norsk Ordbok* i Norge udkommer til sammenligning i oplag på blot 1.000 eksemplarer), kan man alligevel ikke kalde den en udpræget folkelig succes. Derfor er det overordentlig glædeligt at vi kan notere en stor interesse for ODS på nettet. Efter den overvældende interesse der fulgte umiddelbart efter lanceringen, ser det ud til at antallet af besøgende stabiliserer sig mellem 1.000 og 1.500 daglige brugere, der foretager omkring 20.000 søgninger. I en tid hvor public service nævnes stadig oftere i den kulturpolitiske debat, er det ikke nogen uvæsentlig biomstændighed ved projektet.

DET MODERNE SPROG

Et kedeligt træk ved papirordbøger er at de nødvendigvis er forældede allerede når de udkommer. *Den Danske Ordbogs* sidste bind udkom i november 2005, men den samling tekster som udgør det væsentligste grundlag for artiklerne, stammer tilbage fra perioden 1983-1992. Det betyder at udmærkede og gangbare ord som *fugleinfluenza, arbejdspladsvurdering, babybio* og *fladskærm* ikke kan slås op i ordbogen fordi de først er blevet udbredte i sproget efter ordbogens primærperiode. Og af samme grund er hovedparten af de citater der bringes i ordbogen, 10-20 år gamle. Der gælder derimod andre betingelser for onlineredigering. Fordi sproget ændrer sig forholdsvis hurtigt, ikke mindst den del som vedrører ordstoffet, egner inter-



FIGUR 2.
Model af digitaliseringsprocessens to faser. Manuskriptet er først blevet indtastet i et almindeligt tekstbehandlingsprogram i Nanjing, Kina, hvorefter de to versioner er blevet efterbehandlet i Trier. Den elektroniske kopi af det typografiske manuskript er udgangspunkt for det videre arbejde hos DSL. Modellen er en tilpasning af den tilsvarende projektskitse for *Deutsches Wörterbuch*.

nettet sig godt som medie for ordbøger over nutidssproget. Indholdet kan nemt ændres og suppleres med nye artikler og friske citater når der er behov for det.

Når den første version af *Den Danske Ordbog* lægges i ordnettet i slutningen af 2006, vil det derfor ikke være en kopi af papirordbogen. Netudgaven vil på nogle områder indeholde flere oplysninger, på andre områder færre. Den vil indeholde flere opslagsord dels fordi den vil indeholde et antal nyskrevne artikler over nye ord der er kommet til i sproget, dels fordi der vil indgå et stort antal små artikler over sammensætninger og afledninger som i papirordbogen af pladsøkonomiske hensyn kun nævnes som eksempler på orddannelsesmuligheder. Fordi plads ikke er noget problem i de digitale medier, har vi besluttet at gøre disse ord til rigtige opslagsord i netudgaven. Til gengæld vil netordbogen af hensyn til forlagets salg af bogen ikke have alle papirversionens oplysninger med i den første version.

Ligesom ordbogen løbende forsynes med nye artikler, vil tekstsamlingen blive forsynet med nye tekster. Nye tekster sikrer at beskrivelsesgrundlaget holdes opdateret, og kan hjælpe redaktørerne med automatisk at blive opmærksom på nye artikelkandidater. På længere sigt kan materialet også blive vigtigt til undersøgelser af den sproglige udvikling over tid. Af samme grund er det vigtigt også at forøge tekstmængden bagud i tiden. Et oplagt materiale ligger allerede tilgængeligt i Arkiv for Dansk Litteratur, en samling skønlitterære tekster af klassiske, danske forfattere, udarbejdet i et samarbejde mellem DSL og Det Kongelige Bibliotek og velegnet som supplerende tekstmateriale til en overvejende litterært baseret ordbog som ODS.

SØGEMULIGHEDER

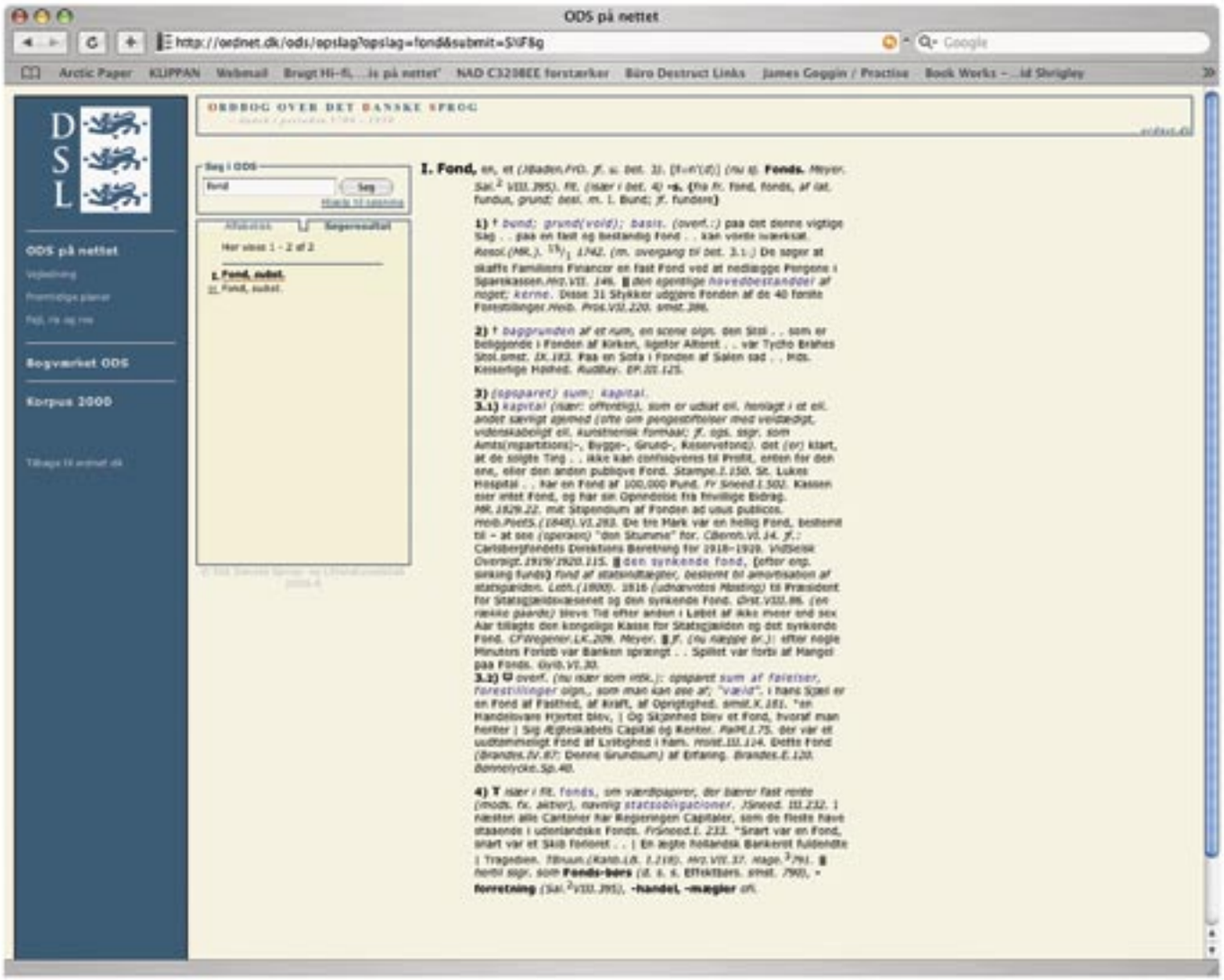
De muligheder som mediet tilbyder, skal udnyttes til forskellige former for avanceret søgning. Bortset fra søgning med jokertegn og de tværgående søgninger i databaseformatet som selvfølgelig skal være mulig i begge ordbøger, vil det især være i den moderne del eksperimenterne kommer til at foregå. Korpusdelen vil blive forsynet med nyt design og med flere søgefaciliteter der først og fremmest skal udnytte at teksterne er blevet syntaktisk opmærket – eller *parset* som det også hedder. Det betyder at det bliver muligt at søge på bestemte syntaktiske mønstre i materialet (man kan fx være interesseret i at undersøge om et bestemt verbum kan tage objekt, eller hvilke substantiver der typisk er subjekt for verbet). Et område der har høj prioritet, er at øge integrationen mellem korpus- og ordbogsdel så man ikke blot kan skifte mellem søgninger de to steder som en generel facilitet, men sådan at resurserne udnyttes netop på det sted man har slået op. Hvis man har slået et ord op i ordbogen, betyder det at man fx skal kunne søge oplysning om den procentvise fordeling af alternative bøjningsformer eller hvilke ord der er almindelige i en bestemt grammatisk konstruktion. Det kan også være yderligere eksempelmateriale, fx på citater, typiske naboord eller syntaktiske mønstre.

Et andet område der prioriteres højt, er muligheden for at lave begrebsorienterede søgninger. Skjult rundt omkring i de mange ordbogsartikler ligger et væld af oplysninger om ordenes relationer til hinanden: synonymer, antonymer, fagmarkeringer, stilmarkeringer osv. De kan bare ikke udnyttes i en traditionel papirordbog hvor ordningsprincippet er alfabetisk – medmindre oplysningerne arrangeres på en ny måde sådan som det gøres i en tesaurus, en begrebsordbog eller et krydsogtværsleksikon. I et elektronisk opslagsværk er der derimod ikke noget teknisk til hinder for det. Det kræver dog en ikke ubetydelig redaktionel bearbejdning at beskrive nettet af betydningsforbindelser så det bliver elektronisk søgbart. DSL har derfor siden 2005 samarbejdet med Center for Sprogteknologi på Københavns Universitet om at skabe et sådant sprogteknologisk betydningsnet (projektet *DanNet*, finansieret af Forskningsrådet for Kultur og Kommunikation) på grundlag af *Den Danske Ordbog*s artikler. Resultatet vil dels kunne anvendes til de begrebsorienterede søgninger i ordnettet, dels have en række selvstændige anvendelsesmuligheder inden for forskellige former for intelligent informationshåndtering i sprogteknologiske it-systemer.

FREMTIDEN

At udvikle disse nye funktionaliteter til det elektroniske opslagsværk er ikke noget der bare sker af sig selv. Det indebærer en ikke ringe forskningsindsats, og i flere henseender er der tale om en ny type forskning som kræver den særlige kombination af ekspertise inden for både leksikografi, korpuslingvistik og sprogteknologi som DSL har opbygget og udviklet inden for de senere år. Vi giver med *ordnet.dk* vores bud på nogle af de muligheder som vi ser for fremtidens elektroniske ordbøger og opslagsværker.

Ligesom edderkoppens net har ordnettet ikke nogen på forhånd fastsat størrelse; det er fleksibelt, lader sig strække



FIGUR 3.
Netudgaven af *Ordbog over det danske Sprog*.

og kan udvides når der er behov for det. Der kan spindes flere tråde både i længden og på tværs. I DSL håber vi at ordnettet kan udvides med alle de mange ord der hele tiden opstår i sproget, og på længere sigt også bagud i tiden til at omfatte beskrivelser af det danske ordforråd fra de ældre og ældste sprogtrin. Og parallelt med den leksikografiske beskrivelse bør der foreligge tekstudgivelser fra de tilsvarende perioder. En del af det indholdsmæssige arbejde er allerede udført, men der er endnu lang vej før resultatet foreligger i et passende elektronisk format. Nu håber vi i første omgang at de grundlæggende tråde viser sig holdbare.

HENVISNINGER:

Arkiv for Dansk Litteratur: www.adl.dk

DanNet: www.wordnet.dk

Deutsches Wörterbuch: http://germazope.uni-trier.de/Projects/DWB

Oxford English Dictionary:www.oed.com

Svenska Akademiens Ordbok: http://g3.spraakdata.gu.se/saob