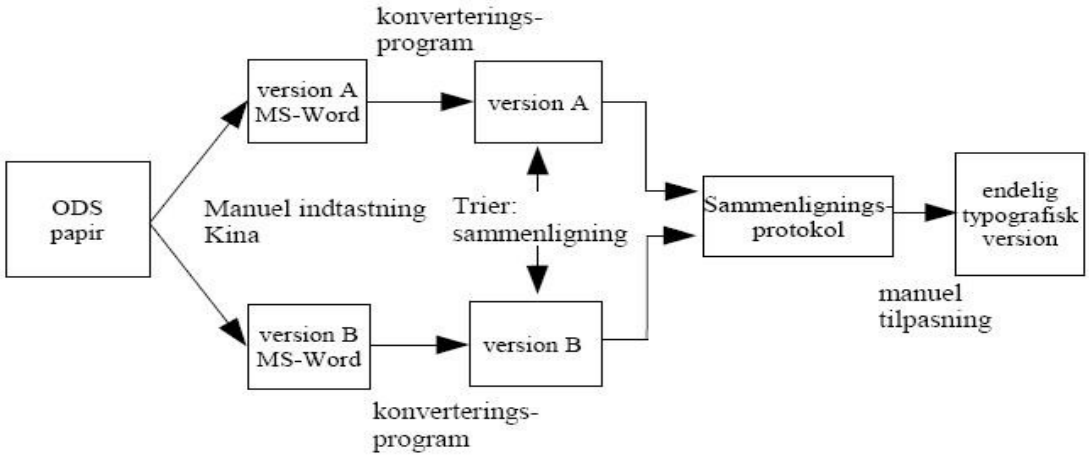
ordnet.dk – ordbøger og korpus på internettet

Af Henrik Lorentzen og Lars Trap-Jensen, Det Danske Sprog- og Litteraturselskab

Ordnet.dk er et websted der giver samtidig adgang til to ordbøger og et tekstkorpus. Siden 2004 har Det Danske Sprog- og Litteraturselskab for en bevilling fra Carlsbergfondet og Kulturministeriet og støttet af Det Elektroniske Forskningsbibliotek arbejdet på at organisere data i en hensigtsmæssig struktur og udvikle en hjemmeside med en række af de funktioner der hører til i en digital ordbog. De enkelte elementer i resursen er blevet offentliggjort i den takt de er blevet klar til det. Med lanceringen af Den Danske Ordbog i en netudgave den 2. november 2009 er tiden kommet til at gøre status over projektet og redegøre for de valg der har skullet træffes undervejs. I det følgende gennemgås de enkelte dele, idet der dog gøres mest ud af Den Danske Ordbog som det senest tilkomne element på hjemmesiden.

1. Ordbog over det danske Sprog

Ordbog over det danske Sprog (ODS) udgør både det første og det sidste element af *ordnet.dk*. Værket er blevet digitaliseret inden for rammerne af projektet og blev som noget af det første sendt til Kina for at blive indtastet af firmaet TQY DoubleKey i Nanjing.



Figur 1: Digitaliseringsmodel for ODS

Det skete i et samarbejde med den afdeling ved universitetet i Trier som havde været ansvarlig for den tilsvarende digitalisering af brødrene Grimms *Deutsches Wörterbuch*. Digitaliseringsmodellen er gengivet i skematisk form i figur 1, men eftersom de enkelte dele af processen er blevet beskrevet i andre artikler, henvises interesserede læsere hertil for nærmere detaljer (se fx Bojsen & Trap-Jensen 2005, Trap-Jensen 2006, hvorfra figur 1 er gengivet).

Siden november 2005 har det digitaliserede værk været offentligt tilgængeligt, dog kun i en foreløbig version. Dels har de fem supplementsbind ikke kunnet ses i netudgaven, dels har værket kun været ret nødtørftigt indholdsopmærket. I begyndelsen var kun opslagsord, varianter og visse sammensætninger isoleret i den underliggende database; siden er oplysninger om ordklasse kommet til, mens resten af en artikel blot optræder som én tekstblok. I de indtastede filer er til gengæld alle den trykte bogs typografiske oplysninger bevaret så netudgaven har kunnet præsenteres i en bognær, læselig form.

Det har imidlertid hele tiden været planen at de typografiske oplysninger skulle bruges som udgangspunkt for en egentlig indholdsmæssig opdeling af artiklerne. Kun derved bliver det muligt at præsentere indholdet i mere overskuelige blokke på skærmen og at søge i bestemte elementer, fx søgning efter långivende sprog, citater af en bestemt forfatter eller ord der er opstået i en bestemt periode. Denne opgave bliver også den sidste del af projektets første fase sammen med offentliggørelsen af supplementsbindene. I skrivende stund pågår dette arbejde stadig, og det forventes at en ny udgave af ODS på nettet kan præsenteres til sommeren 2010. Heri vil flere dele af artikelblokken være genkendt så artiklerne kan vises i ordnettets generelle design, og supplementsartiklerne vil ligeledes indgå. Det vil dog også i lang tid fremover være en opgave at raffinere indholdsstrukturen yderligere. Det samme gælder supplementsartiklerne: Kun helt nye artikler vil uden problemer kunne flettes sammen med hovedværkets artikler, mens artikler der tilføjer, ændrer eller sletter noget i eksisterende artikler, i første omgang blot bringes som et selvstændigt supplement til de pågældende artikler.

Endelig bør det nævnes at ODS er blevet revideret indholdsmæssigt på et enkelt punkt. Mange af ordbogens citater er hentet fra samleværker der er udkommet væsentlig senere end originaludgaven, undertiden endda efter forfatterens død. Det kan nok forvirre mange brugere at læse kildeangivelser som *Blich.(1920)* og *HCAnd.(1919)*, og vi har derfor opdateret kildebasen sådan at det af kildeforkortelsen fremgår hvornår det citerede værk er udkommet første gang. Derudover kan man også se årstallet for den udgave der citeres fra, hvis denne afviger fra førsteudgaven. Også dette vil slå igennem når den nye version offentliggøres.

2. KorpusDK

En ny version af DSL's korpusresurser har siden foråret 2008 været tilgængelig under *ordnet.dk*. Set fra brugerens side er der ikke sket store, synlige ændringer i indholdsdelen. Det er de samme tekster man får adgang til som på det gamle site *Korpus 2000*. Ikke desto mindre skal navneændringen til *KorpusDK* signalere at det er noget nyt. Indtil videre er det mest hjemmesidens udseende og søgefunktionerne der har ændret sig for brugerne. Det har været prioriteret at give både eksperter og mere almindeligt interesserede mulighed for at søge præcist efter bestemte ord. Der tilbydes derfor tre forskellige måder at søge på: Standardsøgningen er den hurtige måde at søge efter et bestemt ord og se det i en konkordans. Hvis man vil lave mere raffinerede søgninger, tilbydes der en udvidet søgning hvor brugeren selv kan bygge søgeudtrykket op ved at tilføje flere ord og specificere antallet af mellemliggende ord, ordklasse og bøjningsform. Denne mulighed skulle gerne være selvforklarende og lige til at bruge af de fleste. Hvis man skal have det fulde udbytte og kunne formulere alle de søgninger som systemet kan håndtere, kræves det at man behersker det særlige søgesprog som korpusprocessoren CQP benytter. Søgemuligheden "Formel søgning" giver adgang til alle de søgninger der kan formuleres i CQP. Til gengæld kan man ikke forvente at almindelige

brugere er fortrolige med den formelle syntaks, og derfor tilbydes denne søgning blot som en mulighed på linje med de to øvrige søgemåder.

Figur 2: De tre søgemuligheder i KorpusDK

Selvom det ikke er umiddelbart synligt, er der dog også sket en del på indholdssiden. En række fejl og uhensigtsmæssigheder i opmærkningen er blevet fjernet, ligesom der er sket en tilpasning af samspillet mellem data, søgemaskine og hjemmeside. Endelig skal det siges at vi løbende har modtaget nye tekster siden 2005, primært avis- og fagblade, men også i et vist omfang skøn- og faglitterære tekster. I øjeblikket arbejder vi på en model til forbedret opmærkning af de indkomne tekster som vil sætte os i stand til at offentliggøre dem i KorpusDK. Denne opgave udføres til dels i et samarbejde med Dansk Sprognævn inden for forskningsinfrastrukturprojektet *DK-CLARIN*. Dette projekt udløber med udgangen af 2010 eller forår 2011, og til den tid forventer vi at være klar med nyt indhold i KorpusDK.

3. Den Danske Ordbog

Arbejdet med at klargøre Den Danske Ordbog (DDO) til internetudgaven kan opdeles i to hovedområder, dels revision af datastrukturen, dels tilføjelse og revision af indhold.

3.1. Datastruktur

Overgangen til at vise data i en elektronisk version betød at vi ønskede at nedlægge mange af papirordbogens henvisninger. De giver god mening i det papirbundne, lineære format hvor brugeren slår ord op på alfabetisk plads og må hjælpes videre hvis den ønskede oplysning ikke befinder sig der. Det gælder fx stavevarianter som de officielle dobbeltformer og almindelige fejlstavninger; i den trykte ordbog optræder de som opslagsord der blot skal lede brugeren på rette vej: *majonæse* henviser til *mayonnaise* og *akceptere* til *acceptere*. Den slags tomme artikler er der ikke brug for i en elektronisk visning hvor man sikrer sig at variantformer knyttes til den korrekte opslagsform i data således at man altid kommer til den rigtige artikel hvis man som bruger indtaster en variantform der er registreret i ordbogen. Fx vil man ved indtastning af formen *alexandriner* finde opslagsordet *aleksandriner* fordi variantformen med x er registreret i denne artikel. Det kan diskuteres hvordan dette skal formidles til brugeren. Skal man blot få vist artiklen med den korrekte staveform uden kommentarer, eller skal man oplyse om at formen med x er en uofficiel, men udbredt variant? Man kunne også være

endnu mere eksplicit og lade brugeren komme til en mellemstation hvor det oplyses at formen *alexandriner* ikke findes i ordbogen, men at vi i stedet kan tilbyde formen *aleksandriner*. Vi har skønnet at det er mest brugervenligt at komme direkte til artiklen, men med oplysning om den indtastede forms status. Hvis man søger på en fejlstavet form eller laver en slåfejl, kommer man ikke til en artikel, men får under overskriften "Mente du ..." hjælp i form af et antal ordbogsartikler der ligner det indtastede.

Figur 3: Artiklen *aleksandriner* i DDO

En anden type papirhenvisninger der er blevet nedlagt, er de rene synonymhenvisninger hvor to ord har nøjagtig samme betydning og definitionen derfor kun bringes i den ene artikel, mens der henvises fra den anden. Et eksempel er *absolutisme*, som i den politiske betydning er synonymt med *enevælde* og peger på den artikel. I de tilfælde hvor henvisningsartiklen ikke indeholder supplerende information i form af fx brugsrestriktioner ("især talesprog", "gammeldags", "uformelt" osv.), har vi ment at det var uproblematisk at hente definitionen og vise den i den oprindelige henvisningsartikel; samtidig gives henvisningsordet som synonym. Det sparer brugeren for at klikke sig videre. I data foregår det på den måde at den gamle henvisning, som tidligere gik til et ord, nu går præcist til en betydning ved hjælp af et unikt id-nummer.

Arbejdet med henvisningerne har ikke kun handlet om at nedlægge henvisninger; for at muliggøre klik til korrekt opslagsord og korrekt betydning er der tværtimod indført en lang række nye henvisninger som findes i data, men kun er interessante og synlige for brugeren som klikmuligheder. I den trykte DDO optræder der i mange artikler synonymer, antonymer og ord fra samme ordfelt, de sidste indledt med markøren JF. Brugeren af papirordbogen forventes at slå en sådan henvisning op og selv kunne lokalisere den rigtige betydning, evt. efter at have valgt mellem homografer. I netudgaven vil vi gerne hjælpe brugeren helt hen til det korrekte sted ved at tilbyde en klikbar henvisning. I mange tilfælde, nemlig ved monosemе og ikke-homografe ord, har vi indført henvisningen automatisk, men hvis målet for henvisningen ikke var entydigt, har vi manuelt indsat en reference til den rigtige artikel og allerhelst også den rigtige betydning. Et eksempel: I artiklen *strygeinstrument* optræder synonymet *stryger*; dette ord har imidlertid to betydninger, dels 'musiker', dels 'musikinstrument'. Henvisningen skal kun gå til den sidste betydning så man ved klik lander netop der. I øjeblikket fungerer en del af disse henvisninger som de skal, men der er også en del fejl eller

upræcise henvisninger, så det er et område der skal arbejdes videre med.

Et andet arbejdsintensivt område er disambiguering af ord i flerordsforbindelser. I den trykte DDO eller de underliggende data var det ikke angivet hvilket lemma der indgik i et givet flerordsudtryk, men det ønsker vi at gøre i den elektroniske version. Det har to formål, dels søgeteknisk, dels klikrelateret. I en trykt ordbog vil et flerordsudtryk typisk stå i én artikel, evt. med henvisning fra visse af de andre indgående ord. Den begrænsning er ikke aktuel i en elektronisk version hvor vi principielt gerne vil præsentere en ordforbindelse i alle de artikler den henter ord fra, således at talemåden *råbe ulven kommer* optræder i 3 forskellige artikler uden at brugeren behøver spekulere over hvilket ord vi har valgt at anbringe udtrykket under. For at muliggøre en sådan søgning og præsentation må vi forsyne de enkelte ordformer i udtrykket med en oplysning om hvilket lemma formen hører til. Første fase i den proces er automatisk, så der kan forekomme fejl, men målet er at minimere disse. Målet er også at de enkelte ord i en ordforbindelse skal være klikbare så man kan hoppe hen til den relevante artikel og også gerne den rigtige betydning.

Ved meget almindelige ord bliver listen over flerordsforbindelser meget lang, fx optræder *kommer* (dvs. former af lemmaet *komme*) i ikke mindre end 179 faste udtryk. I brugergrænsefladen findes de i højre side under overskriften "Faste udtryk", som i udgangspunktet er foldet sammen og kan foldes ud ved tryk på knappen +. I selve artikelvisningen, i midterfeltet, har vi ment at det blev for uoverskueligt med så mange faste udtryk, så hvis antallet overstiger 50, vises kun de udtryk der også optræder i den pågældende artikel i den trykte bog. Men de er fortsat klikbare fra den lange liste i højre side. På den måde håber vi at imødekomme forskellige brugeres forskellige behov, men det kræver selvfølgelig at systemet og opstillingen er gennemskuelige.

3.2. Indhold

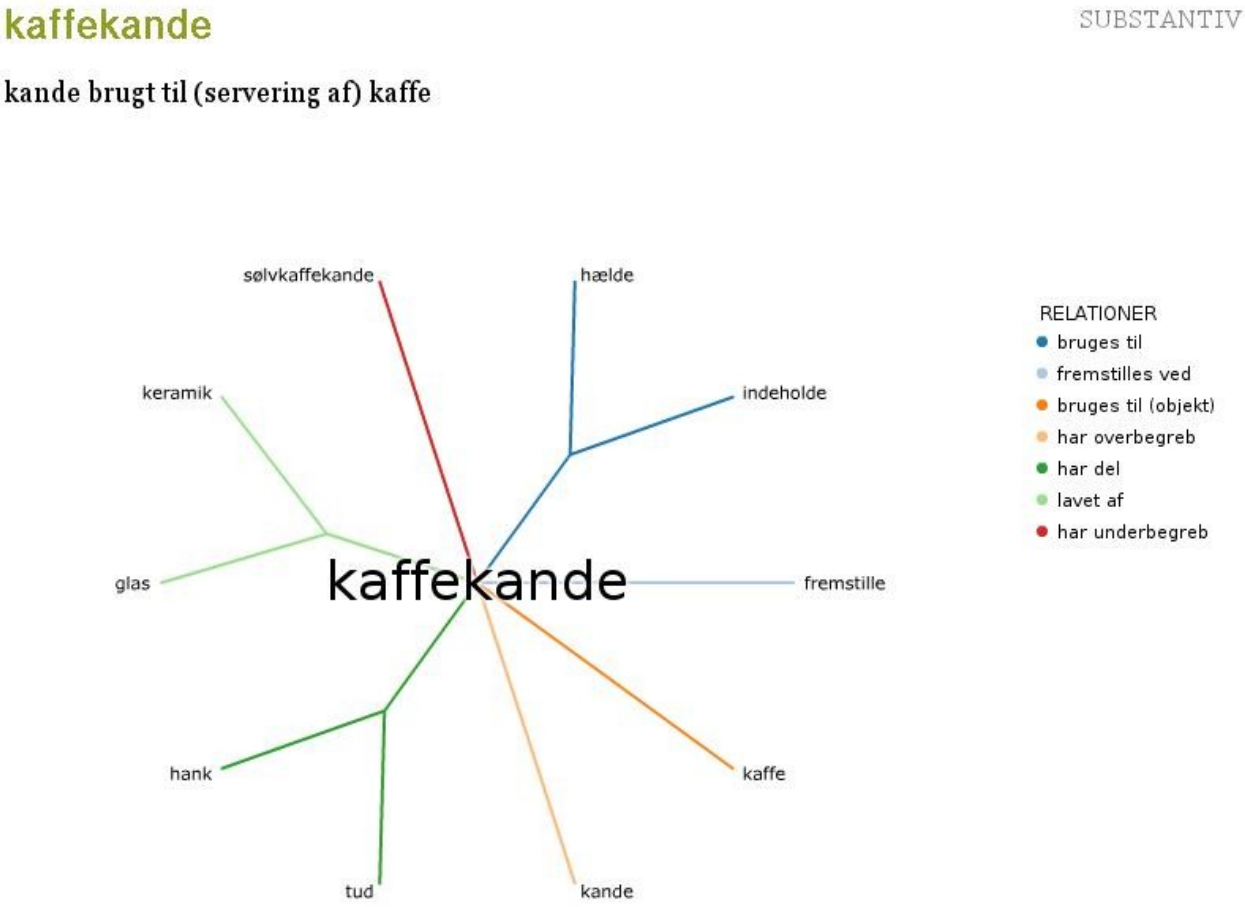
Hidtil har vi beskrevet hvordan datastrukturen er blevet forbedret og beriget, men der er også sket meget med ordbogens indhold så det ikke blot er en tro kopi af den trykte DDO der nu ligger på nettet. Det mest interessante er nok de nytilkomne artikler, dvs. ord der slet ikke er med i DDO på papir. Det drejer sig dels om nye ord og betydninger der ikke var dukket op eller etablerede da DDO gik i trykken, dels om ord og betydninger der principielt godt kunne have været med i DDO, men fx var lavfrekvente i ordbogens korpus. Den første type har vi valgt at promovere lidt mere ved at lade dem optræde i højre side af grænsefladen på DDO's forside; der præsenteres en tilfældig håndfuld som skifter når forsiden genindlæses, og det kan fx være *trainsurfe*, *break even*, *voldsromantik, podcaste* og *bagstiv*.

En anden type opslagsord er de ca. 37.000 ord der optræder som orddannelseseksempler i bunden af de trykte DDO-artikler. Redaktionen er af Bergenholtz & Vrang (2004) blevet kritiseret for at tælle dem med som opslagsord, og i Lorentzen & Trap-Jensen (2004) imødegår vi kritikken ved at gøre gældende at selvom de ikke har egen artikel, har de dog fået redaktionel behandling i og med de er placeret under et (evt. flere) af de ord de er dannet af, og det er angivet hvilken betydning de knytter til sig. Vi har dog ønsket at give dem en mere udførlig behandling i netudgaven, og i første omgang er de nu alle beskrevet i selvstændige artikler med oplysninger om variantformer, ordklasse, evt. genus samt bøjning. En del af dem er også forsynet med definition og undertiden med citat. Det er selvfølgelig ikke helt nok, men i de kommende år vil en af redaktionens opgaver være at bringe denne gruppe artikler op på et mere tilfredsstillende niveau. I februar 2010 var der 20.058 artikler uden betydningsbeskrivelse.

Et særligt problem er lydskriften. I papirordbogen bruger vi Dania, der som bekendt er skræddersyet til dansk. Desværre kan den ikke uden videre vises på hjemmesider fordi den ikke er en del af Unicode. Vi har derfor valgt at bruge IPA i halvgrov notation som til gengæld kan vises i alle browsere. Under alle omstændigheder er det nok de færreste almindelige

brugere der overhovedet er i stand til at tyde lydskrift. På længere sigt vil vi gerne tilbyde en eller flere "levende" udtaler sådan som det allerede kendes fra danske og udenlandske ordbogssider, og til den tid bliver lydskriften måske en sekundær oplysningstype som man kan vælge at undertrykke.

En ny oplysningstype gemmer sig under overskriften "Beslægtede ord". Det Danske Sprog- og Litteraturselskab har siden 2004 arbejdet sammen med Center for Sprogteknologi om at opbygge et leksikalsk-semantisk ordnet for dansk, et wordnet, ud fra primært definitionerne i DDO (se fx Pedersen & Asmussen 2006). Et wordnet er groft sagt et hierarkisk system hvor begreberne er ordnet efter over- og underbegreb, idet der samtidig gives andre semantiske oplysninger om fx genstandes dele, deres formål og typiske anvendelse, personers beskæftigelse m.m. Denne resurse blev offentliggjort i marts 2009 under navnet DanNet, og nu kommer de berigede oplysninger herfra hjem til deres udgangspunkt. I første omgang får man ved klik på knappen "Beslægtede ord" vist et begrebs overbegreb, eventuelle underbegreber samt sidebegreber, dvs. andre begreber med samme overbegreb. Fx har *kaffekande* overbegrebet *kande*, mens det har *sølvkaffekande* som underbegreb. Andre ord med *kande* som overbegreb er bl.a. *termokande, mælkekande* og *tinkande*. Knappen der trykkes på, er efter udviklerens ønske mærket "Beta" fordi funktionen endnu er ganske eksperimentel, og fordi der kan være fejl eller uhensigtsmæssigheder i dataene fra DanNet. Planen er dog på længere sigt at udvikle en søgeside til specifik begrebssøgning hvor man ikke bare søger på over- og underbegreber, men kan tilføje søgeparametre som fx funktion, indhold eller materiale. Det ville fx anbringe de tre ovennævnte kander i hver sin gruppe.



Figur 4: Visuel fremstilling af *kaffekande*'s relationer

4. Integration af resurser

Tanken bag *ordnet.dk* har fra starten været at de forskellige elementer skulle udnyttes på tværs og på den måde berige hinanden. DDO er jo en korpusbaseret ordbog, og en del af ordbogens korpus indgår i KorpusDK, så derfor var det nærliggende at lade den omstændighed komme brugerne til gode på forskellig vis. I den nuværende version sker det overvejende under overskriften "Relaterede søgninger" i skærmens venstrespalte. Her har brugeren altid mulighed for at komme fra et søgeresultat til den tilsvarende søgning i en af de andre baser. For KorpusDK's vedkommende kan man vælge mellem en konkordanssøgning – enten svarende til den indtastede streng eller til det pågældende lemma – og søgning efter kollokationer, eller *naboord* som de kaldes i KorpusDK. Når man står i en artikel i DDO har man desuden mulighed for at se eksempler fra korpus på de kollokationer der bringes under overskriften "Eksempler" – det sker ved at klikke på det lille Kikon der står efter kollokationen i det store midterfelt på skærmen.

I KorpusDK kan man vælge at se en liste over faste udtryk med et bestemt ord der forekommer i DDO (jf. figur 5). Et klik på et af udtrykkene slår det pågældende udtryk op i korpus og viser resultatet i en konkordans.

Figur 5: Faste udtryk fra DDO i KorpusDK

Endelig har vi allerede set hvordan oplysninger fra DanNet er brugt som en ny oplysningstype i DDO under overskriften "Beslægtede ord". De nævnte muligheder er langt fra udtømmende.

Man kan forestille sig frekvensangivelser for opslagsord, varianter og bøjningsformer, mulighed for korpussøgning på syntaktiske mønstre fra DDO's konstruktionsoplysninger, de tilsvarende muligheder fra ODS til et korpus over historiske tekster. Listen kan fortsættes yderligere, men indførelsen af mulighederne hører en senere fase til.

Litteratur

Bergenholtz, Henning & Vibeke Vrang 2004: "Ny dansk ordbog i seks bind for sekretærer og forskere". *LexicoNordica 11*, s. 165-189.

Bojsen, Else & Lars Trap-Jensen 2005: "ODS, ODS-S og fremtiden (sammen med Else Bojsen)". Peter Widell og Mette Kunøe (red.): *10. Møde om Udforskningen af Dansk Sprog*, Århus 2005, s. 58-67.

Lorentzen, Henrik & Lars Trap-Jensen 2004: "Kommentarer til Henning Bergenholtz & Vibeke Vrang: Ny dansk ordbog i seks bind for sekretærer og forskere". *LexicoNordica 11*, s. 191-201.

Pedersen, Bolette Sandford & Jørg Asmussen 2006: "DanNet – fra ordbog til et leksikalsk-semantisk WordNet for dansk". *LEDA-Nyt 42*, s. 3-11.

Trap-Jensen, Lars 2006: "Digitaliseringen af den store danske ordbog – et kapitel i historien om ODS på nettet". *Referencen* nr. 1, 2006, 36. årgang. Faggruppen for Reference-arbejde, Bibliotekarforbundet, s. 5-10.

Henrik Lorentzen
seniorredaktør
Det Danske Sprog- og Litteraturselskab
hl@dsl.dk

Lars Trap-Jensen
ledende redaktør
Det Danske Sprog- og Litteraturselskab
ltj@dsl.dk