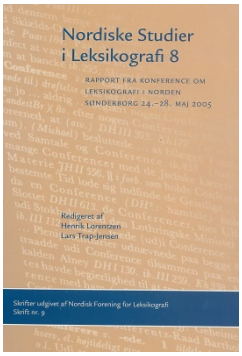
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: ordnet.dk – et nyt sprogligt opslagsværk på internettet

Forfatter: Henrik Lorentzen og Lars Trap-Jensen

Kilde: Nordiska Studier i Leksikografi 8, 2006, s. 253-264
Rapport från Konferens om lexikografi i Norden, Göteborg 27.-29. maj 1999

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.

Henrik Lorentzen og Lars Trap-Jensen

ordnet.dk – et nyt sprogligt opslagsværk på internettet

***ordnet.dk* – a new lookup tool on the Internet**. We give an account of the architecture and functionalities of a web-based application that gives joint access to three existing resources and seeks to combine them in a new way. The project includes digitisation of the historical 28-volume dictionary *Ordbog over det danske Sprog* (Dictionary of the Danish Language) as well as the integration with its five supplementary volumes. The modern *Den Danske Ordbog* (The Danish Dictionary) will continuously be supplied with new material and adapted to a concept for a digital dictionary which will among other things introduce new features such as onomasiological queries and larger integration with the *Korpus 2000* component. The corpus module is supplied with additional material, and a revision of both the interface and functionalities of the module is foreseen as part of the project which runs until the end of 2009.

1. Introduktion

I denne artikel vil vi redegøre for et opslagsværk som er under udarbejdelse i Det Danske Sprog- og Litteraturselskab (DSL). Det drejer sig om et elektronisk opslagsværk som skal koble tre eksisterende værker sammen på hjemmesiden www.ordnet.dk, nemlig *Ordbog over det danske Sprog* (ODS), *Den Danske Ordbog* (DDO) og *Korpus 2000*. Projektet løber i årene 2004-2009 og finansieres af Carlsbergfondet og Kulturministeriet med støtte fra Det Elektroniske Forskningsbibliotek.

Projektet indebærer at ODS først digitaliseres og derefter sammenflettes med *Supplementet til ODS*, hvis femte og sidste bind udkom i efteråret 2005.

DDO foreligger allerede som digitalt manuskript, og det er derfor her mulighederne for at eksperimentere med det elektroniske medie kommer til at foregå. En af fordelene ved netpublicering er muligheden for at koble flere værker sammen til ét og springe imellem dem. Den mulighed vil vi udnytte ved at knytte ordbogsdel og korpustekster sammen, både sådan at korpusmaterialet udnyttes mere effektivt i den beskrivelse man får i ordbogsartiklerne, og sådan at brugeren får mulighed for at lave sine egne undersøgelser med udgangspunkt i det ord han/hun lige har slået op. Derudover skal både ordbog og korpus forsynes med flere artikler og tekster så basen holder sig opdateret med den sproglige udvikling.

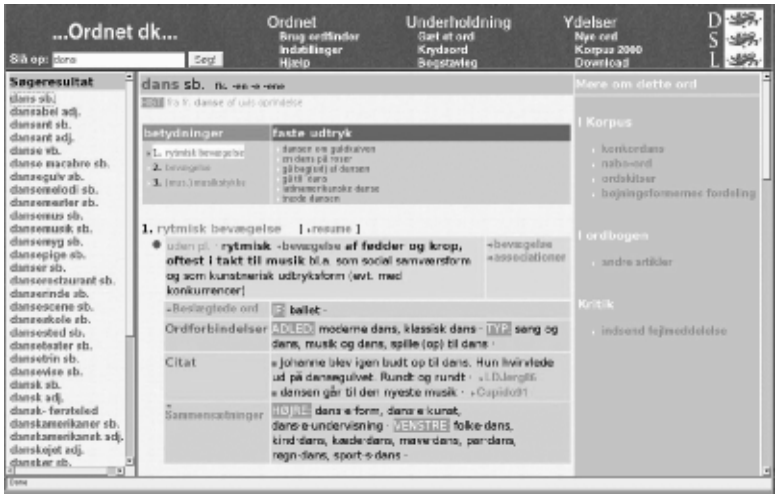
2. Den Danske Ordbog på nettet – en prototype

Som et første skridt i retning af at lægge DDO på internettet har vi udarbejdet en intern prototype til en elektronisk version af ordbogens data. Og det er med overlæg at vi taler om ordbogens data, og ikke bare ordbogen, for det er afgørende for os at versionen på nettet ikke blot er en tro kopi af den trykte bog vist på en computerskærm. De første generationer af elektroniske ordbøger bestod netop blot af de

trykte ordbogsspalter forsynet med visse søgefunktioner, oftest begrænset til søgning på opslagsord og fritekstsøgning i den fulde tekst. Vi vil gerne gå et par skridt videre og ud over de nævnte tilbyde mere avancerede søgninger, fx søgninger der ikke tager udgangspunkt i et bestemt opslagsord eller udtryk (semasiologisk søgning), men i et bestemt betydningsindhold (onomasiologisk eller begrebsorienteret søgning). Det er vigtigt at understrege at der er tale om en prototype der lægger vægt på mulighederne og ikke på præsentationen, så derfor er det ikke den endelige grænseflade der ses på de følgende figurer.

Som ved de fleste opslagsværker og andre tjenester på nettet, er der et indtastningsfelt hvor man indtaster det ord man vil slå op, fx substantivet *dans*. Resultatet af denne søgning er en liste med ord der begynder med opslagsordet, altså en højretrunkeret søgning, der også inkluderer opslagsord som fx *dansemyg* og *dansk*. Det er også muligt at lave en venstretrunkeret søgning, nemlig på strengen **dans*, hvilket ud over rigtige andetledssammensætninger som fx *krigsdans* og *kædedans* også giver *impedans* og det for korpusleksikografer nok så vigtige ord *konkordans*. Der er i dette tilfælde altså tale om rent tegnorienterede søgninger.

Hvis man klikker på ordet *dans* i resultatlisten, får man et skærmbillede der minder om en bog, men alligevel er lidt anderledes (se figur 1). Noget af det nye er at man i begyndelsen af artiklen får vist et resumé over ordets forskellige betydninger, for *dans*' vedkommende 3 betydninger. Til højre for dette ses en liste over de faste udtryk som *dans* indgår i. På den måde får man hurtigere end i en trykt bog et overblik og kan relativt let gå til det ønskede sted med et enkelt klik. Fordelene er naturligvis endnu større ved opslagsord med mange betydninger og mange faste udtryk.



Figur 1. *Artiklen* ***dans*** *i ordnet-prototypen*

Definitionerne i Den Danske Ordbog følger i stor udstrækning den klassiske aristoteliske definitionsmåde med genus proximum og et eller flere differentiae specificae. Betydningsresuméerne er automatisk genereret ud fra dette genus proximum, som blev udvalgt af redaktørerne i redigeringsprocessen og skrevet i et særligt, usynligt element der oprindelig var tænkt som et eksperiment, men nu viser sig at være yderst anvendeligt, ikke bare til betydningsresuméer. Disse resuméer bygger enten på genus proximum alene eller genus proximum i kombination med et par af de omgivende ord. Hvis dette system skal fungere hensigtsmæssigt i alle tilfælde, kræver det en manuel redigering, men som et foreløbigt bud på hvordan man hurtigt kan præsentere betydninger, synes vi det er ganske godt. I de store engelsksprogede learner's dictionaries som fx Macmillan og Longman har man allerede denne resuméfacilitet i de elektroniske udgaver, især for lange artiklers vedkommende, men det har tydeligvis også krævet menneskelig redigering da der ikke altid er én til én-overensstemmelse mellem tekststrenge i definitionerne og i resuméerne.

De faste udtryk fremgår som nævnt af en liste der som i den trykte bog er alfabetiseret efter første ord i udtrykket. Alle udtrykkene er klikbare, dvs. at man fx kan klikke på udtrykket *latinamerikanske danse* og blive ledt derhen i strukturen. Når man har læst forklaring m.m., har man mulighed for at klikke sig tilbage til resuméet i artiklens begyndelse.

I den trykte udgave af DDO er der ofte anvendt traditionelle strategier for at spare plads, fx brug af bindestreger eller tilde til at erstatte opslagsordet. I det digitale medium er pladshensynet ikke længere afgørende, så derfor har vi allerede i prototypen valgt at lade alle sammensætninger være skrevet helt ud, sammenlign sammensætningerne i tabel 1.

Trykt DDO	Prototype
danseform, -kunst, -undervisning; folke-, kind-, kæde-, mave-, par-, regn-, sportsdans	danseform, dansekunst, danseundervisning; folkedans, kinddans, kædedans, mavedans, pardans, regndans, sportsdans

Tabel 1. *Udsnit af artiklen* ***dans*** *(sammensætninger)*

Vi har ligeledes droppet tilden som erstatning for opslagsordet i ordforbindelser (kollokationer), sammenlign eksemplerne i tabel 2.

Trykt DDO	Prototype
moderne ~, klassisk ~; sang og ~, musik og ~, spille (op) til ~	moderne dans, klassisk dans; sang og dans, musik og dans, spille (op) til dans

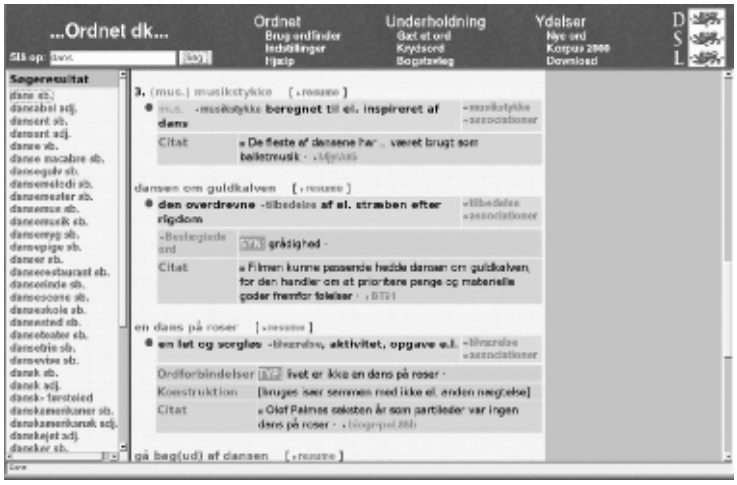
Tabel 2. *Udsnit af artiklen* ***dans*** *(kollokationer)*

Forøgelsen i pladsforbrug er beskeden, mens læsbarheden til gengæld vinder betragteligt ved afskaffelsen af bindestreg og tilde. I den endelige netversion vil man også kunne skrive sigler som SYN, ANT, JF helt ud og tilføje nye elementbetegnelser såsom *ordforbindelse*, *konstruktion* og *citat*.

Forkortelsesstrategien har også spillet en rolle ved udformningen af kildeangivelserne til citaterne (ca. 100.000 i alt), idet der i den trykte ordbog ved alle citater er anført en kilde, men i forkortet og komprimeret form. Idéen har været at forkortelsen skulle være så gennemskuelig som muligt (med mulighed for at slå forkortelsen op i ordbogens bind 6), men i prototypen er kildeangivelserne i stedet gjort klikbare så man ved et klik med musen får vist et lille vindue med mere detaljerede bibliografiske oplysninger. Det vil være en stor hjælp for mange (danske) brugere der måske nok kan afkode *BerlT90* (avisen Berlingske Tidende 1990), men vil have svært ved at regne ud hvad der gemmer sig bag fx *LDJørg86* eller *biogr.-pol.86b*.[1]

De faciliteter og de ændringer i forhold til den trykte bog som vi har nævnt indtil nu, anser vi for forholdsvis almindelige og forventelige. I det følgende vil vi omtale nogle mere specielle funktioner som vedrører den tidligere nævnte genus proximum-opmærkning og sammenkoblingen med korpustekster.

Vi vil igen se på artiklen **dans**, nærmere betegnet betydning 3, der er defineret som 'musikstykke beregnet til el. inspireret af dans', se figur 2.



Figur 2. *Udsnit af artiklen* ***dans*** *(betydning 3)*

Der er her knyttet to søgemuligheder til ordet *musikstykke*, der altså er genus proximum i denne definition. Klikker man på *musikstykke* i selve definitionen, bliver dette ord slået op i ordbogen. Hvis man klikker på *musikstykke* til højre for

[1] Henholdsvis Lone Diana Jørgensen: *Svanedans*, Gyldendal 1986 og Jørgen Poulsen: *Olof Palme*, Samleren 1986

definitionen, får man i venstre side vist en liste over alle ord der i ordbogen er defineret som 'musikstykke', dvs. har dette ord som genus proximum. Denne liste er genereret automatisk og sorteret efter hvilke ord der typisk optræder i definitioner hvor *musikstykke* er genus proximum. Som kriterium for typiskhed er brugt den velkendte mutual information-statistik, som først blev introduceret af Church og Hanks (1989), og som groft sagt lister ord efter hvor stor den indbyrdes tiltrækningskraft er. I dette tilfælde giver det os en liste hvor ord der betegner musikstykker, er ordnet efter hvilke karakteristiske træk (differentiae specificae) der indgår i deres definition. På denne måde får vi musikstykker der hører til danse, muntre musikstykker, musikstykker inddelt i satser, lyriske musikstykker, musikstykker for klaver osv. (se tabel 3).

Ord der betyder 'musikstykke'

Træk: **dansen**
- cha-cha-cha sb.
- polonæse sb.
- sarabande sb.
- samba sb.
- rumba sb.
- tarantel sb. bet. 2
- bolero sb.
- menuet sb.
- mazurka sb.
- mambo sb.
- hopsa sb.
- gigue sb.
- gavotte sb.
- tango sb.
- fandango sb.
- polka sb.
- czardas sb.

Træk: **muntert**
- capriccio sb.
- scherzo sb.
- humoreske sb.
- polka sb.
- bagatel sb.
- rondo sb.

Træk: **satser**
- divertimento sb.
- symfoni sb.
- suite sb. bet. 2
- serenade sb.
- sonate sb.
- klavertrio sb.

Træk: **sangstemmer**
- septet sb. bet. 2
- kvintet sb. bet. 2
- sekstet sb. bet. 2
- kvartet sb. bet. 2

Træk: **lyrisk**
- impromptu sb.
- romance sb. bet. 2
- nocturne sb.
- ballade sb. bet. 3

Træk: **rytme**
- cha-cha-cha sb.
- tarantel sb. bet. 2
- mazurka sb.
- menuet sb.
- tango sb.
- polonæse sb.
- sarabande sb.
- samba sb.
- rumba sb.
- mambo sb.
- vals sb.
- bolero sb.
- gavotte sb.
- gigue sb.
- hopsa sb.
- polka sb.
- fandango sb.
- czardas sb.
- sørgemarch sb.
- march sb. bet. 4

Træk: **klaver**
- klaverstykke sb.
- impromptu sb.
- præludium sb.
- novellette sb. bet. 2
- klavertrio sb.
- ballade sb. bet. 3
- toccata sb.

Træk: **instrumenter**
- trio sb. bet. 2
- oktet sb. bet. 2
- septet sb. bet. 2
- kvintet sb. bet. 2
- sekstet sb. bet. 2
- kvartet sb. bet. 2
- koncert sb. bet. 2

Træk: **orkester**
- ouverture sb.
- kammerkoncert sb. bet. 2
- dødsmesse sb.
- symfoni sb.
- ouverture sb. bet. 2
- passion sb. bet. 2
- kantate sb.
- solokoncert sb. bet. 2

Træk: **takt**
- galop sb. bet. 2
- siciliano sb.
- sørgemarch sb.
- march sb. bet. 4
- chaconne sb.
- passacaglia sb.

Tabel 3. *Uddrag af listen over ord med* musikstykke *som genus proximum*

Det er klart at der her er tale om et eksperiment: Listen er genereret og sorteret automatisk og ikke gennemgået af et menneske; fx er det ikke indlysende at korværker som dødsmesse, passion og kantate kun er placeret under *orkester* og ikke også under *kor*, som slet ikke optræder som signifikant træk.Vi synes dog alligevel den rummer perspektiver, og ved manuel tilretning vil resultatet blive langt bedre (jf. Asmussen 2004, Trap-Jensen 2004). Et næste skridt i arbejdet vil netop være at indbygge mulighed for begrebsorienterede søgninger, hvor udgangspunktet er begrebet og ikke ordet. Hvis man eksempelvis er interesseret i at finde ord for mørke øltyper, vil man kunne indtaste et genus proximum (*øl*) og et eller flere specifikke træk (*mørkt*), og man vil hurtigt få foreslået kandidater som *stout*, *lagerøl*, *porter*, *maltøl* og *nisseøl.*

En yderligere facilitet i prototypen er koblingen med korpus. Til højre for ordbogsartiklen findes nemlig en mulighed for at se mere om ordet i korpus. Det giver foreløbig en konkordans fra Korpus 2000 (http://korpus.dsl.dk): et korpus med i øjeblikket ca. 56 millioner løbende ord fordelt på to delkorpusser fra omkring 1990 og omkring år 2000. I prototypen er de to resurser adskilt fra hinanden, men vi ønsker at skabe en større integration mellem korpus og ordbog, en friere navigering mellem de to resurser. Fx kan en konkordans, fremkaldt af et ordbogsopslag, udløse ønsket om at slå op i ordbogen et nyt sted. I de definitioner eller eksempler man finder der, ser man måske ord eller udtryk der kan give anledning til en ny korpussøgning osv. Korpus 2000's interface tilbyder i dag visse grundlæggende korpusfunktioner som fx sortering på nøgleordet selv eller på dets venstre eller højre kontekst og lister over ord der er hyppige naboer til opslagsordet. I et nyt korpusinterface, som er under udvikling i DSL, regner vi med at tilbyde mere avancerede funktioner som fx lister der på grundlag af en syntaktisk opmærkning af korpus giver overblik over et ords typiske syntaktiske og leksikalske medspillere[2] og mulighed for søgning i brugerdefinerede delkorpusser som fx avistekster før 1995 eller talesprog fra radio og tv.

3. Omfang, lemmaudvælgelse og redigeringsprincipper

Overgangen fra papir til skærm ændrer på flere punkter betingelserne for redigering, og det vil betyde at de elektroniske artikler efterhånden vil ændre sig i forhold til bogværkets artikler. Præsentationen af oplysninger vil ændre sig: Ud over selve designet, som i sagens natur vil tage sig anderledes ud i netversionen, har vi allerede set at pladsøkonomi ikke er så vigtig ved netpublicering, og mange forkortelser, brug af tilde og bindestreger og andre pladsbesparende foranstaltninger vil derfor blive gjort overflødige. Vi har også tidligere set at betydningsresuméer i begyndelsen af

[2] Det der af Kilgariff et al. (2004) kaldes 'word sketches'.

artiklerne er nyttige til at skaffe overblik over en elektronisk artikel. I netversionen vil man også få mulighed for med teleskopløsninger enten at skjule visse oplysningstyper for at få større overblik eller udfolde oplysningerne, alt efter éns behov. Og endelig forestiller vi os at nogle oplysningstyper skal præsenteres på en anden måde end i bogværket. Det gælder især bøjningsoplysninger og syntaktiske oplysninger, som også af sprogteknologiske grunde bør gives en anden struktur i den elektroniske udgave (se Trap-Jensen, under udgivelse). Det samme gælder til en vis grad de etymologiske oplysninger.

På indholdssiden vil artiklerne også ændre sig, dels ved at det bliver muligt at bringe flere autentiske eksempler end det stramt redigerede udvalg som man finder i bogværket, dels ved at nye funktioner bliver tilbudt som kun er mulige i det elektroniske medie. Koblingen til korpus har været nævnt som noget vi prioriterer højt. Korpus skal udnyttes ved at brugeren ud over den generelle mulighed for at slå det tilsvarende ord op i korpus også skal have adgang til nogle specifikke oplysninger baseret på korpusundersøgelser der knytter sig til den artikel brugeren befinder sig i. Fx kan det være interessant at se fordelingen i korpus af dobbeltformer eller alternative bøjningsformer, eller at se autentiske eksempler på de syntaktiske mønstre der gives i artiklerne. Begge dele kan lade sig gøre fordi korpusteksterne er blevet opmærket både morfologisk og syntaktisk.

Det andet område vi prioriterer højt, er muligheden for at lave begrebsorienterede søgninger. I gennemgangen af prototypen viste vi hvordan ordbogsmanuskriptet i forvejen indeholder mange af de forbindelser mellem ordene som man er interesseret i at få oplyst når man søger efter begrebet snarere end betegnelsen, en såkaldt onomasiologisk søgning. Ud over over-/underbegrebsrelationen bestemt ved genus proximum kan man finde oplysning om synonymi, antonymi, del-helhed og flere andre relationer. I prototypen kan man på baggrund af genus proximum-opmærkningen finde mange interessante begrebsområder, men det er et problem at systemet af begrebsområder ikke har været behandlet ud fra en konsistent, ontologisk synsvinkel. Ordbogsartiklerne er nemlig først og fremmest skrevet til mennesker og ikke til en computer. En artikel kan derfor fungere fint som oplysning til en menneskelig bruger, men hvis den forsynder sig mod det ontologiske hierarki, vil computeren ikke registrere ordet som en del af det pågældende begrebsområde. I DDO er en *hare* fx defineret som 'et gråbrunt pattedyr med hvid bug, to store fortænder i over- og undermunden, lange ører og kraftige bagben', hvilket fungerer fint for en menneskelig bruger, men udelukker ordet fra en søgning på begrebet gnaver. Og når en *benvarmer* er defineret som 'en slags ulden strømpe uden fod til at trække på benet', er det så underbegreb til strømpe eller til beklædningsgenstand? Den slags problemer kommer man hurtigt ud i hvis de begrebsorienterede søgninger foretages direkte på artiklernes genus proximum-opmærkning. Vi har derfor indset at der kræves temmelig meget redigering af ordstoffet, og har derfor indledt et samarbejde

med Center for Sprogteknologi ved Københavns Universitet om at lave et decideret leksikalsk-semantisk ordnet for dansk på baggrund af artiklerne i DDO, efter modellen fra de internationale *wordnets*. Det danske projekt kalder vi DanNet (se www.wordnet.dk), og tanken er at det skal kunne integreres i *ordnet.dk* og få de begrebsmæssige søgninger til at fungere hensigtsmæssigt – foruden at det naturligvis vil få en række selvstændige, sprogteknologiske anvendelser ligesom søsterprojekterne for en række andre sprog.

Hvor mange ordbogsartikler man vil kunne slå op i *ordnet.dk*, ved vi ikke helt sikkert endnu. Under indtryk af den kritik der har været rettet mod den trykte DDO,[3] vil vi dog godt oplyse at det er planen at både denne ordbogs ca. 60.000 hovedopslagsord og de ca. 40.000 underopslagsord (primært sammensætninger og afledninger) skal kunne slås op som selvstændige artikler. Derudover vil der som sagt blive nyskrevet et antal helt nye artikler, især om nye ord der kommer til i sproget, men ideelt set vil det også være nyttigt at supplere ordstoffet med andre typer artikler. Det kan være ord som blev sorteret fra under redigeringen af DDO (faglige eller forholdsvis sjældne ord), eller ord fra tiden mellem ODS' afslutning og DDO's primære periode (dvs. fra ca. 1950 til ca. 1980) som ikke er repræsenteret i DDO's korpus. Der er også en vis usikkerhed om lemmabestanden i ODS. Den første netudgave indeholder 180.000 opslagsord, men hertil skal muligvis lægges en række sammensætninger der gemmer sig nede i artiklerne – sammensætninger som vi overvejer at gøre søgbare – og i hvert fald de nye artikler der er indeholdt i supplementsbindene. Og endelig ved vi endnu ikke hvor stor fællesmængden mellem de to ordbøger er.

For Korpus 2000's vedkommende er planen som nævnt ovenfor at udarbejde en ny brugergrænseflade hvor flere nye og forbedrede faciliteter stilles til rådighed, herunder ikke mindst udnyttelse af teksternes syntaktiske opmærkning. Det er også tanken at man skal kunne søge mere fleksibelt i korpus, fx ved at kunne vælge at søge i samtlige tekster eller kun relevante udsnit. Endelig vil mængden af tekster løbende blive forøget så også tekstbasen holdes opdateret. Det er det der på længere sigt vil gøre det muligt at foretage diakrone undersøgelser af sprogets udvikling, og redaktionsinternt vil det være et nyttigt redskab til at generere automatiske lister over nye lemmakandidater.

4. Ordbog over det danske Sprog på internettet

Som nævnt i indledningen drejer en væsentlig del af ordnet-projektet sig om at gøre Ordbog over det danske Sprog (ODS) tilgængelig i et elektronisk format, ligesom andre tilsvarende store nationalordbøger er blevet digitale, fx Svenska Akademiens ordbok (SAOB) i Sverige, Oxford English Dictionary (OED) i den engelsktalende

[3] Jf. bl.a. Bergenholtz og Vrang 2004.

verden og Grimms Deutsches Wörterbuch i Tyskland. Norsk Ordbok og Norsk Riksmålordbok er også trådt ind i den digitale tidsalder, jf. Guttu 2005 og Grønvik og Tvedt 2006.

I den indledende fase skulle det besluttes om bogværket skulle scannes eller indtastes. Begge modeller er blevet brugt af andre store ordbogsprojekter, fx scanning af SAOB og indtastning af OED, men vi besluttede at lægge os op ad den model der blev brugt i forbindelse med Grimms Deutsches Wörterbuch. Den går ud på at manuskriptet indtastes i to uafhængige versioner, der siden sammenlignes elektronisk. På den måde kommer man så godt som samtlige tastefejl til livs uden en lang og sej korrekturlæsning, der ellers ville være nødvendig ved scanning.

Denne del af arbejdet foregår i samarbejde med en afdeling ved universitetet i Trier ved navn Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften (se http://germazope.uni-trier.de/Projects/KoZe2). Først er bogen tastet ind fysisk, hvilket er foregået i et firma i Kina der har specialiseret sig i denne type opgaver og også stod for indtastningen af Grimm. Efter indtastningen er de to parallelle versioner så sammenlignet automatisk i Trier, idet uoverensstemmelserne er blevet bearbejdet manuelt, dvs. at en person med kompetence i germanske sprog har slået op i ODS for at se hvad der skulle stå.

Denne fase af arbejdet er nu afsluttet. Vi har modtaget filer hvor alle 28 bind er indtastet og den automatiske sammenligning er foretaget. Teksten til ODS foreligger altså nu i et rent tekstformat spækket med typografiske koder, og i figur 3 kan man få et indtryk af hvordan det ser ud.

```
$0040.12 <P>__<A+1>#F+herske-syg,#F-</A+1> #/+adj. meget, alt for be-
$0040.13 g#.^orlig efter at herske (jf.#/- -k#.^ar, -lysten#/+).
$0040.14 vAph.(1759).#/- Emilie var herskesyg og #.on-
$0040.15 skede en mere glimrende Stilling i Sta-
$0040.16 rostens Huus.#/+Hauch.I.265.#/- Det er en Ytring
$0040.17 af Tyranni og herskesygt Sindelag. #/+PM#.oll.
$0040.18 II.388.#/- det at hj#.^alpe #/+(er)#/- ikke . . at v#.^are
$0040.19 den Herskesygeste men den Taalmodigste.
$0040.20 #/+Kierk.XIII.533.#/- <A+1>#F+-syge,#F-</A+1> en. #/+det at v#.^ore
$0040.21 herskesyg; magtbrynde (jf.#/- -lyst#/+). Moth.H
$0040.22 173.#/- Alexander Magnus . . havde og store
$0040.23 Fejl. Hans Herske-Syge alleene . . kand
$0040.24 tiene til Beviis derpaa. #/+Holb.Ep.III.168.
$0040.25 Bagges.L.I.279. Brandes.VIII.151.#/-</P>
```

Figur 3. *Artiklerne **herskesyg** og **herskesyge** i indtastet format*

Det drejer sig om artiklerne **herskesyg** og **herskesyge**, og som det fremgår, kan det godt være vanskeligt at læse for mennesker. Til gengæld er det typografiske billede

repræsenteret fuldt ud: Hver linje i bogen indledes med en kode der oplyser om spaltenummer og linjenummer; alle typografiske markeringer såsom skriftstørrelse, fed, kursiv, spatiering er gengivet, ligesom alle specialtegn, dansk æ, ø og å samt de små symboler der brugtes dengang i form af ankre, noder, fugle, tandhjul osv., har hver sin specifikke kode.

Der er endnu ikke foretaget nogen systematisk kontrol af indtastningen, men alting tyder på at der er meget få indtastningsfejl, så det er vi meget tilfredse med. Der er dog ting i den trykte bog som har været vanskelige at afkode både for de kinesiske indtastere og ved den efterfølgende sammenligning i Trier. Disse tvivlstilfælde er markeret i filerne, og der vil være henved 2.000 steder hvor vi skal ind og vurdere og rette individuelt.

Der er bl.a. tale om steder hvor vi som indfødte danskere umiddelbart kan se hvad der skal stå, mens det ikke har været muligt for en ikkedansker, fx tilfælde hvor mellemrum mellem ord er blevet meget små eller nærmest væk. En bestemt kildeangivelse er indtastet på disse to måder i filerne:

Ris paa150Maader
Rispaa150Maader

for sådan ser det faktisk ud i bogen, men en person med modersmålskompetence ser straks at det rigtige ser sådan ud:

Ris paa 150 Maader

I andre tilfælde er der "flueklatter" i satsen, hvor indtasterne ikke har vidst om klatten hørte med til bogstavet, fx et n med flueklat over, der mest ligner et n med tilde. I andre tilfælde igen er satsen helt udvisket, og man kan som dansker måske rekonstruere den, men ellers må man søge tilbage til ordbogssedlerne hvis der er tale om et citat. Et eksempel på dette er et indtastet:

Løvemank{??}{?A}

som bør være

Løvemanken

Den næste fase i arbejdet skal foregå hos DSL, og det er den strukturelle fortolkning af de typografisk opmærkede data. Det er den fase vi skal til at begynde på, og vi ved endnu ikke hvilke problemer vi står over for, men vores forventning er at vi kan nå meget langt ved hjælp af de typografiske signaler der er givet videre fra bogformatet.

Vi regner med at gå trinvis frem sådan at vi kan offentliggøre en foreløbig version i efteråret 2005, en version hvor i hvert fald opslagsordene er opmærket, og hvor præsentationen er gjort mere overskuelig med flere linjeskift og brug af farver. Det er selvfølgelig primitivt og slet ikke endemålet, men for mange brugere af ODS vil det være et kæmpe fremskridt.

I en senere fase vil vi prøve at komme dybere ned i strukturen og opnå en finere opmærkning af indholdet i de forskellige oplysningstyper. En anden hovedopgave, som bestemt ikke må undervurderes, er sammenfletningen af ODS og de 5 supplementsbind. Fordelene ved supplementsbindene er at de allerede foreligger i elektronisk form, omend i et andet format end hovedværket; til gengæld er oplysningerne i supplementet ganske heterogene – der kan være tale om elementer der skal ind på alle niveauer i hovedværkets struktur: helt nye artikler, tilføjelse af betydninger eller citater, ændring eller sletning af eksisterende oplysninger. I første omgang kunne man tænke sig at supplementets artikler og andre oplysninger blot bringes efter hovedværkets, men i samme skærmbillede (ligesom man har gjort i OED med de nye oplysninger der endnu ikke er integreret i de eksisterende artikler). Senere vil man, formentlig manuelt, skulle ind og placere oplysningerne det rigtige sted i strukturen.

5. *ordnet.dk* og fremtiden

Med *ordnet.dk* er Det Danske Sprog- og Litteraturselskab for første gang gået ind i at publicere ordbøger elektronisk. Men det behøver ikke slutte med ODS og DDO. Selskabet råder over flere resurser både mht. ordbøger og tekster: En ordbog over gammeldansk er under udarbejdelse, Holberg-Ordbog I-V udkom 1981-1988, og et elektronisk renæssanceprojekt der indledes fra 2006, omfatter foruden tekstudgivelser på både latin og dansk også begyndelsen til en ordbog over ældre nydansk. På tekstsiden råder DSL i kraft af projektet *Studér Middelalder på Nettet* (http://smn.dsl.dk) over en række centrale danske middelaldertekster i digital form. I det elektroniske *Arkiv for Dansk Litteratur* (www.adl.dk) findes desuden et meget stort antal tekster af klassiske danske forfattere, baseret på DSL-udgaver, som ville være oplagt at bruge som tekstkorpus for en overvejende litterær ordbog som ODS. Selvom det altså stadig er i afdelingen for visioner og ønsketænkning, finder vi det ikke helt urealistisk at forestille sig *ordnet.dk* udvidet til engang i fremtiden at omfatte både tekstudgaver og leksikografiske beskrivelser af dansk sprog lige fra runerne til i dag.

Litteratur

Ordbøger:

DDO = *Den Danske Ordbog. Bind 1-6*. Udgivet af Det Danske Sprog- og Litteraturselskab. Hovedredaktører: Ebba Hjorth og Kjeld Kristensen. København: Gyldendal 2003-2005.

ODS = *Ordbog over det danske Sprog. Bind 1-28*. Udgivet af Det Danske Sprog- og Litteraturselskab. Grundlagt af Verner Dahlerup. Ledende redaktører: H. Juul-Jensen og Jørgen Glahder. København: Gyldendalske Boghandel, Nordisk Forlag 1918-1956.

Anden litteratur:

Asmussen, Jørg 2004: Feature Detection – A Tool for Unifying Dictionary Definitions. I: Williams, Geoffrey and Vessier, Sandra: *Proceedings of the 11th EURALEX International Congress*. Lorient. 63-69.

Bergenholtz, Henning/Vibeke Vrang 2004: Den Danske Ordbog imponerer og skuffer. I: *Hermes 33,* 149-178.

Church, Kenneth og Hanks, Patrick 1989: Word association norms, mutual information and lexicography. I: ACL Proceedings, 27th Annual Meeting, Vancouver. 76-83.

Grønvik, Oddrun og Tvedt, Lars Jørgen 2006: Norsk Ordbok 2014 – presentation av eit komplekst leksikografisk verktøy. I denne rapport.

Guttu, Tor 2005: Videreføringen av *Norsk Riksmålordbok* – om redigeringsarbeidet. I: Fjeld, Ruth V. og Worren, Dagfinn (red.): *Nordiske Studiar i Leksikografi 7. Rapport frå Konferanse om leksikografi i Norden, Volda 20.-24. mai 2003*. Oslo. 166-174.

Kilgariff, Adam; Rychly, Pavel; Smrz, Pavel; Tugwell, David 2004: The Sketch Engine. I: Williams, Geoffrey and Vessier, Sandra: *Proceedings of the 11th EURALEX International Congress*. Lorient. 105-115.

Trap-Jensen, Lars 2004: Et net af ord – *ordnet.dk*. I: *Mål & Mæle 4*, 20-26.

Trap-Jensen, Lars, under udgivelse: Making Dictionaries for Paper or Screen: Implications for Conceptual Design. I: *Proceedings of the 12th EURALEX International Congress*. Torino 2006.

Websider:

ADL = *Arkiv for Dansk Litteratur.* Det Kongelige Bibliotek og Det Danske Sprog- og Litteraturselskab. København. www.adl.dk

DanNet. Center for Sprogteknologi og Det Danske Sprog- og Litteraturselskab 2005-2008. København. www.wordnet.dk

Korpus 2000. Det Danske Sprog- og Litteraturselskab. København. http://korpus.dsl.dk

Studér Middelalder på Nettet. Det Danske Sprog- og Litteraturselskab. København. http://smn.dsl.dk