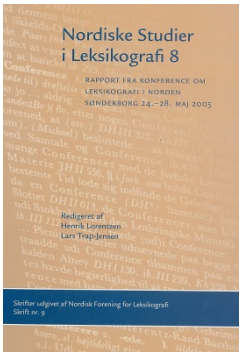
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Den danske Sprogteknologiske Ordbase og dens anvendelse i værktøj til leksikografiske formål

Forfatter: Anna Braasch

Kilde: Nordiska Studier i Leksikografi 8, 2006, s. 25-38
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.

Anna Braasch

Den danske Sprogteknologiske Ordbase og dens anvendelse i værktøj til leksikografiske formål

The Danish Lexicon for Language Technology Applications (STO) and its use in a tool for lexicographic purposes. This article deals with the largest and most comprehensive computational lexicon for Danish. Firstly, the development principles, the lexical coverage and the linguistic content of this lexicon are presented. This part focuses on the treatment of inflectional morphology by means of the Remove/Add computing method. Secondly, the development and functionalities of a flexible and effective lemmatiser program for Danish are discussed; the rules of the lemmatiser have been derived from the STO morphology data. A few examples illustrate the use of the lemmatiser in solving lexicographic tasks. Finally, the user-interface for online look-ups in the STO database is described: it transforms the computational lexicon into an electronic dictionary making it a useful source of lexical knowledge for lexicographers and other interested users. Also a number of useful web addresses, viz. to the STO database, the lemmatiser and relevant documentation, also in English, are provided.

1. Indledning

Den Sprogteknologiske Ordbase (STO) for dansk er udviklet til anvendelse som ordbogskomponent i programmer til datamatisk sprogbehandling, fx morfologisk eller syntaktisk analyse af tekster og applikationer hvori der indgår sådanne analyser. Dermed kan en sådan ordbase også udgøre kernen i nogle programmer som med fordel kan bruges i leksikografers og lingvisters arbejde, eksempelvis i et værktøj der automatisk identificerer lemmaer i en tekst, en såkaldt lemmatiser. For at opnå de bedst mulige resultater er det vigtigt at værktøjernes ordbogsmodul har en vis størrelse, og at det er leksikografisk og lingvistisk velfunderet.

STO er allerede blevet anvendt som leksikonmodul i flere værktøjer, foruden at den også er blevet brugt i en række lingvistiske forskningsprojekter, endda inden arbejdet med basen var afsluttet (fx Ørsnes 2004). For interesserede er der via internettet adgang til at søge i store dele af ordbasen, og der registreres allerede nu mange opslag – op til flere tusinde om ugen, hvilket tyder på stor almen interesse.

I det følgende beskrives først STOs indhold med hensyn til dens ekstensionelle dækning (antallet og arten af lemmaer) og intensionelle dækning (oplysningstyper). Derefter præsenteres lemmatiseringsværktøjet for dansk der er udviklet til det leksikografiske arbejde i STO-projektet, og som nu er tilgængelig for andre projekter. Til sidst beskrives kort hvordan denne ordbase der, selv om den er udarbejdet til datamatiske formål, også kan anvendes som elektronisk ordbog over internettet ved hjælp af en brugergrænseflade.

2. STO som dataleksikografisk produkt

STO er, som nævnt i indledningen, en ordbase der er udviklet til anvendelse i sprogteknologisk, datalingvistisk og dataleksikografisk forskning. Ordbasens materiale er korpusbaseret både med hensyn til lemmaselektionen og til beskrivelsen af ordenes egenskaber. De metodiske overvejelser der ligger til grund for basen med hensyn til dens indhold, struktur og beskrivelse, er styret af de planlagte anvendelser.

Det er fælles for elektroniske ordbøger og datamatiske ordbaser (også kaldet *leksikon*) at de beskriver et nærmere afgrænset ordforråd i overensstemmelse med faste leksikografiske retningslinjer. En computer kan ikke uddrage oplysninger fra eksempler ved hjælp af analogier eller på anden måde udnytte tekstlig information (henvisninger, citater, forklaringer osv.), og derfor stilles der andre krav til en ordbase til datamatisk anvendelse end til en ordbog for mennesker. Forskellen mellem en elektronisk ordbog og en ordbase består primært i at den leksikografiske beskrivelse i et leksikon skal være meget mere detaljeret og opdelt i dens mindste bestanddele. Desuden skal den være formaliseret i et fast beskrivelsessprog og struktureret i klare oplysningstyper. Den skarpe grænse mellem de to typer leksikalske datasamlinger udviskes dog noget ved at der benyttes flere og flere datamatiske metoder og værktøjer i den traditionelle leksikografi som påvirker både arbejdet og produktet. Omvendt genbruges data fra ordbøger ved udarbejdelsen af ordbaser, foruden at leksikografens håndværk vinder indpas i leksikografi for datamater.

På adressen http://cst.dk/sto/referencer/index.html er der en række artikler om STO som beskriver forskellige aspekter i projektet. Yderligere information fås ved direkte henvendelse til Center for Sprogteknologi, Københavns Universitet.

2.1. Ordforrådet i STO

Ved udarbejdelsen af STO er der blevet lagt vægt på at medtage et bredt sammensat udvalg af ord. Ordbasen indeholder i alt mere end 81.500 lemmaer, fortrinsvis fra almensprog. En stor hjælp i selektionen af de 68.000 almensproglige lemmaer var Den Danske Ordbogs (DDO) foreløbige, frekvensbaserede lemmaliste som blev stillet til rådighed for arbejdet i 2001. De resterende ca. 13.500 lemmaer stammer fra fagsproglige tekster, men de er ikke egentlige eksperttermer. De seks udvalgte fagområder var edb/it, miljø, sundhed/helse, finans/økonomi, forvaltning samt handel/erhverv. Formålet med at inddrage fagrelaterede tekster fra internettet var at udbrede ordbasens ekstensionelle dækning sådan at ordforrådet også omfatter almene fagord, også kaldet gråzone-ord, der ligger tæt op ad det almensproglige ordforråd. En detaljeret oversigt over sammensætningen af ordforrådet og statistik over de enkelte ordklasser mv. kan ses på hjemmesiden www.cst.dk/sto.

2.2. Lingvistiske oplysninger i STO

2.2.1. Principper

Udgangspunktet for valget af oplysninger i den lingvistiske beskrivelse er styret af den datalingvistiske tilgang til behandling af sprog; denne opererer traditionelt med to typer grundmoduler, nemlig et ordbogsmodul (det såkaldte leksikon) og et grammatikmodul. Den traditionelle, skarpe grænse mellem grammatikken (der beskriver *generelle* – grammatiske – regler) og ordbogen (der indeholder ordene og beskrivelsen af deres *individuelle* egenskaber) udviskes dog mere og mere. Siden 1990'erne er oplysninger der beskriver ordenes syntaktiske konstruktionspotentiale, medtaget i den leksikalske beskrivelse; dette gælder i stigende grad også deres semantiske kompatibilitet. Det grundlæggende princip er at reglerne i et sprogteknologisk systems grammatik og oplysningerne i dets ordbog tilsammen skal udgøre en samlet formaliseret beskrivelse og dermed dække det ønskede segment af sproget. Dette princip stammer fra de leksikalistiske syntaksteorier, herunder den såkaldte Head-Driven Phrase Structure Grammar (HPSG, se Pollard & Sag 1994). Det udmønter sig i dag i den såkaldte leksikalisme ("the lexicalist approach"), der indebærer at en ordbase bør indeholde mange og detaljerede oplysninger om syntaktiske generaliseringer (Ørsnes 2004:213). Derved bliver grammatikken i stor udstrækning integreret i den leksikalske beskrivelse. Sådanne generaliseringer er fx den regelbundne dativalternation (1a), eller aktiv/passiv-alternationen (1b), som i traditionel lingvistik beskrives af grammatikken.

(1a) Marie gav Peter et kys/Marie gav et kys til Peter
(1b) Peter betaler udgifterne/Udgifterne betales af Peter

Mange moderne ordbøger for mennesker, som for eksempel Den Danske Ordbog (DDO, 2003-05), følger også denne tendens og opererer med såkaldte skabeloner eller konstruktionsmønstre, som er en slags formaliseret beskrivelse af ordets typiske nærkontekst, samtidig med at traditionelle brugseksempler illustrerer ordets konstruktionsmønstre.

2.2.2. Oplysningstyperne

STO indeholder en lang række strukturerede og formaliserede oplysninger fordelt på tre beskrivelseslag: morfologi (for hele ordforrådet), syntaks (for mere end 45.000 lemmaer, udvalgt efter frekvens) og semantik (for en mindre del af ordforrådet, ca. 8.000 lemmaer med i alt 10.000 læsninger) til eksperimentelle formål.

I det følgende beskrives først en del af det morfologiske lag i detaljer, om end ikke udtømmende. Derefter gives et overblik over oplysningerne der hører til det syntaktiske lag.

Administrative og andre ikke-lingvistiske oplysningstyper bliver ikke omtalt her.

Grunden til den detaljerede præsentation af de bøjningsmorfologiske oplysninger er at disse danner basis for det lemmatiseringsværktøj der skildres i afsnit 3. Præsentationen i nedenstående afsnit 2.3 og dets underafsnit er i høj grad baseret på dokumentet STOs Lingvistiske Specifikationer (Braasch et al. 2004-2005).

2.3. Morfologi

Den samlede morfologiske beskrivelse af et lemma er fordelt på flere blokke der indeholder hver sin type oplysninger som vedrører lemmaets ordklasse, stavning, bøjning, sammensætning (kun for substantiviske komposita) og "autonomi" (med værdien "NO" angives at ordet kun forekommer i faste udtryk som fx *[gå i] skuddermudder*).

2.3.1. Ordklasseangivelserne

STO følger på dette punkt i alt væsentligt Retskrivningsordbogens (RO 2001) ordklasseinddeling med ganske få undtagelser. En sådan undtagelse er at STO behandler talord som adjektiver, med subkategorierne "cardinal", fx *fem* og "ordinal", fx *femte*. Desuden er der indført en kategori "unique" der dækker over subkategorierne formelt subjekt (*der*), infinitivmarkør (*at*) og lemmaet *som* i ikke-konjunktionsfunktionen.

2.3.2. Stavning

Hvis et ord har flere stavemåder, er disse anført i ordbasen. I visse tilfælde er også enkelte alternative stavemåder, der ikke er godkendt i RO 2001, medtaget. Begrundelsen herfor er følgende (jf. Braasch & Olsen 2005):

> Det drejer sig først og fremmest om stavemåder der har ændret status fra godkendt til ikke-godkendt eller omvendt i de seneste udgaver af RO. Da STO skal kunne bruges til automatisk genkendelse [af ord i tekster], og da dette nødvendigvis også må omfatte ældre tekster end den sidste udgave af RO, er der i STO medtaget former i overensstemmelse med RO 86 og frem. Disse former mærkes som ikke-godkendte. Andre ikke-godkendte stavemåder i STO er fx 'canarisk' og 'sclerose' som begge er uhyre hyppigt forekommende. Også bøjningsmønstre kan være ikke-godkendte i tilknytning til visse ord, som er ligeledes meget hyppige, som fx 'test' som har aldrig måttet bøjes med '-s' i pluralis, formen forekommer ikke desto mindre i mange tekster, og derfor er den medtaget i STO.

2.3.3. Bøjningsoplysninger

I den morfologiske beskrivelse er det væsentligste krav at beskrivelsesapparatet skal kunne rumme og håndtere alle danske bøjningsformer samt andre relevante, morfologirelaterede oplysninger. Dette krav er opfyldt ved at fastlægge det tilstrækkelige sæt af træk med tilhørende værdimængder, uanset om der er tale om regelmæssig eller uregelmæssig bøjning. Der opereres med morfemer (typerne *rod*

og *endelse*) og individuelle produktionsregler (af typen "fjern/tilskriv") til beregningen af de enkelte ordformer.

Set fra en datalingvistisk synsvinkel er denne metode meget effektiv da den sikrer en ensartet og økonomisk håndtering af den samlede bøjningsmorfologi. Et væsentligt punkt er at ord der traditionelt anses for at have uregelmæssig bøjning, i STO håndteres på en meget enkel måde helt på linje med regelmæssig bøjning.

Bøjningsoplysningerne er udtrykt i bøjningsmønstre. Hver ordklasse har sin specielle kombination af formbestemmende træk, og for hvert træk er der defineret en liste af relevante værdier. Hvert mønster er unikt og omfatter ordets sammenhørende bøjningsformer. De enkelte bøjningsformer produceres ved hjælp af de såkaldte *beregningsregler*. Et ord kan naturligvis have mere end et bøjningsmønster, som fx ordet *tallerken*, med former uden (2a) eller med synkope (2b og 2c):

(2a) med beregningsreglerne (+en, +er, +erne) => tallerkenen/tallerkener osv.
(2b) med beregningsreglerne (+en, [en]ner, [en]nerne) => tallerkenen/tallerkner osv.
(2c) med beregningsreglerne ([en]nen, [en]ner, [en]nerne) => tallerknen/tallerkner.

En beregningsregel tager udgangspunkt i grundformen og udpeger *roden* ved at angive hvad der skal fjernes fra grundformen (notation i [ ]), og hvad der derefter skal tilføjes for den pågældende form. Det specielle ved systemet er altså at begrebet *rod* her er forstået som den absolut længste del af et ord som er uforandret i den beskrevne bøjningsform. Således får man et ords operationelle rod ved at fjerne, begyndende bagfra, den del af ordet der ændres. I dette beskrivelsessystem er der ingen lingvistisk baserede regler for hvad *rod* er (det samme gælder begrebet *bøjningsendelse*); man går rent formalistisk til værks. Det er – set fra datamatisk synspunkt – en udmærket måde at håndtere ord med uregelmæssig bøjning på, nemlig på lige fod med de regelmæssigt bøjede, men det afviger fra den traditionelle lingvistiske anskuelsesmåde.

Nedenfor gives der et eksempel fra STOs Lingvistiske Specifikationer (op.cit.) på konstruktion af substantivers bøjningsmønstre, som omfatter de ordklassespecifikke egenskaber. I et mønster håndteres foruden bøjningsendelserne alle formrelaterede egenskaber, såsom synkope, fordobling af stamkonsonant og ændring af stammevokal, idet de inkluderes i beregningsreglerne.

Eksempelvis beskrives ordet 'tid' med mønster *MFG0016* som kombinerer oplysningerne om ordklasse (substantiv), køn (fælleskøn), bøjningsendelser for tal og bestemthed i umarkeret kasus *(+0, +en, +er, +erne)* som lægges til ordets rod der i dette tilfælde er identisk med opslagsordets grundform. Kasusendelsen (+*s*) for genitiv tilskrives hver af de fire nævnte former hvilket giver i alt 8 former. På denne måde laves et nyt mønster for hver unik kombination af træk/værdi-par; og der laves kun én enkelt udtømmende beskrivelse (ét mønster) af hver kombination. Der refereres under opslagsordet ved et

> nummer til det passende mønster […]. Mange substantiver bøjes på samme måde som 'tid', [fx 'stol', 'citron'] dvs. mønstret har et stort antal forekomster, såkaldte instantieringer.

Et mønster beskriver i de fleste tilfælde mange ords bøjning, mens en række mønstre kun har en enkelt eller nogle få instantieringer. Det er dem den traditionelle morfologi beskriver som undtagelser, fx *barnebarn* med flertalsformen *børnebørn*. Her fjernes intet fra grundformens rod (*barn*) for at danne formen ental/bestemt/genitiv, blot tilføjes *-ets*, hvorimod der for at generere formen flertal/ubestemt/umarkeret kasus skal fjernes *-arnebarn* og tilføjes *-ørnebørn*. I dette mønster er roden for pluralis ligeledes reduceret til et enkelt bogstav, *b*. Mønstret har kun denne ene instantiering, og i traditionel leksikografi håndteres ordet som særtilfælde eller undtagelse. En væsentlig fordel ved den her anvendte metode er, som allerede nævnt, at alle ord håndteres vha. samme mekanisme, hvilket sikrer en enkel og ensartet processering i forbindelse med praktiske anvendelser, eksempelvis i et lemmatiseringsværktøj. Der er naturligvis også en vis ulempe for leksikografen, nemlig det store antal bøjningsmønstre det er nødvendigt at etablere før systemet er fuldt udviklet.

2.3.4. Sammensætningsoplysninger

STO-basen indeholder mange afledte og sammensatte opslagsord. Afledte ord håndteres med samme mekanisme som simple, usammensatte ord, uden oplysning om orddannelse. På den anden side er håndteringen af sammensætningsmorfologi vigtig for dansk da den mest produktive metode for dannelse af nye ord netop er sammensætning af substantiver. Derfor er STO-materialet også forberedt til dynamiske anvendelser hvor eksempelvis nye substantiviske sammensætninger kan genkendes hvis de består af ord der er kodet i basen. Ordbasen indeholder nedenstående to oplysningstyper vedrørende substantiviske sammensætningers morfologi.

Fugeelement i sammensætninger

Når et ord indgår som førsteled i sammensætninger, er der tre muligheder mht. hvordan dets form er i sammensætningen, jf. eksemplerne nedenfor: (3a) ordet forbliver uændret i sammensætninger; (3b) ordet afkortes i sammensætninger; (3c) ordet får tilføjet et fugeelement[1]

(3a) lampe => lampe[0]fod, lampe[0]skærm
(3b) maskine => maskin[e]mester, maskin[e]oversættelse
(3c) afdeling => afdeling+s+leder, afdeling+s+sygeplejerske

[1] Aage Hansen (1967) kalder dette *sammenbindingselement*, i RO 2001 hedder det *bindebogstav.*

Et simpleksord kan have et eller flere forskellige fugeelementer som alle registreres og udtrykkes i overensstemmelse med fjern/tilskriv-metoden (jf. beregningsreglerne ovenfor). Oplysningerne om fugeelementer ved fx ordet *mand* formuleres således på følgende måde:

(4a) mand + 0 => mandtal
(4b) mand + e => mandeår
(4c) mand + s => mandsperson.

Denne oplysningstype er i høj grad korpusbaseret, og den er registreret mere udførligt og systematisk i STO end i Retskrivningsordbogen (jf. RO 2001, Indledningens afsnit 7.)

Dekomponering
Dekomponering er markering af et kompositums (sammensætning) primære bestanddele. Det foretages kun på det øverste niveau, nemlig i to dele: førsteled og sidsteled, også i de tilfælde hvor førsteleddet i sig selv er et kompositum (5a). Et kompositums led markeres ved at sætte '+' mellem leddene og mellem led og fugeelement (5b). I de tilfælde hvor noget fjernes, markeres det på samme måde som i bøjningsmønstrenes beregningsregler, som fx ved lemmaet *arbejdsfordeling* (5c).

(5a) urtepotte + skjuler
(5b) stat + s +sikkerhed
(5c) arbejde + [e]s + fordeling

Der gælder to principielle betingelser mht. om en sammensætning dekomponeres eller ikke. For det første dekomponeres kun sådanne sammensatte ord som består af to dele der hver især er et selvstændigt ord. For det andet skal begge led beholde deres oprindelige betydning. Derfor dekomponeres ord som *makroøkonomi* og *urmager* ikke.

2.4. Syntaks
Kernen i den syntaktiske beskrivelse er valensmønstret. Det indeholder oplysningerne om hvor mange led der knytter sig til ordet (aritet), hvorvidt leddene er obligatoriske eller ikke, hvilken syntaktisk funktion (fx subjekt, objekt) og hvilken syntaktisk kategori det enkelte led har (fx nominal- eller præpositionssyntagme inkl. den styrede præposition, eller en ledsætning). Derudover er der en række oplysninger som fx vedrører verbers refleksivitet, partikel og brug af hjælpeverbum. Desuden er der korpuseksempler som belyser hver syntaktisk konstruktion. Eksemplerne kan ikke bruges af maskiner, men er medtaget for at lette leksikografens arbejde. Der er to typer eksempler, den ene type er standardeksemplet der knytter sig til en given

konstruktionstype. Det indeholder ikke selve lemmaet, det eksemplificerer blot den type nærkontekst som lemmaet kan indgå i. Den anden type er det individuelle eksempel, med lemmaet i den pågældende konstruktion. Standardeksempler er fortrinsvis anvendt i tilfælde af simple konstruktionstyper, eksempelvis ved monovalente substantiver. Hvis lemmaet indgår i en kompleks konstruktion med flere valensbundne led fra forskellige syntaktiske kategorier, er der i de fleste tilfælde indsat et individuelt eksempel. Dette er en fordel for brugeren når han/hun slår op i databasen. En mere detaljeret redegørelse over de syntaktiske træk kan findes bl.a. i Braasch & Pedersen (2002) og i STOs Lingvistiske Specifikationer (op.cit.)

De syntaktiske oplysninger anvendes eksempelvis i automatisk sætningsanalyse, den såkaldte parsning. Parsning benyttes bl.a. som delproces i leksikografiske værktøjer, eksempelvis til at genkende, opmærke og registrere et ords grammatiske strukturer i et tekstkorpus forud for ordets leksikografiske beskrivelse. Et sådant værktøj for engelsk der kan udtrække et ords såkaldte leksikalske profiler fra et korpus, beskrives i Kilgarriff og Rundell (2002).

3. Et sprogteknologisk værktøj for leksikografer: lemmatiser

I moderne datamatstøttet leksikografi bruges sprogteknologien på mange forskellige områder, eksempelvis i arbejdet med et tekstkorpus. En af de grundlæggende arbejdsprocesser er at gennemlæse relevante tekster og finde nye lemmaer til den ordbog der er under udarbejdelse eller opdatering. I denne proces er der god hjælp at hente fra forskellige sprogteknologiske værktøjer. På Center for Sprogteknologis hjemmeside, http://cst.dk/online/index.html, kan der afprøves en række af dem i kombination med hinanden (jf. "Seks værktøjer i tandem"). Kombinationen omfatter bl.a. en såkaldt POS-tagger (som beriger teksten med ordklasseopmærkninger), en navnegenkender og en lemmatiser. Eksempelvis kan der vha. lemmatiseren automatisk produceres en liste af lemmaer der forekommer i en given tekst. I det følgende fokuseres på lemmatiseren fordi den på flere måder er et godt eksempel på forholdet mellem datalingvistisk forskning, sprogteknologisk implementering og dataleksikografisk anvendelse. Nedenstående beskrivelse er mht. de tekniske detaljer baseret på dokumentationen af værktøjet.

Lemmatiseren er udviklet af Bart Jongejan og Dorte Haltrup Hansen i STO-projektet med det formål at dække behovet for et leksikografisk hjælpeværktøj til udtrækning af de ord fra fagrelaterede tekster der endnu ikke indgik i STOs ordforråd (som tidligere beskrevet, se afsnit 2.1). Målet var at lemmatiseren skulle være mere præcis og fleksibel end de traditionelle programmer til lemmatisering der normalt arbejder med trunkering. CST's lemmatiser er regelbaseret, og dens regler kan både håndtere regelmæssige og ikke-regelmæssige bøjninger. Lemmatiseringen omfatter tre opgaver:

- at føre hvert ord i en tekst tilbage til dets grundform (som er et kendt ord – et lemma i ordbogen)
- at vælge lemma hvis mere end et kendt ord er muligt lemma (homografer)
- at gætte lemmaer hvis grundformen ikke er kendt i ordbogen.

I lemmatiserens ordbogsmodul udnyttes STOs ordforråd og de bøjningsmorfologiske oplysninger der er anført for hvert lemma. Udvikling og træning af programmet blev gennemført i flere trin parallelt med udvidelsen af STO på følgende måde. Udgangspunktet i 2002 var ordforrådet på 50.000 lemmaer og deres bøjningsmønstre. Dette materiale blev udfoldet til en fuldformsordbog med i alt ca. 594.000 ordformer. Ud fra dette materiale udledtes bøjningsreglerne ("flex rules"), disse bruges af lemmatiseren til at genkende ordformer og føre dem tilbage til det pågældende lemma. For en detaljeret beskrivelse af produktion og applikation af regler mv. henvises til den fulde dokumentation (Jongejan & Haltrup 2005).

Da det nye tekstkorpus fra det første fagområde, edb/it, først var blevet lemmatiseret, kunne lemmatiseren således sammenholde den producerede lemmaliste med STOs ordliste og identificere hvilke lemmaer der var nye i forhold til STOs almensproglige ordforråd. I processen anvendtes forskellige outputformater, med fokus på lemmaet eller listen over ordformer og deres morfologiske etikette til hvert lemma som output. Listerne blev sorteret efter forskellige kriterier, disse er beskrevet i detaljer i Jongejan & Haltrup (op.cit.).

De relevante nye ord blev derefter integreret i STO-basen og forsynet med en lingvistisk beskrivelse, hvorefter lemmatiseren kunne trænes med det udvidede materiale. Denne proces blev så benyttet i flere gennemløb til at udvide STOs ordforråd med lemmaer fra yderligere fem fagområder. Lemmatiseren er således dels baseret på STO-materialet, dels blevet benyttet i det leksikografiske arbejde i STO til at lemmatisere nye tekster og udpege lemmakandidater til udbygning af ordforrådet.

På grund af STO-materialets størrelse og kvalitet kan lemmatiseren nu beregne lemmaet med 94-98 procents nøjagtighed. Det bedste resultat opnås hvis inputteksten er ordklasseopmærket. Tabel 1 viser en sammenligning af testresultater for lemmatiseren brugt med forskellige optioner udført på et korpus bestående af 250.000 løbende ord (det såkaldte PAROLE-korpus).

	Correct lemmas	Time
Input *with* POS-tags Lemmatisation *with* dictionary = **real lemmatiser**	97,8 %	App. 1 min.
Input *without* POS-tags Lemmatisation *with* dictionary = **discount lemmatiser**	94,5 %	App. 25 sec
Input *with* POS-tags Lemmatisation *without* dictionary = **good stemmer**	97,4 %	App. 48 sec
Input *without* POS-tags Lemmatisation *without* dictionary = **stemmer**	88,4 %	App. 30 sec

Tabel 1. *Sammenligning af testresultater for lemmatiseren. Kilde:* the CST Lemmatiser *(Jongejan & Haltrup 2005)*

Det bør bemærkes at ordbogskomponenten kan udskiftes med brugerens egen ordbog. Den skal blot indeholde de nødvendige oplysninger om lemma og ordformer således at lemmatiseren kan generere bøjningsreglerne ud fra denne ordbog. Desuden kan lemmatiseren også arbejde uden et ordbogsmodul, blot med de regler der er genereret fra ordbogen. Lemmatiseren kan for øvrigt også trænes til at håndtere andre sprog med suffiksbaseret bøjningsmorfologi, fx engelsk og svensk.

Det er indlysende at en sådan fleksibel lemmatiser, der også har flere faciliteter og forskellige input- og outputformater, kan bruges på mange måder i leksikografisk arbejde; nedenfor nævnes blot nogle få konkrete eksempler foruden de ovennævnte generelle funktioner.

Lemmatiseren kan beregne frekvensen af både de enkelte ordformer og samle alle forekomsttal for et lemmas ordformer. Med udgangspunkt i denne facilitet kan man lave automatiske undersøgelser og få svar på spørgsmål som fx

- Med hvilken hyppighed forekommer et givent lemma i det valgte korpus? – vigtigt for at kunne afgøre om lemmaet skal medtages i en ordbog der er under udarbejdelse eller udvidelse.
- Hvilken bøjningsform forekommer hyppigst i teksten/korpusset, fx *tallerkenen* (uden synkope) eller *tallerknen* (med synkope)? – og hvad skal stå først i artiklen? Bruges den græsk/latinske flertalsform så hyppigt, fx *korpora* i stedet for *korpusser*, at den bør det medtages i ordbogen? (Formen er jo ikke RO-godkendt, men er alligevel hyppig, især i lingvistiske fagtekster).
- I hvilke bøjede former forekommer lemmaet *test* – er den danske eller den engelskinspirerede (ikke RO-godkendte) flertalsform *tests* hyppigst?

- Hvad er lemmaet til bøjningsformerne af et nyt fremmedord der forekommer i teksten, fx *beepere*, eller *wannabees/wannabeer*.
- Med hvilket køn bruges fremmedordet hyppigst i korpusset? – fx en/et *website*.

Svaret på spørgsmålene, dvs. resultatet af denne type undersøgelser, kan med fordel inddrages i udformningen af ordbogsartikler fordi oplysningerne baseres på empiri i stedet for på introspektion. Det er indlysende, og et velkendt faktum, at brugen af en lemmatiser og andre sprogteknologiske hjælpeværktøjer øger både effektiviteten og pålideligheden af det leksikografiske arbejde.

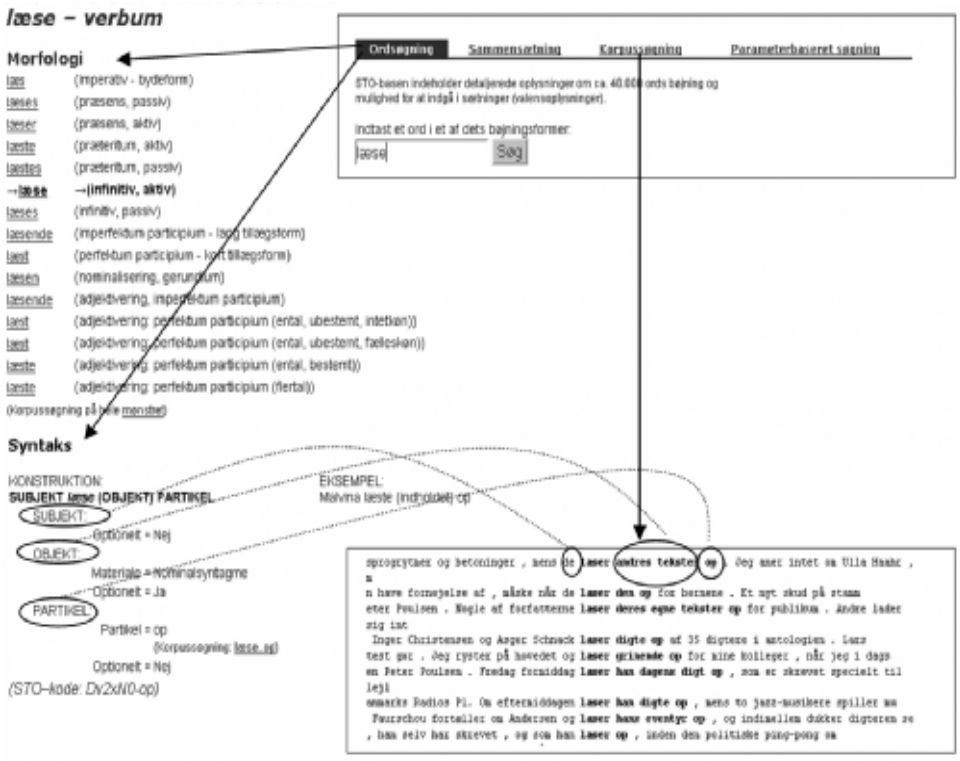
Den detaljerede dokumentation (Jongejan & Haltrup, op.cit.) findes på adressen http://cst.dk/online/index.html, hvor også lemmatiseren kan afprøves.

4. Online opslag i STO

Den Sprogteknologiske Ordbase er frit tilgængelig for online opslag over internettet. Leksikografer, lingvister og alle interesserede sprogbrugere kan derved få et indblik i den største del af STO-databasens indhold, idet semantikdelen dog ikke er tilgængelig. Desuden er der en række oplysningstyper som ikke vises på skærmen fordi de kun er relevante for datamatiske applikationer. Brugergrænsefladen er tilgængelig fra http://cst.dk/sto/webinterface/index.html. På hjemmesiden findes der også en brugervejledning og eksempler på interessante søgeord.

Grænsefladen har fire forskellige søgemuligheder (som det fremgår af figur 1, der viser et udsnit af sammenklippede skærmbilleder); de skitseres kort her:

- *Ordsøgning* med et lemma eller en ordform. Søgeresultatet er opslagsordets bøjningsformer (i alle dets bøjningsmønstre) eller den søgte ordforms lemmaform og dens fulde bøjning. Desuden vises lemmaets – eller lemmaernes – syntaktiske konstruktionsmuligheder med eksempler.
- *Substantiviske sammensætninger.* Søgeresultatet er en liste med alle substantiver som søgestrengen indgår i, dvs. både sådanne ord som er registrerede og mulige sammensætninger (sidstnævnte er markeret med kursiv), fx 'mand' i 'mandeår' og 'havemand' vs. '*konfirmand*'. Desuden kan man se sammensætningens struktur.
- *Korpussøgning* kan ske med et lemma, en enkelt ordform eller et ords udvalgte konstruktion. Søgeresultatet er forekomsterne i kwic-format fra et korpus bestående af artikler fra Berlingske Tidende, årgang 90-92.
- *Parameterbaseret søgning* med kombination af forskellige lingvistiske kriterier (fx ordklasse + aritet + styret præposition). Søgeresultatet er en liste af tilfældigt udvalgte ord (op til 30), hvis egenskaber svarer til den valgte kombination.



Figur 1. *Mulighederne for søgning i STO-basen*

Den sidstnævnte søgemulighed er nok den mest interessante for leksikografer, og den er beskrevet i Braasch & Olsen 2005 på følgende vis:

> […] man kan få vist grupper af ord der har samme lingvistiske egenskaber ud fra den valgte parameterkombination, eksempelvis en gruppe af divalente verber med partikel og/eller præpositionen *for*. Tabel 2 viser et udsnit af resultatet for denne søgning. Den forkortede beskrivelse af konstruktionen kan være lidt svær at gennemskue, derfor er det muligt at klikke på lemmaet og få vist alle ordets konstruktioner i detaljer og med eksempler. I tilfældet *ængste* er nedenstående konstruktion en ud af de 5 registrerede for dette verbum. Ved samme opslag vises også ordets morfologi for at skabe sammenhæng mellem det morfologiske og det syntaktiske beskrivelseslag.

STO er leksikografisk og lingvistisk velfunderet, og grænsefladen på internettet gør den til en online sproglig oplysningskilde, også for leksikografer. For at man kan sammenligne søgeresultaterne fra STO med andre danske sprogressourcer, er der etableret links på nettet til

- Retskrivningsordbogen (3. udgave 1996-2002, Dansk Sprognævn)
- Korpus 2000 (Det Danske Sprog- og Litteraturselskab)
- Google (søgning på danske internetsider).

Nr.	Lemma	Ordklasse	Forkortet beskrivelse af konstruktion (Klik på lemmaet for uddybning)
1	*stå, 1*	*verbum*	*divalent: NP, PP(NP/infinitive subject equi control) prep=for*
2	*ængste*	*verbum*	*divalent reflexive: NP, PP(NP/that-c/infc w. subject control) prep=for*
3	*udgive*	*verbum*	*divalent reflexive: NP, obligatory PP(NP/infc) prep = for*
4	*grue*	*verbum*	*divalent: NP, PP(NP/that-clause) prep=for*

Tabel 2. *Søgeresultat for en parameterbaseret søgning (udsnit af skærmbillede)*

5. Opsummering

Formålet med denne artikel har været at præsentere den nu færdige Sprogteknologiske Ordbase for dansk, STO, med fokus på de oplysningstyper der er relevante for ordbasens anvendelse i et værktøj for leksikografer. Den datamatiske leksikografi har to aspekter. Det ene er udarbejdelsen af leksikalske datasamlinger for sprogteknologiske anvendelser – som sådan er STO den mest omfattende for dansk. Det andet aspekt er anvendelsen af sprogteknologiske (datamatisk baserede) værktøjer i leksikografi. I dette indlæg har jeg forsøgt at vise hvordan disse to aspekter i dag er tæt forbundne med hinanden, og hvilket gavn leksikografien kan have af den sprogteknologiske udvikling der er baseret på datalingvistisk forskning.

Slutnote

Projektet har modtaget bevilling fra Ministeriet for Videnskab, Teknologi og Udvikling for perioden 1. marts 2001-29. februar 2004. Ordbasen er produceret i et samarbejde mellem Center for Sprogteknologi, KUA; Institut for Datalingvistik, CBS; Institut for Almen og Anvendt Lingvistik, KUA; Institut for Fagsprog, Kommunikation og Informationsvidenskab, SDU. Projektledelse og koordinering blev varetaget af CST. Brugergrænsefladens udvikling blev finansieret af Danmarks Elektroniske Forskningsbibliotek (DEF) i 2002.

Litteratur

Internetadresser der er nævnt i artiklen (pr. 15. september 2005):
http://cst.dk/sto/referencer/index.html
http://cst.dk/online/index.html
www.cst.dk/sto

Ordbøger:

DDO = *Den Danske Ordbog.* (Hovedred. Ebba Hjorth & Kjeld Kristensen). København: Det Danske Sprog- og Litteraturselskab og Gyldendal, 2003-2005.
RO = *Retskrivningsordbogen* (2001), elektronisk version.
STO = Braasch, Anna m.fl. (red.): *Sprogteknologisk Ordbase.* København: Center for Sprogteknologi, 2001-2004. www.cst.dk/sto

Anden litteratur:

Braasch, Anna, Costanza Navarretta, Sanni Nimb, Sussi Olsen, Bolette S. Pedersen & Claus Povlsen 2004-2005: *STOs Lingvistiske Specifikationer*. CST: København (Upubliceret dokumentation, kan udleveres efter aftale).
Braasch, Anna & Sussi Olsen 2005: Den SprogTeknologiske Ordbase (STO) for dansk – Leksikografiske aspekter. I: *LEDA-Nyt nr. 39 - Marts 2005,* 13-20. Leksikografer i Danmark, København: Center for Sprogteknologi.
Braasch, Anna & Bolette S. Pedersen 2002: Recent Work in the Danish Computational Lexicon Project "STO". I: Braasch, A. & C. Povlsen (red.), *The Tenth EURALEX International Congress, Proceedings, Vol. I.* Copenhagen: CST, 301-314.
Hansen, Aage 1967: *Moderne dansk.* København: Grafisk Forlag.
Jongejan, Bart & Dorte Haltrup 2005: *the CST Lemmatiser*, version 2.7 (23. august 2005). Center for Sprogteknologi. (Publiceret på http://cst.dk/online/index.html).
Kilgarriff, Adam & Michael Rundell 2002: Lexical Profiling Software and its Lexicographic Applications – a Case Study. I: Braasch, A. & Povlsen, C. (red.) *The Tenth EURALEX International Congress, Proceedings, Vol. II*. Copenhagen: CST, 807-818.
Olsen, Sussi 2005: STO – En sprogteknologisk database. I: Peter Widell & Mette Kunøe (red.): *10. Møde om Udforskningen af Dansk Sprog*. Århus, 289-297.
Pedersen, Bolette S. 2005: Datamatisk leksikografi i Norden – status og visioner. I: Ruth V. Fjeld & Dagfinn Worren (red.): *Nordiske Studiar i leksikografi. Rapport frå Konferanse om leksikografi i Norden, Volda 20.-24. mai 2003*. Oslo: NFL, 302-314.
Pollard, Carl & Ivan Sag 1994: *Head-Driven Phrase-Structure Grammar.* Chicago & London: The University of Chicago Press.
Ørsnes, Bjarne 2004: Automatisk opbygning af et LFG-baseret datamatisk leksikon for dansk. I: Henrik Holmboe (red.): *Nordisk Sprogteknologi. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. København: Museum Tusculanum, 211-237.