OLAM – et semiautomatisk morfologisk og lydstrukturelt kodningssystem for dansk

Thomas O. Madsen, Hans Basbøll og Claus Lambertsen[1]
Syddansk Universitet, Institut for sprog og kommunikation
Campusvej 55, 5230 Odense M.
E-post: tom@language.sdu.dk, hba@language.sdu.dk, claus@lambertsen.de

Abstract

This paper describes the OLAM system. It contains a database containing orthographic forms combined with morphological forms and information pertaining to segmentation and pronunciation. OLAM is first intended for semi-automatic coding of written text in Danish, second for complex linguistic searches and hypothesis testing.

Indledning

OLAM-systemet, der vil blive nærmere beskrevet i det følgende, udvikles i et samarbejde mellem Hans Basbøll, Thomas O. Madsen, begge fra Center for Sprogtilegnelse[2] (herefter OPS), Syddansk Universitet, og Claus Lambertsen, Berlin. Peter Molbæk Hansen, KUA, der er forfatter og copyrightindehaver til Gyldendals røde ordbog "Dansk Udtale" (herefter DU), 1990, deltager ikke i arbejdet med OLAM-systemet, men er orienteret om og har tilladt anvendelsen af DU i en elektronisk version med henblik på udviklingen af et morfologisk og lydstrukturelt søgnings- og kodningsprogram.

Hensigten med OLAM-systemet er på den ene side at udvikle en database, baseret på DU, indeholdende ortografiske former koblet til morfologiske former, segmenteringsoplysninger og udtaleformer. Denne database vil udgøre fundamentet for et kodningsprogram (OLAM-code), der skal anvendes til semiautomatisk kodning af danske standard-ortografiske tekster, bl.a. et stort longi-

tudinalt talesprogskorpus (Odense Tvillinge-Korpus) indsamlet af Odense Projektet i Sprogtilegnelse (se note 2) og delvis transskriberet i løbet af projektperioden. På den anden side vil databasen blive udbygget med yderligere informationer og søgefaciliteter og på den måde udgøre en avanceret søgemaskine (OLAM-search) med mulighed for komplekse lingvistiske søgninger og testning af hypoteser.

OLAM er et projekt under udvikling. Mange af de beskrevne processer er realiseret, og kodningsprogrammet forventes at være operationelt i løbet af sommeren 2002. Beskrivelsen i det følgende vil først og fremmest omhandle udviklingen af kodningsprogrammet OLAM-code, dvs. arbejdet med at omstrukturere databasen, at tilføje flere bøjningsformer, generere morfologiske kodningsformer m.v., og endelig opgaven med at indarbejde databasen i et operationelt kodningsprogram.

1. DU og OLAM

Udgangspunktet for arbejdet med OLAM-systemet, der som nævnt udvikles bl.a. med henblik på et semi-automatisk morfologisk og lydstrukturelt kodningssystem for dansk standardortografi, er en elektronisk version af DU, som Claus Lambertsen for nogle år siden fik stillet til rådighed af Peter Molbæk Hansen. Claus Lambertsen har forarbejdet den elektroniske version af DU, så den senere kunne importeres i en FileMaker-database, som han har udviklet med henblik på bl.a. digraf- og difon-søgeprocedurer.

1.1. Strukturen i DU

Det vil altid være temmelig vilkårligt, hvor mange ord der medtages i en ordbog, uanset om det er en en- eller tosproget ordbog. At dansk er et sprog med vide muligheder for dannelse af sammensatte ord, gør ikke afgrænsningsproblematikken mindre relevant. Peter Molbæk Hansen pointerer i det indledende afsnit om 'Ordbogens principper og dens brug', at han i det konkrete udvalg af ord har

skelet stærkt til Retskrivningsordbogen (1986) "således forstået at der i udtaleordbogen ikke forekommer ret mange opslagsord der ikke også forekommer i Retskrivningsordbogen". Det skal dog pointeres, at det samlede antal opslagord i DU er noget mindre end i Retskrivningsordbogen, hvilket bl.a. skyldes, at DU ikke indeholder lige så mange sammensatte ord, da udtalen af langt de fleste sammensatte ord kan udledes af ordenes enkelte led, fx *kaffekande*[3]. Visse sammensatte substantiver er dog medtaget som opslagsord, fordi et eller flere af deres enkelte led ikke forekommer alene, og/eller deres udtale ikke kan konstrueres efter sammensætningsregler (fx *trykseksten*, *rødgrød*).

1.2. Opslagsord

DU indeholder 41102 ordbogsartikler, der er ordnet alfabetisk efter deres opslagsord. I forbindelse med import af den forarbejdede DU i en FileMaker database valgte vi at lade hvert kort indeholde ét opslagsord med tilhørende bøjningsformer. I en lang række tilfælde indeholder opslagsordet i DU en lodret streg ("|"), der deler ordet i en venstredel bestående af en "grafemisk konstant" (i GC-feltet[4]) og en højredel bestående af en "grafemisk endelse" (da endelserne endnu ikke er morfologisk identificerede, har vi valgt at betegne feltet "GV" for "grafemisk variabel").

Ved termen "grafemisk konstant" henviser vi til den del af et ord, der manifesteres ortografisk ens i alle ordets bøjningsformer, fx *cyk* i substantivet *cyk*|*-el*, *cyk*|*-(e)len*, *cyk*|*-ler* eller *vris* i adjektivet *vris*|*-sen*, *vris*|*-sent*, *vris*|*-ne*[5].

Den "grafemiske endelse" er den del af et ord, der ikke udgør en grafemisk konstant, men indeholder bøjningsspecifikke ortografiske oplysninger, fx endelsen *-pe* i infinitiv af verbet *hop*|*-pe*. Som det fremgår af eksemplet, er der ikke tale om morfologiske endelser, men grafemiske endelser indeholdende bl.a. bindekonsonanter.

Ved de bøjede ordklasser er opslagsordet på sædvanlig leksikografisk vis følgende: singularis ubestemt form af substantiver, singularis fælleskøn ubestemt

form af adjektiver, infinitiv ved verber. I forbindelse med hvert opslagsord er der desuden angivet en af følgende ordklasser: adj., adv., art., interj., konj., num., onom., pron., prop., præp., sb., vb.[6]

Endelig er DU struktureret således, at "homografe opslagsord", dvs. potentielle opslagsord, der manifesteres ortografisk ens, repræsenteres af samme opslagsord, hvis de har samme distributionsmønster i de for ordbogen relevante lydlige, ordklassemæssige og bøjningsmæssige henseender. Hvis de homografe opslagsord derimod har forskellige distributionsmønstre i en eller flere af disse henseender, repræsenteres de af forskellige artikler.

1.3. Bøjningsformer

Hvis vi ser bort fra de fuldt udskrevne særformer, der vil blive behandlet nedenfor, anføres kun den grafemiske endelse i andre bøjningsformer end opslagsformen, og en bindestreg ("-") repræsenterer her den grafemiske konstant, der er fælles for opslagsordets bøjningsformer.

Til opslagsordet følger i de fleste artikler et antal bøjningsformer. Følgende typer af bøjede former er almindeligvis angivet:

• Substantiver: bestemt form singularis og ubestemt form pluralis. Bestemt form pluralis er kun angivet i tilfælde af særformer (fx *børnene*, *gæssene*, se nedenfor) og i substantiver, hvis ubestemt form singularis er lig ubestemt form pluralis (fx *lam*|*-mene*, *kort*|*-ene*). Sidstnævnte kategori er desuden markeret i feltet PL=SG.

• Adjektiver: neutrum. Som følge af DUs (ortografibaserede) principper er der ikke anført en distinktion mellem adjektivers bestemte form og pluralis, hvilket betyder, at e-formen kun forekommer én gang. Komparativ og superlativ anføres, hvis adjektivet kan gradbøjes.

• Verber: præsens, præteritum, perfektum participium (og præsens participium ved verber, der ender på vokal) og imperativ. Imperativ, der altid er lig verbets

grafemiske konstant, markeres ved "-!" i feltet for den grafemiske endelse, fx GC: "*hop*|" og GV: "-!" for verbet *hop*|*-pe*.

• Pronominer: alle bøjede former.

Hvis der i forbindelse med et opslagsord eksisterer morfologiske dobbeltformer, fx flere præteritumsformer (*grinede*, *grinte*), er alle former medtaget, hvis de er forskellige fra alle øvrige former af ordet. Disse morfologiske dobbeltformer (MDF) har vi senere markeret i MDF-felter.

For de fleste ord er desuden medtaget den form, ordet optræder i som førsteled i sammensætninger. Hvis denne form ikke er medtaget, betyder det, at ordet – hvis det overhovedet kan optræde som førsteled – er sammenfaldende med opslagsformen i både ortografisk og lydlig henseende. Hvis førsteledsformen ortografisk, men ikke lydligt, er lig opslagsformen, er det markeret i feltet I SMS. Hvis førsteledsformen derimod ortografisk er lig hele opslagsformen (GC og GV), er den både markeret i feltet I SMS og GV-feltet er udfyldt af opslagsordets GV-indhold efterfulgt af en eller flere bindekonsonanter og et "+":

Opslagsord	GC	GV	eksempel
skab	*skab*|	*-s+*	*skabsdør*
føds| *-el*	*føds*|	*-els+*	*fødselshjælper*
dom|	*dom*|	*+*	*domhus*
	dom|	*-me+*	*dommedag*
	dom|	*-s+*	*domsmand*

Som det fremgår af sidste eksempel, kan et ord sagtens have flere former som førsteled i sammensætninger, fx *land*|*+, -e+, -s+; dag*|*+, -e+, -s+; bøn*|*+, -ne+*.

Endelig behandles visse bøjningsformer som særformer ("SF"). Hvis en bøjningsform er forskellig fra opslagsformen og ikke kan dannes ved en sammenkædning af den grafemiske konstant og en grafemisk endelse, udgør formen en

særform og hele formen er udskrevet, fx *var*, præteritumsformen af *vær|e*, *ænder*, pluralis ubestemt form af *and|*. Særformer indeholder ikke lodrette streger til adskillelse mellem en grafemisk konstant og grafemiske endelser, og indeholder altså heller ikke nogen grafemisk endelse. Vi har valgt at importere særformer i GC-felterne, men har samtidig markeret i feltet SF, at der er tale om særformer.

1.4. Udtaleformer

Til hver af de optagne ordformer er der angivet en eller flere udtaleformer. Lydskriften i DU udgør en tilpasset variant af den internationale standard IPA.

2. Identifikation og udspaltning af morfologiske endelser

Som det fremgår af beskrivelsen ovenfor, var den elektroniske version af DU, som vi fik adgang til, struktureret med henblik på at blive anvendt som udtaleordbog, dvs. ud fra et ortografisk princip og kun med ganske få morfologiske oplysninger. Vores første opgave var derfor at omstrukturere databasen fra primært at være organiseret efter ortografiske principper til at udgøre en morfologisk struktureret database.

Vores første problem i dette arbejde var, at der ikke eksisterede en en-til-en relation mellem felter og feltindhold, hvilket betød, at vi ikke kunne forudsige, hvilke bøjningsformer vi kunne finde i et specifikt (GV-) felt eller hvilke (GV-) felter en specifik bøjningskategori begrænsede sig til. Vi kunne med andre ord ikke blot flytte de morfologiske endelser til bøjningsspecifikke felter, men var nødsaget til at søge morfologiske bøjningsendelser i samtlige GV-felter.

For at løse dette problem blev der oprettet en række felter – ME-felter – der skulle indeholde specifikke identificerede morfologiske bøjningsendelser. For også at kunne segmentere komplekse morfologiske endelser, som fx pluralis bestemt form i *hus|-e-ne*, oprettede vi 2 ME-felter til hver bøjningsform, dvs. et ME01-felt og et ME02-felt pr. linje. Ganske vist kan der ud fra lingvistiske kri-

terier, fx Basbølls ordstrukturmodel baseret på endelsernes produktivitet (1998, 2001), argumenteres for, at der skal anvendes mere end blot to ME-felter pr. bøjningsform, men vi valgte at følge et minimalprincip[7]. Dernæst begyndte vi at søge og fraspalte de specifikke morfologiske bøjningsendelser, så enhver morfologisk endelse kunne placeres i et ME-felt, dog således at intet ME-felt kunne være udfyldt af mere end én morfologisk endelse[8].

I forbindelse med placeringen af morfologiske bøjningsendelser i hhv. ME01- og ME02-felter opererer vi med fire prioriterede generelle principper:

(1) hvis der i en grafemisk endelse kan identificeres to morfologiske bøjningsendelser, placeres den første i ME01 og den anden i ME02; (2) samme grammatiske kategori anbringes gennemført i enten ME01 eller ME02, uafhængigt af bøjningslinje; (3) Ø-endelser indsættes kun i ME-felter hvor andre udtryk for den pågældende morfologiske kategori ville stå; (4) hvis en grafemisk endelse kun indeholder én morfologisk bøjningsendelse (herunder "Ø"), placeres denne i ME01.

Vores overordnede strategi for segmentering af morfologiske endelser var at fraspalte maksimalt bagfra i GV-felterne, dvs. i de grafemiske endelser. Det forhold at opslagsordene er opdelt i en grafemisk konstant og en grafemisk endelse vha. en lodret streg betød, at vi var sikret mod at komme til at segmentere i den grafemiske stamme, fx at fraspalte *en* som ental bestemte form i *sten|, fænomen|, helgen|* eller *er* som præsens i *accepter|, administrer|, adopter|*.

Strategien med at fraspalte maksimalt bagfra i GV-felterne udtrykker en lingvistisk hypotese om, at segmentering af danske bøjningsendelser overvejende kan udføres vha. enkelte formmæssigt identificerbare morfologiske endelser (jf. Basbølls ordmodel, Basbøll 1998, 2001).

Nedenfor har vi opstillet nogle eksempler på søgnings- og fraspaltningsprocedurer samt placering af morfologiske bøjningsendelser i ME01- og ME02-felter.

Infinitiv er altid lig hele opslagsordet (GC og GV) i ordklassen verber og skal derfor kun søges i første linje (i felterne GC_01 og GV_01).

Hvis infinitivsendelsen ("e") overhovedet er til stede, er den altid placeret (sidst) i feltet GV_01, eventuelt efter en bindekonsonant (fx i *hop|p-e*), og kan uden videre flyttes til feltet A_ME01[9]. I tilfælde uden udtrykt infinitivsendelse indsættes et "Ø", fx i *gå|*.

I søgningen af præsens-endelsen "r" er vi pga. defektive former og morfologiske dobbeltformer nødsaget til at udvide søgefeltet til at omfatte alle GV-felter[10]. Alle identificerede præsens-endelser, dvs. "r", fundet ved søgning først på "*er" og derefter på "*r" kan herefter flyttes til feltet B_ME02. I tilfælde med "er" flyttes "e" til B_ME01, ellers indsættes "Ø" i B_ME01 (fx *hop|p-e-r*, *gå|-Ø-r*). Eventuelle bindekonsonanter bliver tilbage i GV-feltet (fx *p* i *hop|p-e-r*). Placeringen af "e" i B_ME01 skal opfattes som udtryk for, at basisformen for præsensformen *hopper* er en infinitivsform.

Principperne for segmentering af morfologiske bøjningsendelser beskrevet ovenfor vedrører udelukkende segmentering i GV-felterne. Når det drejer sig om særformer (markeret i SF-feltet), må vi gå lidt mere forsigtigt til værks, da særformerne ikke indeholder en lodret streg, der kan blokere for segmentering i den grafemiske stamme. Som overordnet strategi har vi valgt at udskrive alle særformer – først sorteret efter ordklasse og derpå efter deres "endelse" (dvs. om særformen ender på en bogstavsekvens, der også findes som fraspaltet bøjningsendelse). Med afsæt i denne liste afprøves, hvilke af søgningsprocedurerne fra GV-felterne, vi uden fejlanalyser kan overføre til særformerne.

3. Dannelse af nye bøjningsformer

Mht. bøjningsformer er udgangspunktet i DU, at bøjede former er udeladt, hvis deres lydskrift kan dannes ved blot at sammenkæde lydskriften af opslagsordet med lydskriften af en lydligt konstant endelse, eller hvis de i alle relevante henseender opfører sig parallelt med en medtagen bøjningsform. Med andre ord er bl.a. flere af de højfrekvente produktive endelser ikke medtaget, fx bestemt form

pluralis af substantiver (*-ne* i *bil|-er-ne*), præsens participium i de tilfælde, hvor verbet ender på en konsonant (*-ende* i *løb|-ende*), gerundium (*-en* i *handl|-en* og *-en* i *gå|-en*) og passiv (*-s* i *hop|p-ede-s*). For at OLAM-systemet kan fungere optimalt som et morfologisk og lydstrukturelt søgnings- og kodningsprogram, skal vi naturligvis indføje disse bøjnings- og udtaleformer i databasen.

I det følgende vil vi med udgangspunkt i verbers præsens participium og bestemt form pluralis af substantiver illustrere, hvilke procedurer vi tager i anvendelse for at danne de omtalte bøjnings- og udtaleformer.

Præsens participium er allerede udfyldt (i feltet F_ME01) for det mindretal af verber, der ender på en vokal. I alle andre tilfælde, dvs. hvor verbet ender på en konsonant, indsættes "ende" i F_ME01. Mht. udtalen kopieres infinitivens udtale og "nə" indsættes efter infinitivens "ə".

Bestemt form pluralis af substantiver er ikke medtaget i normaltilfældet, dvs. hvis der kun er én ubestemt form i pluralis og bestemthedsformen er denne form efterfulgt af “ne”. I de tilfælde, hvor feltet til bestemt form i pluralis (D_ME02) er tomt, kan vi indsætte den ortografiske form for ubestemt pluralis og tilføje “ne”. Mht. udtalen indsættes udtaleformen af ubestemt form pluralis efterfulgt af "nə".

4. MOL-, SOL- og POL-former

Som det er beskrevet i indledningen, er formålet med OLAM-projektet bl.a. at udvikle et kodningsprogram til kodning af danske standard-ortografiske tekster[11]. Til det formål skal OLAM-databasen indeholde mindst tre typer informationer, nemlig morfologiske former, segmenteringsoplysninger og udtaleformer. De tre informationsniveauer betegnes inden for OLAM-systemet MOL-former (Morphological OLAM forms), SOL-former (Segmentized OLAM forms) og POL-former (Phonetic/Phonological OLAM forms).

Eftersom udtaleformerne, hvis 1. form udgør POL-formen, og SOL-formerne, der bygges op vha. grafemisk stamme, eventuelle bindekonsonanter og morfologisk(e) endelse(r), bortset fra grænsesymbolerne, allerede er til stede i databasen, vil vi ikke komme nærmere ind på dem her. Procedurerne til generering af MOL-formerne vil derimod blive behandlet nedenfor.

Som det fremgår af eksemplet sidst i artiklen, N|tank-DEF:en, er en MOL-form typisk bygget op af følgende elementer (fra venstre mod højre): VERSALER, der angiver ordklasse (fx ADJ, N eller NUM[12]); symbolet "|", der adskiller ordklasseangivelse og opslagsordet; almindelige bogstaver, der angiver opslagsordet i standardortografi; et morfologisk grænsesymbol (uddybes nedenfor); VERSALER, der angiver den grammatiske kategori, efterfulgt af et kolon (fx DEF:, POSS:, PAST:); og endelig almindelige bogstaver, der angiver morfologisk(e) bøjningsendelse(r).

Grænsesymbolerne, der overvejende er morfologiske, anvendes på følgende måder:

Symbolet "-" er det umarkerede symbol før (adderet, dvs. ikke-fusioneret) morfologisk endelse i ME01-feltet (fx N|bil-PL:er), hvor "=" er det umarkerede symbol før morfologisk endelse i ME02-feltet (fx N|bil-DEF:en); begge symboler forekommer i N|bil-PL:er=DEF:ne.

Symbolet "&" findes før morfologisk endelse, der er delvist fusioneret, dvs. at der faktisk er udtrykt en segmenterbar endelse, men at den ikke er føjet til ordets normale stamme (fx *bøger*, N|bog&PL:er; *ænder*, N|and&PL:er)[13]; Symbolet "&" kan kun forekomme én gang i en MOL-form (aldrig sammen med "-"): *ænderne* N|and&PL:er=DEF:ne.

Symbolet "+" er placeret mellem leddene i sammensatte ord (fx *landmand*, N|land+N|mand);

"^" findes før "medfødte", totalt fusionerede, semantiske kategorier. Symbolet "^" er i realiteten ikke et morfologisk grænsesymbol, men indikerer en særlig relation mellem to indholdskategorier, nemlig "total fusion". Symbolet "^" an-

vendes i eksempler som *penge,* N|penge^PL, og *skib*, N|skib^NEUT, hvor det angiver, at *penge* er "medfødt" pluralis og *skib* er "medfødt" neutrum (markeret indholdskategori). Teknisk set er symbolet "^" af lavere rang end (de andre) morfologiske grænsesymboler, hvorfor operatoren implicerer, at de to omgivende elementer skal behandles som en enhed i relation til andre operatorer (hvad enten de omgivende elementer består af en stamme og en grammatisk kategori, fx *skib* og NEUT, eller to grammatiske kategorier, fx PL og DEF)[14]. Som illustration skal parenteserne i *musene* N|(mus-(PL^DEF:)ene) og *husene* N|(((hus^NEUT)-PL:e)-DEF:ne) opløses udefra og indad.

5. Kodning af tekster

Som det fremgår af ovenstående beskrivelse vil alle bøjningsformer, når MOL- og SOL-formerne er færdiggenereret, bl.a. bestå af en ortografisk form, en morfologisk form (MOL), en segmenteret form (SOL) og en udtaleform (POL). I forbindelse med kodning af danske standard-ortografiske tekster vil den ortografiske form kunne fungere som indgang til den specifikke bøjningsforms grammatiske og udtalemæssige former.

Det største problem forbundet med en "ord-til-ord"-identifikation af tekster består i det store antal homografer, der eksisterer på dansk. Homografiproblemet udelukker desværre, at kodningsprogrammet kan foretage fuldautomatisk kodning af tekster, med mindre man indbygger en eller anden form for avanceret syntaktisk parser. Vi har valgt, i hvert fald i denne omgang, at løse homografiproblemet på den måde, at det er en operatør, der varetager opgaven med disambiguering i tilfælde af homografi.

I det færdige semiautomatiske kodningssystem vil resultatet af en søgning føre til en af følgende tre situationer:

(1) kun én ortografisk form i databasen matcher inputformen. I dette tilfælde indsættes automatisk morfologisk, segmenterings- og udtaleinformation i de tre

følgende linjer under den identificerede inputform (hhv. i MOL-, SOL- og POL-linjen);
(2) ingen form i databasen matcher inputformen. I det tilfælde indsættes den ortografiske inputform (markeret med et særligt symbol) automatisk i tre følgende linjer under den søgte inputform;
(3) to eller flere ortografisk former i databasen matcher inputformen. I tilfælde af homografi skifter kodningsprogrammet til et layout, der anvendes i forbindelse med disambiguering. I disambigueringslayoutet har operatøren mulighed for at vælge den rette form ud fra oplysninger om MOL-, SOL- og POL-former. I eksemplet nedenfor vælger operatøren den rette af tre mulige MOL-, SOL- og POL-former.

	tanken	*tanken*	*tanken*
MOL-form	N|tanke-DEF:en	N|tank-DEF:en	N|tank-DEF:en
SOL-form	tanke|-n	tank|-en	tank-en
POL-form	tANgən	tAN'gən	ta:Ngən[15]

Som en ekstra sikkerhed har vi valgt, at operatøren skal godkende en hel ytring og dens kodning, før de tre kodningslinjer (MOL-, SOL- og POL-linjerne) overføres til det oprindelige dokument.

Efterord (juni 2002)

Pr. juni 2002 er status for OLAM-projektet følgende: En prototype af kodningsdelen er under fortsat afprøvning og OLAM-code er ved at blive implementeret. I løbet af sommeren 2002 foretages forskellige testninger og tilpasninger af systemet. I efteråret 2002 iværksættes den semiautomatiske kodning af vore tvillingedata (under projektet Danske Børns Sprogtilegnelse som støttes af Carlsbergfondet fra 2002). Parallelt hermed udvikles søgemaskinen OLAM-search

(under samme projekt). Der vil blive arbejdet henimod at etablere internetadgang til dele af systemet.

Noter:

[1] Rækkefølgen af forfatternavne udtrykker en vægtfordeling vedrørende udarbejdelsen af denne artikel, ikke OLAM-projektet som sådan.
[2] OLAM har været tilknyttet Odense-Projektet i Sprogtilegnelse i et projekt (1998-2001) støttet af Statens Humanistiske Forskningsråd og Syddansk Universitet. OLAM indgår nu i projektet *Danske Børns sprogtilegnelse*, der støttes af Carlsbergfondet fra 2002.
[3] Ordets udtale som førsteled i sammensætninger er dog angivet sammen med udtalen af andre bøjningsformer.
[4] Feltbetegnelser vil i det følgende være angivet med VERSALER, hvorimod feltindhold markeres ved dobbelte anførselstegn.
[5] Den lodrette streg ("|") er placeret så langt til højre i opslagsordet som muligt, for at der ved hver bøjningsform skal skrives så lidt som muligt.
[6] Her, som i resten af denne beskrivelse, anvender vi Peter Molbæk Hansens grammatiske betegnelser.
[7] Der findes bøjningsformer med tre morfologiske endelser, fx substantivernes pluralis bestemt form genitiv (*hus| -e-ne-s*), men da vi behandler genitiv som klitisk, skal vi ikke have et særligt ME-felt til genitiv.
[8] Det skal understreges, at antallet af morfologiske endelser her udelukkende vedrører udtrykskategorier, ikke antallet af (semantiske) indholdskategorier (fx udgør endelsen i *mus|-ene* kun én morfologisk endelse på udtrykssiden, men udtrykker på indholdssiden både pluralis og bestemthed).
[9] Af praktiske hensyn har vi valgt at referere til "linjer" i den oprindelige del af databasen, dvs. den grafemisk strukturerede del, med 01, 02, 03, osv., hvorimod "linjer" i den morfologisk strukturerede del af basen betegnes A, B, C, osv.
[10] Både grafemiske og morfologiske dobbeltformer håndteres ved procedurer, der af hensyn til artiklens omfang ikke vil blive beskrevet her.
[11] Med standardortografi henvises der til Retskrivningsordbogen 1986, da denne var udgangspunktet for udviklingen af DU mht. standardformer (jf. afsnit 1.1 Strukturen i DU).
[12] Grammatiske betegnelser og ordklassebetegnelser i det færdige OLAM-system afviger på en række punkter fra betegnelserne i DU, derimod er de en delmængde af de (engelske) betegnelser, der anvendes inden for CHILDES (http://childes.psy.cmu.edu/).
[13] Distinktionen mellem "-" (ingen fusion) og "&" (delvis fusion) består i, hvorvidt der er en ortografisk forskel på opslagsordets stamme og den pågældende morfologiske forms stamme (eller mere præcist: et "&" markerer, at opslagsformens grafemiske konstant ikke er lig begyndelsen af den pågældende bøjningsforms grafemiske konstant et vice versa).
[14] Med andre ord kan de to indholdskategorier, der er forbundet af operatoren "^" (fx i PL^DEF eller hus^NEU), ikke opløses i to selvstændige (ikke-tomme) elementer på udtrykssiden.
[15] Beklageligvis var det af tekniske årsager ikke muligt i nærværende artikel at gengive lydskrift med IPA-symboler. "A" svarer til fuldvokalen i *banke*, "N" svarer til sidste lyd i *streng*, """ symboliserer stød, "a:" svarer til langvokalen i *gade*.

6. Litteratur

Basbøll, Hans. 1998. Nyt om stødet i moderne rigsdansk: om samspillet mellem lydstruktur og ordgrammatik. *Danske Studier 1998*:33-86.

Basbøll, Hans. 2001. Fuldproduktive bøjningsendelser og stød: nogle konsekvenser af en ny model for ordstruktur. In Pia Jarvad et al. (eds.) [Festskrift]. København:Hans Reitzels Forlag.

Dansk Sprognævn. 1986. *Retskrivningsordbogen*. København:Gyldendal.

Hansen, Peter Molbæk. 1990. *Dansk udtale*. København:Gyldendal.