SIMULERING AF RELATIONEL DATABASE

Bodil Nistrup Madsen
Institut for Datalingvistik

Indlæg på Symposium for datamatstøttet leksikografi og terminologi, 5.-6. november 1987, Handelshøjskolen i København

I det følgende rapporteres om et forsøg, som er gennemført med henblik på at afprøve, om et informationssøgningssystem med forholdsvis begrænsede datastruktureringsmuligheder kan bringes til at fungere således, at man opnår de samme fordele som i et relationelt databasesystem.

Forsøget er beskrevet i detaljer i et særskilt LAMBDA-nummer, LAMBDA Nr. 6 (Nistrup Madsen 1988), hvorfor indlæggets indhold her gengives i forkortet form uden oplysninger og eksempler af system- eller programmeringsteknisk art.

Jeg vil gerne takke mine kolleger i DANLEX-gruppen, uden hvis opmuntring og støtte forsøget ikke kunne gennemføres. En speciel tak til Hanne Ruus for gode forslag og til Ebba Hjort, som har leveret eksempelmateriale.

1. BAGGRUND

Ordbogsartikler i videnskabelige ordbøger indeholder ofte et meget stort antal informationstyper, som indgår i forskellige relationer med hinanden. Ved edb-behandling af leksikografiske data, f.eks. ved lagring i et databasesystem, skal de logiske forbindelser mellem de forskellige oplysninger afspejles, således at relationerne kan anvendes ved søgning og præsentation af data.

Ved Handelshøjskolen i København har man i en årrække arbejdet med systemet DANSTATUS til forskellige terminologi- og ordbogs-

Simulering af relationel database
Bodil Nistrup Madsen
Proceedings of NODALIDA 1987, pages 286-300

projekter. DANSTATUS er den danske version af det engelske informationssøgningssystem STATUS II. I forbindelse med DANLEX-gruppens projekt "Lagring og behandling af maskinlæsbare leksikografiske data i databasesystemer" blev der gjort forsøg med lagring af data fra en videnskabelig ordbog, Gammeldansk Ordbog i DANSTATUS. Forsøget er beskrevet i Vestergaard (1987) og konklusionen er, at DANSTATUS ikke kan betragtes som et ideelt system, da afspejlingen af relationer mellem data ikke umiddelbart er mulig. Det konkluderes endvidere, at den systemtype, der skal bygges videre på, må være relationel.

DANSTATUS har imidlertid en række fordele, som er så vægtige, at det er interessant at undersøge, om man ved hjælp af nogle særlige programmeringsfaciliteter (macrofaciliteter) i DANSTATUS kan simulere en relationel database og derved opnå den ønskede strukturafspejling.

2. ARTIKELSTRUKTUREN I GAMMELDANSK ORDBOG (GLDO)

Som et led i projektet "Edb-behandling af videnskabelige ordbogsdata" har DANLEXgruppen udarbejdet en taksonomi til klassificering af leksikografiske data. Denne taksonomi beskrives i Descriptive Tools for Electronic Processing of Dictionary Data (1987). På basis af taksonomien er der udarbejdet et format til GLDO, som er anvendt ved indtastning af en række artikler ved hjælp af ordbogsredigeringssystemet Compulexis.

GLDO-formatet er inddelt i 4 afsnit:

I: identifikationsafsnit
B: bøjningsafsnit
S: semantisk afsnit
E: etymologisk afsnit

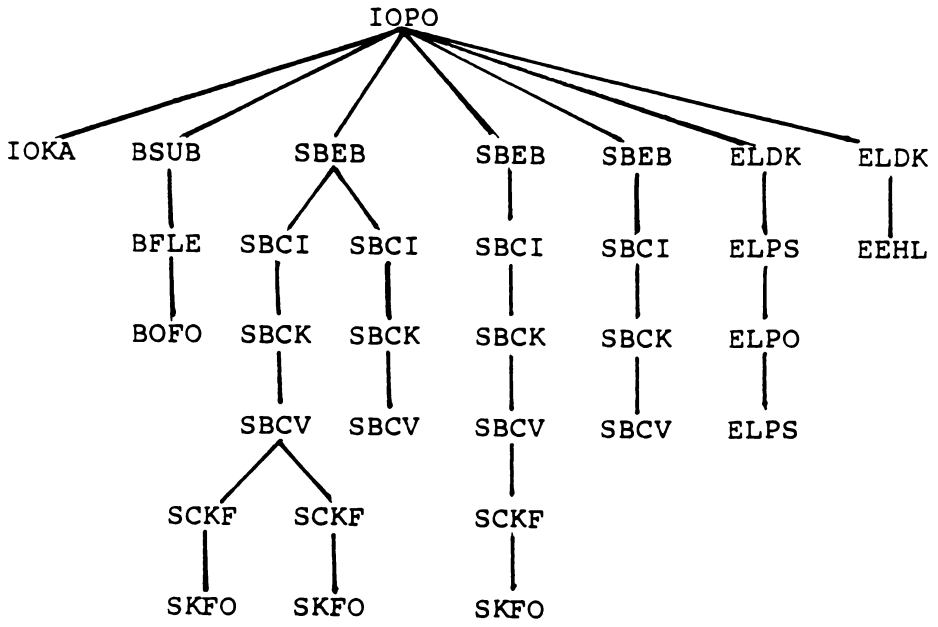
Formatet er beskrevet i Descriptive Tools (1987) og Ruus (1988). Til det her beskrevne forsøg anvendtes kun S-afsnittet, jfr. LAMBDA Nr. 6.

Nedenfor vises til eksemplificering en ordbogsartikel fra GLDO for opslagsordet "dræk", figur 1.

IHOM	IOPO dræk
IOKA	sb.
SEC	§
BSUB	n.
BØKK	Kv Rosg.
BØKV	76,15
SEC	§
BFLE	sg. bek.
BØFO	-ket
SEC	§
SBEB	snavs, smuds, skarn, spec. om ekskrementer
SBCI	ther efter skulle han al thenne verldens lyst vyrde sosom drek
SBCK	Suso.
SBCV	50,8
SCKF	træk
SKFO	Sv.
SCKF	stercus
SKFO	Lat. (jf. stercora.Filip.3,8(Vulg.))
SBCI	then indwolff, som vdsender eller vd skyuder drecket ok skarnet aff mennisken
SBCK	Kv Rosg.
SBCV	76,15
SEC	§
SBEB	overf.
SBCI	skrøbelig mænniske, som ær drek oc madek
SBCK	Suso.
SBCV	175,29
SCKF	putredo et vermis *(jf. Sir.19,3(Vulg.))*
SKFO	Lat.
SEC	§
SBEB	måske sammenblandet med dræg
SBCI	fex .. dreck eller berme .. fecula .. lyden drek vel berme
SBCK	Chr. Ped. Voc. 1510.
SBCV	63r
SEC	§
ELDK	fra
ELPS	mnt.
ELPO	dreck
ELPB	skarn etc.;
ELDK	egl. samme ord som thræk,
EEHL	jf. Bland.I.43

Figur 1: Ordbogsartikel fra GLDO for opslagsordet "dræk"

For tydeligere at vise den hierarkiske opbygning gengives ligeledes en træstruktur for artiklen "dræk".



Figur 2: Træstruktur for artiklen "dræk"

Et af de grundlæggende krav ved lagring af data er, som ovenfor nævnt, at relationerne mellem oplysningerne skal afspejles. Det vil f.eks. betyde, at sammenhængen mellem betydninger (SBEB) og tilhørende citater (SBCI) og kilder (SBCK) skal være entydig. Ved søgning i en database skal det således være muligt at få en selektiv udskrift af artiklen, omfattende f.eks. IOPO (opslagsord), IOKA (ordklasse), SBEB, SBCI og SBCK, jfr. figur 3.

IHOM	IOPO dræk
IOKA	sb.
SBEB	snavs, smuds, skarn, spec. om ekskrementer
SBCI	ther efter skulle han al thenne verldens lyst vyrde sosom drek
SBCK	Suso.
SBCI	then indwolff, som vdsender eller vd skyuder drecket ok skarnet aff mennisken
SBCK	Kv Rosg.
SBEB	overf.
SBCI	skrøbelig menniske, som ær drek oc madek
SBCK	Suso.
SBEB	måske sammenblandet med dræg
SBCI	fex .. dreck eller berme .. fecula .. lyden drek vel berme
SBCK	Chr. Ped. Voc. 1510.

Figur 3: Selektiv udskrift af artiklen "dræk"

Compulexis-artiklerne kan uden problemer overføres til DANSTATUS-poster, idet én Compulexis-artikel svarer til én DANSTATUS-post. I Vestergaard (1987) findes en detaljeret beskrivelse af overførslen. Her skal blot vises et eksempel, nemlig artiklen "dræk", figur 4.

```
IOPO        dræk
IOKA        sb.
BSUB        n.
BOKK        A089
BOKV        76,15
BFLE        sg. bek.
BOFO        -ket
SBEB        snavs, smuds, skarn, spec. om ekskrementer
SBCI        ther efter skulle han al thenne verldens lyst vyrde sosom drek
SBCK        A148
SBCV        50,8
SCKF        træk
SKFO        Sv.
SCKF        stercus
SKFO        Lat. (jf. stercora. Filip.3,8(Vulg.))
SBCI        then indwolff, som vdsender eller vd skyuder drecket ok skarnet
          aff mennisken
SBCK        A089
SBCV        76,15
SBEB        overf.
SBCI        skrøbelig mænniske, som ær drek oc madek
SBCK        A148
SBCV        175,29
SCKF        putredo et vermis (jf. Sir.19,3(Vulg.))
SKFO        Lat.
SBEB        måske sammenblandet med dræg
SBCI        fex .. dreck eller berme .. fecula .. lyden drek vel berme
SBCK        A020
SBCV        63 r
ELDK        fra
ELPS        mnt.
ELPO        dreck
ELPB        skarn etc.;
ELDK        egl. samme ord som thræk ,
EEHL        jf. Bland.I.43
```

Figur 4: Artiklen "dræk" overført til DANSTATUS

Hvis man imidlertid beder om en selektiv præsentation på skærmen, svarende til den i figur 3 viste, fås ikke det ønskede resultat, men i stedet udskriften i figur 5.

```
IOPO        dræk
IOKA        sb.
SBEB        snavs, smuds, skarn, spec. om ekskrementer
            overf.
            måske sammenblandet med dræg
SBCI        ther efter skulle han al thenne verldens lyst vyrde sosom drek
            then indwolff, som vdsender eller vd skyuder drecket ok skarnet
          aff mennisken
            skrøbelig mænniske, som ær drek oc madek
            fex .. dreck eller berme .. fecula .. lyden drek vel berme
SBCK        A148
            A089
            A148
            A020
```

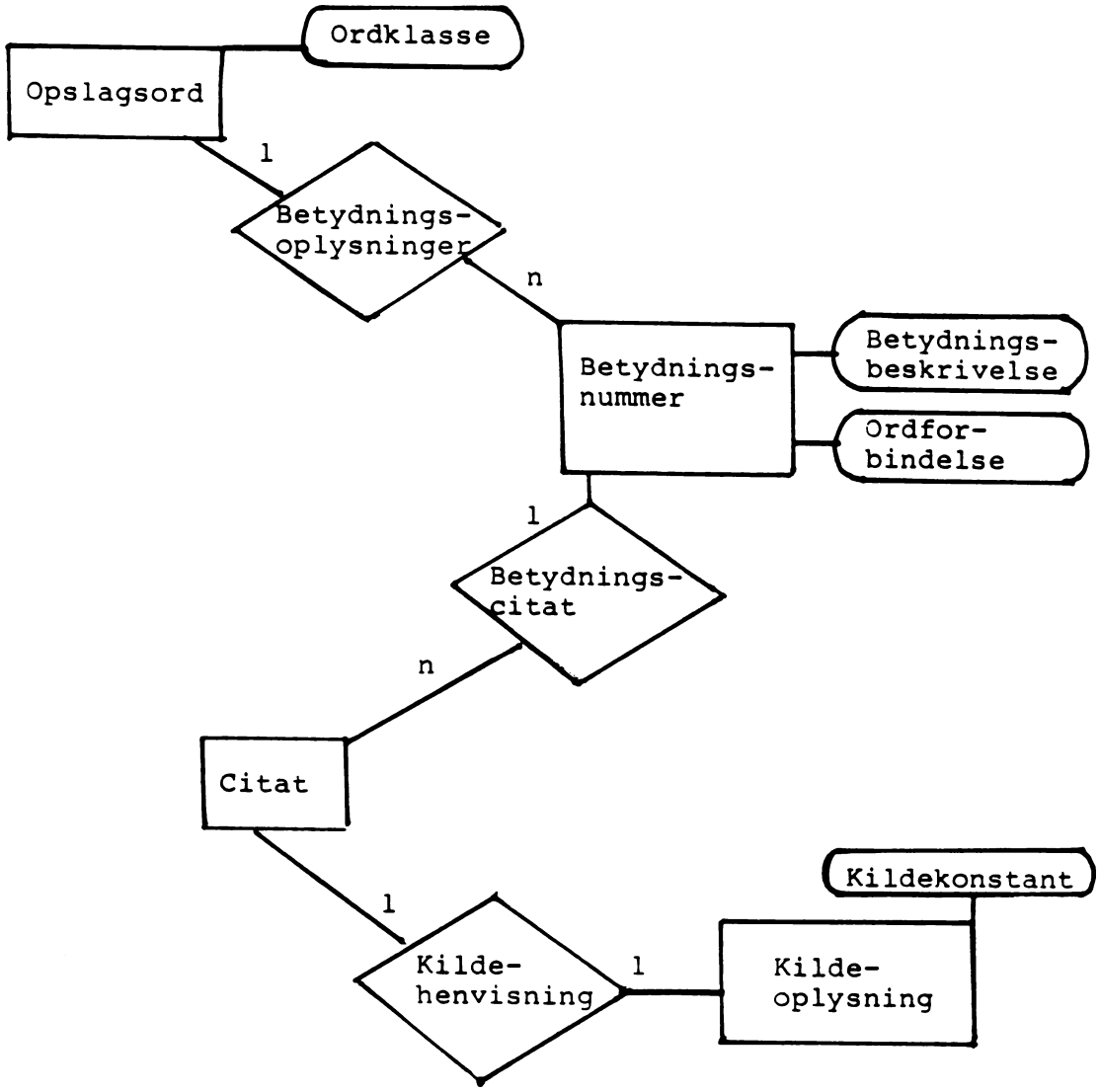
Figur 5: Selektiv udskrift af artiklen "dræk" i DANSTATUS

Denne udskrift er naturligvis utilfredsstillende, idet det ikke klart fremgår, hvilke betydninger, citater og kilder, der hører sammen. Årsagen er, at DANSTATUS opfatter alle forekomster af samme feltnavn i en artikel som ét felt fordelt på forskellige steder i posten.

3. DATASTRUKTURERING MED HENBLIK PÅ UDARBEJDELSE AF EN RELATIONSMODEL

I Ruus (1988) findes et Entitets-Relationsdiagram for S-afsnittet i GLDO. I Descriptive Tools (1987) er redegjort for hvorledes der til udarbejdelsen af et E-R-diagram for en ordbog kan anvendes både et indholds- og strukturbeskrivelsesværktøj, hhv. DANLEX-taksonomien og Warnier & Orr's klammediagram.

Med henblik på at udarbejde en relationsdatamodel, som skulle realiseres i DANSTATUS, valgtes et udsnit af S-afsnittet i GLDO, svarende til E-R-diagrammet i figur 6.



Figur 6: Udsnit af Entitets-Relationsdiagram for S-afsnittet i GLDO

Svarende til dette diagram kan der f.eks. opstilles en relationsdatamodel, som den i figur 7 viste.

I relationsdatamodellen såvel som i den efterfølgende beskrivelse er GLDO-feltnavnene udskiftet med navne, som er umiddelbart forståelige uden særligt kendskab til GLDO. Der anvendes følgende navne:

OPSL	opslagsord	= IOPO
ORDKL	ordklasseangivelse	= IOKA
BETYD	betydningsbeskrivelse	= SBEB
CITAT	betydningscitat	= SBCI
KILDE	betydningskildekonstant	= SBCK

opslagsord

OPSL	ORDKL

betydnings-
oplysninger

OPSL	BETNR	BETYD

betydnings-
CITAT

BETNR	CITAT	KILDE

Figur 7: Relationsdatamodel for udsnit af S-afsnittet i GLDO

I DANSTATUS opereres ikke med forskellige tabeller, indeholdende forskellige typer poster, sådan som det er tilfældet i et relationelt databasesystem. En DANSTATUS database består af én tekstfil og ét indeks, hvori der søges. En DANSTATUS database kan imidlertid indeholde poster med forskellig struktur, dvs. forskellige felter, og et feltnavn kan gentages inden for én post. Der afsættes ikke feltnavne eller plads til ikke udfyldte felter. Det er disse faciliteter, der er udnyttet ved forsøget med implementeringen af den relationelle datamodel.

I figur 8 er vist de tre tabeller, som er udarbejdet specielt med henblik på DANSTATUS. Tabellerne er opstillet således, at de muligheder der ligger i DANSTATUS macrofaciliteter udnyttes bedst muligt. Alle tre tabeller realiseres i én DANSTATUS database, som opbygges af tre forskellige posttyper (én for hver tabel).

opslagsord

OPSL	ORDKL	≠OPSL	≠GREN

betydnings-
beskrivelse

BETYD	≠OPSL	≠GREN

citat

CITAT	KILDE	≠OPSL	≠GREN

Figur 8: Relationstabeller for udsnit af S-afsnittet i GLDO

I tabellerne er der tilføjet nogle identifikations- eller nøglefelter:

≠OPSL	ordbogsartikelnummer (entydig identifikation af ordbogsartiklerne)
≠GREN	betydningsgrennummer (entydig identifikation af betydningerne i en ordbogsartikel)

I overensstemmelse med tabellerne i figur 8 oprettes tre posttyper i DANSTATUS, jfr. figur 9.

Posttype 1: opslagsordsposter
OPSL
ORDKL
ID
  ≠OPSL
  ≠GREN

Posttype 2: betydningsbeskrivelsesposter
BETYD
ID
  ≠OPSL
  ≠GREN

Posttype 3: betydningscitatposter
CITAT
KILDE
ID
  ≠OPSL
  ≠GREN

Figur 9: Posttyper i DANSTATUS

I figur 10 vises de 8 indlæste poster, som tilsammen udgør det valgte udsnit af den tidligere viste ordbogsartikel "dræk".

```
   1.
OPSL      dræk
ORDKL     sb
ID           #OPSL 1     #gren (1 2 3)

   2.
BETYD     snavs, smuds, skarn, spec. om ekskrementer
ID           #OPSL 1     #gren 1

   3.
CITAT     ther efter skulle han al thenne verldens lyst
          vyrde sosom drek
KILDE     Suso
ID           #OPSL 1     #gren 1

   4.
CITAT     then indwolff, som vdsender eller vd skyuder
          drecket ok skarnet aff mennisken
KILDE     Kv Rosg
ID           #OPSL 1     #gren 1

   5.
BETYD     overf.
ID           #OPSL 1     #gren 2

   6.
CITAT     skrøbelig mænniske, som ær drek og madek
KILDE     Suso
ID           #OPSL 1     #gren 2

   7.
BETYD     måske sammenblandet med dræg
ID           #OPSL 1     #gren 3

   8.
CITAT     fex .. dreck eller berme .. fecula .. lyden drek vel berme
KILDE     Chr Ped Voc 1510
ID           #OPSL 1     #gren 3
```

Figur 10: DANSTATUS-poster for artiklen "dræk"

4 SØGEMEMU BASERET PÅ MACROFACILITETERNE I DANSTATUS

Sammenkædningen af de enkelte poster, som udgør en ordbogsartikel, sker ved hjælp af søgninger på ≠OPSL og ≠GREN. Disse søgninger foregår skjult for brugeren, idet de er lagt ind i en række macroer ved hjælp af hvilke der er opbygget en særlig søgemenu.

I posttype 1, opslagsordsposterne, indføres i ≠GREN nummeret på samtlige grene i ordbogsartiklen. Herved opnås at man ikke blot kan kæde alle poster hørende til én ordbogsartikel sammen (ved hjælp af ≠OPSL), men at man også kan kæde udvalgte betydningsposter (grene) og opslagsordsposten fra én artikel sammen (ved hjælp af ≠GREN).

Menuen består af 4 hovedfaser:

(1) Søgning, herunder
- valg af søgeprofil,
- oplysning om antal svar,
- valg mellem at se
  - hele artiklen eller
  - udvalgte felter

(2) Valg mellem at se
- alle grene eller
- kun de grene hvori søgeordet findes

(3) Præsentation, herunder
- valg af profil

(4) Valg mellem at
- foretage ny søgning
- slutte

For en detaljeret gennemgang af søgemenuen og de udnyttede macrofaciliteter henvises til LAMBDA Nr. 6.

Søgemenuen bygger bl.a. på de erfaringer, der er indhøstet ved udvikling af en søgemenu til DANTERM, Dansk Termbank (Wegener 1986).

I figur 11 gengives et eksempel på en søgning ved hjælp af søgemenuen til GLDO.

```
**************************************
*                                    *
*   MENU til søgning i GLDO          *
*                                    *
**************************************
skriv søgeord
dreck
-
skriv søgefelt(er adskilt af komma)
citat

 søgeordet er fundet 1 gang i citat
 vil du se hele artikeln (y) eller nogle udvalgte felter (n)
y

1
OPSL      dræk
ORDKL     sb

BETYD     snavs, smuds, skarn, spec. om ekskrementer

CITAT     ther efter skulle han al thenne verldens lyst
          vyrde sosom drek
KILDE     Suso

CITAT     then indwolff, som vdsender eller vd skyuder
          drecket ok skarnet aff mennisken
KILDE     Kv Rosg

BETYD     overf.

CITAT     skrøbelig mænniske, som ær drek og madek
KILDE     Suso

 vil du se mere ja (y) eller nej (x)
y

BETYD     måske sammenblandet med dræg

CITAT     fex .. *dreck* eller berme .. fecula .. lyden drek vel berme
KILDE     Chr Ped Voc 1510

 vil du fortsætte søgningen (y) eller slutte (x)
y

skriv søgeord
dreck

skriv søgefelt(er adskilt af komma)
citat

 søgeordet er fundet 1 gang i citat
 vil du se hele artikeln (y) eller nogle udvalgte felter (n)
n

 vil du se alle grene (y) eller kun grene indeholdende søgeord (x)
x

 vælg præsentationsprofil (1, 2 eller 3):
1

1
OPSL      dræk
ORDKL     sb

BETYD     måske sammenblandet med dræg

CITAT     fex .. *dreck* eller berme .. fecula .. lyden drek vel berme
KILDE     Chr Ped Voc 1510
```

Figur 11: Eksempel på søgning ved hjælp af søgemenuen til GLDO

5 UDVIDELSER AF SØGEMENU

Som nævnt ovenfor er kun en del af S-afsnittet i GLDO-formatet inddraget i det beskrevne forsøg. Hvis søgemenuen skal udvides til at omfatte alle oplysningstyper i alle afsnit af GLDO-formatet, er der behov for mange nye posttyper og en betydelig udvidelse af macroerne. Såvidt det kan overskues, vil dette dog ikke medføre nye principielle problemer.

En forudsætning for at anvende metoden til et konkret projekt, er endvidere, at der udarbejdes særlige procedurer til indlæsning og ajourføring af data.

Endvidere bør menuen udvides med hjælpetekster og sikring mod forkerte svar fra brugerens side, jfr. Wegener (1986).

6 KONKLUSION

Forsøget med simulering af relationel database har vist, at der er langt bedre muligheder for datastrukturering i DANSTATUS, end den hidtidige anvendelse af systemet har tydet på.

Som tidligere nævnt anvendes DANSTATUS til en række terminologi- og ordbogsprojekter. Som eksempler kan nævnes DANTERM (Dansk Termbank) og Dansk-Fransk ordbogsbase (Blinkenberg & Høybyes Dansk-Fransk Ordbog). I begge projekter er der tale om hierarkisk strukturerede data, dog med færre niveauer end i GLDO. Der er således også her behov for en bedre afspejling af relationerne mellem data, end det er muligt at opnå ved den almindelige anvendelse af DANSTATUS, hvor én termbank- eller ordbogsartikel svarer til én post i systemet.

Forudsat at der kan udarbejdes hensigtsmæssige indlæsnings- og ajourføringsprocedurer, forekommer det derfor hensigtsmæssigt at anvende den beskrevne metode til simulering af relationel database til de to ovenfor nævnte, såvel som til andre lignende projekter.

REFERENCER

Descriptive Tools for Electronic Processing of Dictionary Data (1987), Studies in Computational Lexicography, The DANLEX Group, Danish Working Group on Computational Lexicography: Ebba Hjorth, Jane Rosenkilde Jacobsen, Bodil Nistrup Madsen, Ole Norling-Christensen, Hanne Ruus. (Lexicographica Series Maior 20), Tübingen, Niemeyer.

Nistrup Madsen, Bodil (1988): Simulering af relationel database. (LAMBDA Nr. 6), Institut for Datalingvistik, Handelshøjskolen i København.

Ruus, Hanne (1988): Lexical Data Structures. Indlæg på XIV ALLC konference i Göteborg 1987. Udkommer i Literary and Linguistic Computing 1988.

Vestergaard, Bodil (1987): Undersøgelse af databasesystemer til ordbøger. (LAMBDA Nr. 2), Institut for Datalingvistik, Handelshøjskolen i København.

Wegener, Helle (1986): Forslag til et menubaseret brugerinterface til DANSTATUS opbygget ved hjælp af systemets macrofaciliteter med henblik på søgning i HHK´s termbank DANTERM, Prøve 3, Anvendt Sprogvidenskab, Linie 1: Datamatisk Lingvistik.