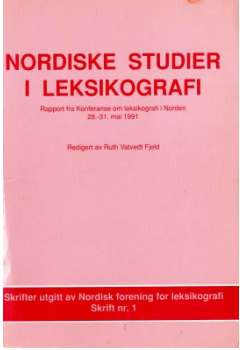


NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	Hvad får man skudt i skoene? Flerordsenheder i aktive ordbøger for mennesker og maskiner	
Forfatter:	Frede Boje og Anna Braasch	
Kilde:	Nordiske Studier i Leksikografi 1, 1992, s. 483-500 Rapport fra Konferanse om leksikografi i Norden, 28.-31. mai 1991	
URL:	http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive	

© Nordisk forening for leksikografi

Betingelser for brug af denne artikel

Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

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.

Frede Boje og Anna Braasch

Hvad får man skudt i skoene? Flerordsenheder i aktive ordbøger for mennesker og maskiner

Det arbejde, vi her fremlægger, er udsprunget af et samarbejde inden for den danske EUROTRA-gruppe, men har udviklet sig til et samarbejde mellem artiklens to forfattere som repræsentanter for hhv. EUROTRA-DK (som nu er en del af det nyoprettede Center for Sprogteknologi) og en projektgruppe ved Handelshøjskolen i København. Formålet med det igangværende arbejde er at samle og klassificere et så omfattende materiale af eksempler på flerordsenheder, at vi vil kunne registrere de mønstre, der forhåbentlig vil tegne sig, og på grundlag heraf udvikle en beskrivelsesmodel, som er eksplicit og udtømmende nok til både at være implementerbar i NLP og kunne danne basis for en aktiv ordbog for mennesker. Dette materiale hentes fortrinsvis i maskinlæsbare korpura som DK87-90 og sammenholdes med det materiale, som er beskrevet i Erik Bruun: Dansk Sprogbrug. Metoden og målsætningen omfatter både leksikografiske og datamatiske aspekter, og udbyttet af vekselvirkningen imellem disse aspekter anser vi for at være det mest spændende i dette projekt.

0. Indledning: To indfaldsvinkler: human- og maskinoversættelse

Når ovenstående udtryk er valgt som titel, er det for at illustrere den ene af de to indfaldsvinkler til vores arbejde; det er nemlig et eksempel på et konkret oversættelsesproblem, vi er stødt på i forbindelse med undervisningen af danskere i oversættelse fra dansk til tysk. På dansk har vi udtrykket:

at skyde nogen noget i skoene

og på tysk udtrykket:

jemandem etwas in die Schuhe schieben.

Det er nærliggende at tro, at de to udtryk betyder det samme og bruges på samme måde, og den opfattelse får man bestyrket, hvis man slår op i Gyldendals røde ordbøger Dansk-tysk og Tysk-dansk:

I. *sko*: Schuh, [...]; skyde én ngt i -ene: e-m etw in die Schue [sic!] schieben, e-m etw unter' stellen; [...]

(Bork 1990a)

Schuh m -e. 1. sko, [hoher] ~ støvle (e-m etw in die -e schieben); [...]

(Bork 1990b)

Vores erfaring siger os imidlertid, at der er en forskel i anvendelsen af udtrykket på dansk og tysk, idet de leksikalske og syntaktiske restriktioner for objektet er forskellige, men problemet er: Hvor finder man tilstrækkelig præcise oplysninger om, hvordan udtrykket bruges på dansk og/eller tysk?

Der er os bekendt ikke nogen, som har foretaget omfattende og systematiske undersøgelser af, hvorledes den slags udtryk er behandlet i danske ordbøger, mens der er i hvert fald to, som - næsten samtidig, men uafhængigt af hinanden - har undersøgt tyske ordbøgers behandling af fraseologiske udtryk. Det drejer sig om Anne Lise Kjær og Anna Braasch, som begge først har skrevet en specialeafhandling (Kjær 1986) og (Braasch 1987) og dernæst en artikel, som sammenfatter afhandlingens resultater (Kjær 1987) og (Braasch 1988).

Ikke uventet konstaterer man, at de eksisterende opslagsværker ikke giver tilfredsstillende svar på vores spørgsmål om restriktionerne for objektet¹.

I vores korpusundersøgelser har vi fundet meget få eksempler på brugen af udtrykket i overskriften, men sammen med de oplysninger, som kan stykkes sammen fra de forskellige ordbøgers, underbygger de vores opfattelse: Mens objektet på tysk normalt er substantivet "die Schuld"², føles "skylden" som objekt på dansk som "dobbelts konfekt"³. På dansk er objektet det, man får skylden for - som regel, men ikke nødvendigvis, udtrykt ved en at-sætning⁴, hvilket tilsyneladende ikke er muligt på tysk.

Oversættelsesproblemer som dette, som især dukker op ved oversættelse fra modersmål til fremmedsprog (men ikke kun dér), har fået os til at stille spørgsmålet: Hvordan udarbejder man en leksikalsk beskrivelse, som er detaljeret og præcis nok til at give svar på den slags spørgsmål?

Vores anden indfaldsvinkel vil vi i det følgende give en noget grundigere redegørelse for. Det drejer sig om dét behov for en detaljeret og præcis beskrivelse af leksikalske enheder, man har i arbejdet med maskinoversættelse. Men først vil vi uddybe, hvad det er for udtryk, vi ønsker at beskrive (afsnit 1), og kort omtale de særlige krav, maskinoversættelsen stiller til ordbøger og grammatikker (afsnit 2).

1. Definition af flerordsenheder - kriterier - typologi

Desværre bruger forskellige forfattere et utal af forskellige betegnelser om de fænomener, vi undersøger, samtidig med, at mange af betegnelserne, fx "kollokation", bruges med forskelligt indhold. Pilz (1978:8) klager over "diese geradezu chaotische terminologische Vielfalt", og situationen synes ikke at være blevet bedre siden da⁵.

Det, vi ønsker at beskrive, er, hvad vi - som et forhåbentlig neutralt overbegreb - vil kalde *flerordsenheder*. Ved en flerordsenhed forstår vi:

En sproglig størrelse, som består af mere end ét ord, og hvis betydning og/eller brug ikke kan beskrives udtømmende alene ved generelle syntaktiske regler og en leksikalsk beskrivelse af enkeltordene.

Når vi nedenfor giver et forslag til, hvorledes man kan inndele flerordsenhederne, betyder det ikke, at vi har et gennemarbejdet system byggende på klare definitioner, men at vi anser det for nødvendigt at skitsere, hvordan de forskellige fænomener efter vores mening er relateret. Vores hovedinteresse ligger imidlertid ikke i denne "grov"-typologi, men, som det vil fremgå

af det følgende, i en systematisk klassifikation af de forskellige typer kollokationer og fraseoleksemer.

For brugeren - inklusive edb-maskinen - er det underordnet, om et givet udtryk kaldes en kollokation, et fraseoleksem eller noget tredje. Det afgørende for ham (eller den) er, at fænomenet er registreret og beskrevet på en måde, som sætter brugeren - og altså også maskinen - i stand til at vælge den korrekte form (evt. en af de korrekte former). Med udtrykket "at vælge" tænker vi her både på receptionsfasen - inklusive den maskinelle analyse - og produktionsfasen.

Ved arbejdet med registreringen og klassifikationen af udtrykkene er det efter vores mening hensigtsmæssigt at bruge størrelse og stabilitet som hovedkriterier. Men ved en udtømmende teoretisk *beskrivelse* af flerordsenhederne er der mange flere faktorer, der skal inddrages. En sådan beskrivelse ville imidlertid fuldstændig sprænge rammerne for denne artikel. Vi vil derfor begrænse os til at nævne de vigtigste faktorer i stikordsform og kort kommentere dem. Det er oplagt, at der er en nær sammenhæng og indbyrdes afhængighed mellem flere af faktorerne.

- 1 Størrelse og ordklassetilørsforhold: 2 ord - sætningsled - sætning - periode - flere perioder.
- 2 Stabilitet: fra fuldstændig fri kombination til helt fast forbindelse.
- 3 Metaforik: konkret - metaforisk.
- 4 Opacitet: transparent - opak.
- 5 Kompositionalitet: kompositionel - ikke-kompositionel.
- 6 Hyppighed: højfrekvent - lavfrekvent kombination.
- 7 Normbundethed: afhængighed af sproginterne og/eller sprogeksterne normer.
- 8 Oversættelighed.
- 9 Forudsigelighed.

Hertil kommer et par faktorer, som ikke er væsentlige for, hvordan de forskellige typer flerordsenheder defineres, men som til gengæld har betydning for, hvilke af typerne man især vil beskæftige sig med, nemlig:

- 10 Teksttyper.
- 11 Emneområder.

Ad 1. Pr. definition må en flerordsenhed bestå af mindst to ord, mens definitionen ikke sætter nogen fast grænse opad. Da vi ønsker at registrere udtryk, som det vil være relevant at have som ordbogsindgange i en maskinlæsbar ordbog - og det vil i praksis stort set svare til lemmaer og sublemmaer i en almindelig ordbog - interesserer vi os især for størrelser, der tilsammen enten udgør ét sætningsled, typisk et NP uanset ledfunktion, eller danner kernen i en verbalfrase, typisk verbum + objekt(er). Derfor har vi heller ikke foretaget en underinddeling af længere idiomatiske udtryk, som utvivlsomt er relevant i anden sammenhæng⁶.

Ad 2. Stabilitet. De to yderpunkter er ret uproblematiske. Til gengæld er det beskrivelsen af de fænomener, som ligger derimellem, der især er interessant. Er de restriktioner, der gælder for variationsmulighederne for de enkelte elementer i flerordsenhederne morfologiske eller syntaktiske? Er de bundet til en lukket liste af ord eller til klasser af ord med fælles semantiske egenskaber? Er der en afhængighedsrelation mellem udfyldningen af

forskellige pladser i et komplekst udtryk? Det er alle disse faktorer, vi først og fremmest vil belyse, monolingvalt og kontrastivt.

Ad 3. Metaforik. Mange flerordsenheder er mere eller mindre levende metaforer, og den rutinerede oversætter ved, at metaforer ofte giver særlige oversættelsesproblemer, men metaforik er næppe egnet som inddelingskriterium.

Ad 4-5. Opacitet og kompositionalitet er to betegnelser, der - med forskellig synsvinkel - bruges til at beskrive det samme fænomen. Samtidig har de nær relation til pkt. 9, forudsigelighed. Der skelnes gerne mellem forskellige grader af opacitet, mens kompositionalitet/ikke-kompositionalitet oftest ses som en dikotomi. En kompositional konstruktion vil være transparent, en ikke-kompositional mere eller mindre opak. Begge begreber står i relation til afkodning (reception). Selv om kompositionalitetsbegrebet er problematisk, mener vi, at det er rimeligt at bruge det som kriterium for, om en leksikalisk betinget flerordsenhed er et idiomatisk udtryk eller en kollokation.

Ad 6. Hyppigheden af en kombination anses af nogle for et kriterium for, om den er en kollokation, se fx Bergenholtz (1989:777). Efter vores mening er høj forekomstfrekvens et stærkt indicium for, at en ordkombination er en flerordsenhed, men ikke anvendeligt som kriterium.

Ad 7. Normbundethed. Vi mener, at den tredeling i "system - norm - sprogbrug", som Coseriu har indført og v. Polenz har videreudviklet⁷, er bedre egnet til at forklare nogle af de mekanismer, der er på spil, end den Saussure'ske todeling i langue - parole, men tror ikke, at det er nødvendigt at inddrage normbegrebet som klassifikationskriterium. I sin afhandling om "Normbetingede ordforbindelser i tysk juridisk sprog" (Kjær 1990a) har Anne Lise Kjær påvist, at retsvirkninger kan være afhængige af, at en sags parter overholder præskriptive normer for formuleringer, men så vidt vi kan skønne, er der her tale om noget, som især (og måske kun) er relevant for juridisk sprogbrug, så vi vil tillade os i første omgang at lade problemet ligge.

Ad 8-9. Forudsigelighed og oversættelighed hører til en vis grad sammen. Hvis et givet syntagma er fuldstændig forudsigeligt, altså hvis det kan konstrueres korrekt alene ud fra kendskab til generelle syntaktiske regler og enkeltordenes betydning, skulle udtrykket ud fra den ovenfor nævnte definition falde uden for emnet. Problemet er, at det ofte først er i forbindelse med oversættelse, at man opdager, at der er nogle restriktioner. Hausmann (1989:1013) sætter sagen på spidsen ved at sige, at hvis man sammenligner ækvivalerende udtryk på alle verdens sprog, vil resultatet nok blive, at et hvilket som helst udtryk, som formodes at være en fri forbindelse, viser sig at være en kollokation. Selv om det selvfølgelig er en vild overdrivelse, er der mere om snakken, end folk, der ikke arbejder kontrastivt, i almindelighed tror.

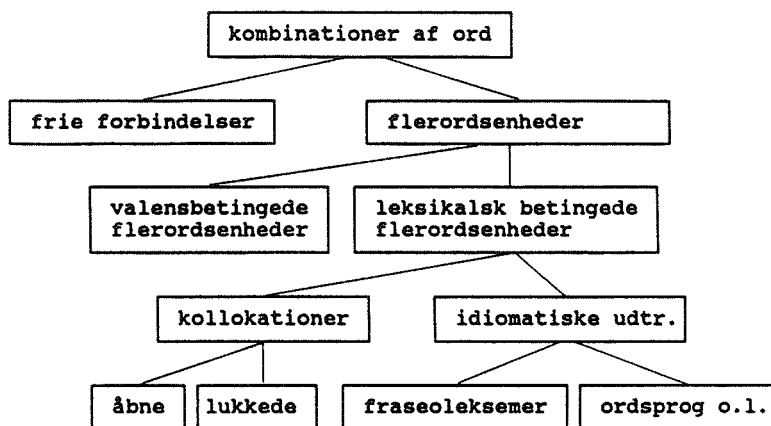
I en redegørelse for sit arbejde med Grunddeutsch nævner Pfeffer (1979:311f), at i ca. 80 % af korpusbelæggene for verbet *halten* indgår det i "redensartige Bindungen", og at hverdags sproget er mere end 15 % "redensartig gebunden". Selv om disse oplysninger skal tages med et vist forbehold, antyder de problemets omfang, specielt for de to former for oversættelse, vi fokuserer på: maskinoversættelse og oversættelse til fremmedsprog.

Forudsigelighed er et begreb, som er temmelig umuligt at operere med, i hvert fald hvis man ikke begrænser det til kun at være forudsigelighed i forbindelse med oversættelse, og i så fald skal man tage stilling til, hvilken type oversættelse mellem hvilke sprog.

Konklusionen på dette er, at selv om netop den manglende forudsigelighed er en af hovedårsagerne til, at der er behov for en fyldestgørende beskrivelse af flerordsenheder, er forudsigelighed uanvendelig som egentligt klassifikationskriterium.

Ad 10-11. På grund af formålet med vores projekter er det klart, at vi ikke beskæftiger os særligt med fænomener, som udelukkende eller helt overvejende optræder i talesprog eller i skønlitteratur. Vi vil her henvise til Skyum-Nielsen (1991), som ud fra en anden indfaldsvinkel behandler mange af disse fænomener. Vi lægger hovedvægten på, hvad der er vigtigt i fagsproglige tekster samt andre non-fiktive tekster, altså fx avisartikler. Det indebærer bl.a., at forbindelser, som er okkasionelle, altså ad hoc-dannelser, og anden kreativ sprogbrug principielt ikke inddrages⁸.

Den typologi, som er det foreløbige resultat af vores overvejelser, er vist i fig. 1:



Figur 1

Skemaet er bl.a. inspireret af Sonja Poulsens og Peder Skyum-Nielsens arbejder⁹, men er mere detaljeret end Poulsen og mindre detaljeret (og principielt anderledes) end Skyum-Nielsen. Vi må igen her afstå fra en fuldt udbygget argumentation og blot kort kommentere de vigtigste distinktioner.

Det, vi her har kaldt "valensbetingede flerordsenheder", svarer stort set til, hvad BBI (1986) kalder "grammatical collocations". Når vi skiller dem ud som en særlig gruppe, skyldes det ikke kun, at de af de fleste forskere med god grund ikke regnes med til kollokationerne, men også, at det er fænomener, som kan beskrives eksakt og udtømmende med allerede kendte metoder, også til brug ved maskinel analyse og oversættelse. Inden for EUROTRA er der udviklet et meget fintmasket system til valensbeskrivelse. Systemet er implementeret i ordbøger og grammatikker, og det fungerer.

Kriteriet for distinktionen mellem kollokationer og idiomatiske udtryk er idiomaticiteten: kollokationerne er ganske vist uforudsigelige, men kan forstås kompositionelt, mens de idiomatiske udtryk hverken i produktion eller reception er kompositionelle. Fraseoleksemene er (som de almindelige leksemer) "byggeklodser" i sætninger, men aldrig hele sætninger.

2. Aktive ordbøger for mennesker og maskiner

Ordbøger for mennesker, både et- og tosprogede, aktive og passive udnytter (omend på forskellig måde og i forskellig grad) brugerens evne til at

- * generalisere ud fra givne sproglige eksempler og dermed danne analogier;
- * udlede oplysninger der er implicit givne i eksempler og dermed få adgang til 'skjulte' regulariteter eller uregelmæssigheder;
- * overføre sin viden om sprogs udtryksmuligheder fra modersmålet til målsproget i form af sprogkompetence og intuition;
- * inddrage sin almenviden om verden.

I automatisk behandling af naturlige sprog (Natural Language Processing, NLP), herunder maskinoversættelse (Machine Translation, MT), kan man ikke bygge på disse evner. Dette betyder, at maskinen skal have kompenseret for det manglende vidensgrundlag med eksplicite oplysninger i en eller anden form. Derfor vil en ordbog til NLP altid indeholde flere og mere detaljerede oplysninger, også om kildesprogets enheder, end en aktiv ordbog for mennesker ved samme enhed (opslagsord).

2.1 EUROTRAs ordbogssystem

EUOTRAs oversættelses- og dermed også dets ordbogssystem bygger på princippet om multilingvalitet. Dette indebærer på det overordnede plan, at der for hvert sprog udarbejdes en monolingval ordbogs-komponent. Vi vil her ikke gå nærmere ind på, hvordan denne komponent er opbygget, blot nævne, at systemet fordeler oversættelsesprocessen på flere niveauer. Hvert niveau har sin egen ordbog, og der er en "oversættelses"ordbog mellem hvert niveaupar. For overskuelighedens skyld omtales i det følgende kun grænsefladeordbøgerne, der indholdsmæssigt mest ligner aktive ordbøger for mennesker¹⁰.

Den egentlige oversættelsesordbog, den såkaldte transferordbog, knytter ideelt kun ordbogsindgangene fra to monolingvale grænsefladeordbøger sammen. Dette betyder, at de monolingvale MT-ordbogs-komponenter stiller større krav til detaljeringsgrad og systematik end ordbøger for mennesker.

Ved analysen af en tekst, inden den kan oversættes, skal også flerordsenhederne identificeres, deres komponenter samles til en leksikalsk enhed; dette er særlig vigtigt for sådanne flerordsenheder, hvis betydning ikke kan forstås ud fra komponenternes.

Principielt skulle de monolingvale ordbøger kunne kombineres på kryds og tværs i EUROTRA-systemet. Der er dog en del væsentlige problemer, som endnu er uløste, det gælder især på området flerordsenheder.

2.2 Grammatik og ordbog i MT

Rollefordelingen mellem grammatik og leksikon (dvs. ordbog) er nøje defineret i MT-systemer (og for NLP i almindelighed). Grammatikken beskriver det generelle (regler for ordstilling, bøjningsmønstre...), leksikonet beskriver det individuelle, ordenes egenskaber hver for sig (ordets betydning(er), rektion, kompatibilitet med andre ord, genus, bøjning...). Som individuelle egenskaber gælder såvel subkategorisering, dvs. ordets syntaktiske forbindelighed,

som selektionsregler, dvs. ordets semisk/sememiske forbindelighed. Det er velkendt, at syntaktiske egenskaber lettere lader sig systematisere og beskrive i regler end semantiske egenskaber.

Interessant er det i denne sammenhæng at iagttage de bestræbelser inden for den moderne leksikografi, der har til hensigt at afklare, hvilke grammatiske oplysninger der er brug for i store monolingvale ordbøger (jf. fx bidragene i Bergholtz/Mugdan 1985).

2.3 Støtteverbumbonstruktioner

Som nævnt ovenfor (i afsnit 2.2) udgør flerordsenheder et stort problem i maskinoversættelse, især når systemet er multilingvalt og ikke kan udnytte "tilfældige" ligheder mellem to bestemte sprog. Men der er en velafgrænset gruppe af flerordsenheder, som det inden for EUROTRA er lykkedes at behandle på en både teoretisk og praktisk tilfredsstillende måde, nemlig en konstruktionstype, som på engelsk kaldes "support verb constructions". Vi fordanser det til "støtteverbumbonstruktioner", men bruger for nemheds skyld også forkortelsen SVC. Det drejer sig om udtryk som:

træffe en beslutning om noget
foretage en undersøgelse af noget.

Disse konstruktioner henregnes i reglen til kollokationer (altså leksikalske), men kan beskrives med et system, der ligner valensbeskrivelsen, og ligesom den egentlige valens er systemet implementeret i EUROTRA og har vist sig at fungere.

Når man inden for EUROTRA har valgt først at koncentrere sig om denne konstruktionstype, skyldes det to ting. For det første er konstruktionstypen meget almindelig i alle EF-sprogene, specielt i administrative tekster, og giver ofte problemer ved oversættelse. For det andet har det vist sig muligt at lave definitioner og beskrivelsesmodeller, som er anvendelige for alle sprogene.

SVC'erne er ikke identiske med det, der i den germanistiske tradition kaldes "Funktionsverbgefüge", men er groft sagt en delmængde af dem. Kernen i en SVC er et substantiv, som har en argumentstruktur, altså er valensbærende. Typisk vil der være tale om et logisk subjekt (kaldet argument 1) og et logisk objekt (argument 2), fx:

Folketingets (arg1) *beslutning* om nye skatter (arg2)

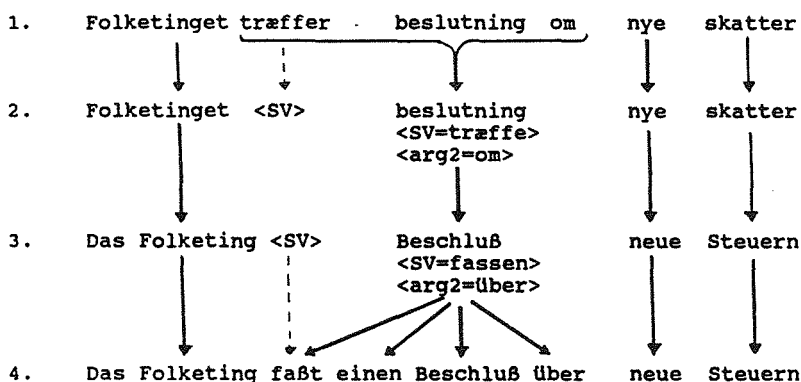
Hvis et sådant valensbærende substantiv står som objekt for et verbum, som i konstruktionen er uden egentligt leksikalsk indhold, kaldes verbet for et *støtteverbum* og helheden for en SVC, fx:

Folketinget *træffer* beslutning om nye skatter

Oplysningen om, hvilket støtteverbum et substantiv forbindes med, anbringes i ordbogsindgangen for substantivet. Det gør, at man i oversættelsesordbøgerne slipper for at oversætte støtteverberne, hvilket ellers er en meget besværlig opgave ved maskinoversættelse, fordi målsprogsverbet ikke er forudsigeligt ud fra kildesprogsverbet, men bestemmes af substantivet på målsproget.

Når substantivet er oversat, kan man derfor blot hente det rigtige støtteverbum i ordbogsindgangen for substantivet på målsproget, hvor man også henter oplysninger om bl.a. valensbundne præpositioner.

Stærkt forenklet ser oversættelsesprocessen ved oversættelse fra dansk til tysk altså ud som vist i fig. 2:



Figur 2

Der er mange aspekter i forbindelse med SVC'er, som vi ikke kan komme ind på her, fx hvorledes man har løst problemet med oversættelse mellem SVC'er og simplex-verber, men et væsentligt formål med vores projekt er at gøre det muligt at beskrive andre typer flerordsenheder tilsvarende eksplicit og implementerbart.

3. Metode

Der findes på dansk et opslagsværk, som - uden at begrænse sig til flerordsenheder - indeholder en ret omfattende samling af den slags udtryk, nemlig Erik Bruuns *Dansk Sprogbrug* (DS). Vi har derfor fundet det naturligt at tage udgangspunkt i denne bog. Ud over - som nævnt i afsnit 1 - at forsøge at underbygge det teoretiske fundament for typologisering og klassifikation af flerordsenheder består arbejdet af følgende faser:

- 1 En undersøgelse af, hvordan DS beskriver flerordsenheder: Hvordan er makro- og mikrostrukturen? Hvordan er variabiliteten angivet, først og fremmest m.h.t. syntaktiske og semantiske oplysninger om de variable elementer?
- 2 En empirisk undersøgelse af, hvilke eksempler på flerordsenheder der findes i et større maskinlæsbart dansk korpus (DK87-90)¹¹.
- 3 En sammenligning af korpuseksemplerne med beskrivelsen i DS og i et vist omfang andre opslagsværker med henblik på at skabe det empiriske grundlag for næste fase.

- 4 En udtømmende, eksplicit og entydig registrering af disse enheders syntaktiske og semantiske egenskaber, specielt deres variabilitet. Registreringen udformes således, at den kan bruges i maskinel analyse og oversættelse, men samtidig kan omsættes til en "menneskevenlig" beskrivelse i en aktiv ordbog for mennesker.

Faseopdelingen skal ikke forstås således, at fase 1 afsluttes, før fase 2 påbegyndes osv., men således, at faserne gennemløbes flere gange. Vi har afsluttet første gennemløb, der nærmest har haft karakter af et pilotprojekt, idet vi har gennemløbet faserne med et begrænset materiale, og bruger i øjeblikket bl.a. resultaterne af dette arbejde til at justere det teoretiske apparat.

Vi har altså valgt at tage udgangspunkt i DS's systematisering, der er baseret på opslagsordets syntaktiske funktion i den pågældende flerordsenhed. Denne typologi er velegnet som grundlag for vore undersøgelser, fordi den er så overskuelig, at den også kan appliceres på store datamængder såsom ekscerpter fra *DK87-90*. På den anden side er den så grovmasket, at der i det videre arbejde må foretages yderligere opdelinger.

Desværre indeholder de indledende afsnit i DS ingen oplysning om, hvilke overvejelser der ligger til grund for udvalget af de sprogbrugseksempler, der indgår i denne "stil- og konstruktionsordbog" (som undertitlen lyder), altså om forfatteren har haft en anden definition (eller intuitiv opfattelse) af, hvad der udgør flerordsenheder, eller om han bevidst ikke har ønsket at skelne mellem flerordsenheder og frie forbindelser. I hvert fald er det tydeligt, at DS anfører eksempler på begge dele uden at markere nogen skelnen mellem dem. At få noget nærmere at vide om indsamlingsmetoden og -kriterierne ville også have været af interesse for vort projekt. Vi ved, at der ligger en del yderligere materiale hos forfatteren i mere eller mindre bearbejdet form. Frasorteringen af dette materiale skete af redaktionelle hensyn.

Systematikken i DS har to sider: For det første omfatter den ordbogsartiklernes struktur: "Eksempler på ensartede anvendelser er samlet under samme bogstavbetegnelse", dvs. at ordbogsartiklen er delt op i afsnit på grundlag af opslagsordets syntaktiske funktion:

Ved substantiver og adjektiver:

- a: som subjekt (formelt og egentligt);
- b: som objekt (direkte og indirekte) samt prædikativ;
- c: som styrelse for præposition og efter konjunktion;
- d: som underled;
- e: med underled og ved sideordning.

Ved verber:

- a: med subjekt;
- b: med objekt og prædikativ;
- c: med adverbial;
- d: som underled;
- e: som styrelse, [ikke-verbalt] sætningsled samt ved sideordning.

For det andet angives flerordsenhedernes variabilitet, altså hvilke led der er obligatoriske og fakultative led; hvilke elementer der er faste og variable; samt de enkelte leds alternationsmuligheder. Dette sker ved hjælp af et markeringssystem med parenteser og skråstreger:

- [] for fakultativ komponent;
- [] [] sidestilling af to eller flere fakultative komponenter;
- () obligatorisk komponent, eksemplificeret;
- [()] fakultativ komponent, eksemplificeret;
- / alternerende komponenter (enkeltord);
- // alternerende komponenter (ordforbindelser);
- skilletegn inden for samme afsnit, der kan omfatte flere syntaktiske funktionstyper.

3.1 Bearbejdning af DS-materiale

Vi har udvalgt nogle artikler fra DS, der umiddelbart så ud til at rumme mange forskellige slags ordforbindelser. Disse artikler (med opslagsordene *hånd, hoved, slutte, holde, hjælp, hjælpe* m.fl.) har vi fået adgang til i maskinlæsbar form.

Artiklerne er blevet foldet ud, dvs. at den komprimering, der er nødvendig af pladshensyn i en trykt ordbog, er blevet ophævet: Tilde og tankestreg er blevet erstattet med opslagsordet, og i alle tilfælde, hvor der er angivet alternative eller fakultative udfyldninger af en plads i et udtryk, er det samlede udtryk anført én gang for hver udfyldningsmulighed. Denne udfoldning, som kun delvis kan automatiseres, er nødvendig for at muliggøre en følgende maskinel strengsammenligning med korpusbelæg.

3.2 Korpusarbejde

Som nævnt ovenfor, er det korpus, vi har anvendt, DK87-90, som rummer 1 million løbende ord pr. årgang, altså i alt 4 millioner ord¹². Ved hjælp af et søgeprogrammel excerperede vi alle forekomster af et givet ord i alle bøjningsformer. Eksemplerne blev hentet som KWIC (Key Word in Context) belæg, med tilstrækkelig lang venstre og højre kontekst, således at belæggene altid gav meningsfyldte eksempler på søgeordets brug.

Det viste sig, at vi fik en temmelig stor mængde data ud for hver søgning (i alt ca. 3 MB). Bearbejdelsen af disse data var ganske tidkrævende. Vi vil her kun komme ind på eksemplerne med ordet *hånd*.

Vi har undersøgt ordet og dets kontekst systematisk ud fra følgende overordnede kriterier:

- * Sortering af eksempler i følgende grupper:
 - frie ordforbindelser
 - okkasionelle konstruktioner, herunder også kontaminationer
 - ægte flerordsenheder.

- * Grovtypologi inden for flerordsenheder på grundlag af:
 - opslagsordets (mulige) syntaktiske rolle(r)
 - opslagsordets kontekstpartnere.
- * Sortering inden for grovtypologiens klasser i to typer:
 - lukkede enheder (med høj leksikalsk stabilitet)
 - åbne enheder (med leksikalsk variabilitet af forskellig art og grad).
- * Registrering af morfosyntaktiske varianter.

At denne undersøgelse kun kan give et fingerpeg om, hvor mange forskellige kombinationer af variabilitet og restriktioner der findes i flerordsenheder, er vi godt klar over. Det viste sig imidlertid ret tydeligt ved sammenligningen af de bearbejdede korpusbelæg og DS's typologi hhv. markeringssystem, på hvilke punkter der først og fremmest er brug for en videregående analyse. Dette vil blive omtalt ved sammenligningen af de to datasæt.

Det er indlysende, at en så detaljeret gennemgang af det samlede ordforråd vil kræve en meget stor arbejdsindsats. Den her valgte metode er først og fremmest egnet til at danne sig et systematisk overblik.

3.3 Begrundelse for metodevalg

Af ovenstående fremgår det, at undersøgelsen primært er syntaksbaseret. Det er almindeligt kendt, at syntaksen er det felt, hvor man kan bedst skaffe sig et systematisk overblik ved overfladeanalyse.

Syntaksen alene er imidlertid på længere sigt utilstrækkelig som grundlag for en systematisering. En semantisk beskrivelse, der kan redegøre for affiniteten mellem ords betydninger, kræver en dyberegående analyse og et mere detaljeret beskrivelsesapparat. I første omgang nøjes vi med at liste de semantiske mønstre, vi umiddelbart kan iagttage ved gennemgang af materialet, til senere brug ved opstilling af et fuldt udbygget beskrivelses-system.

4. Sammenligning af korpusbelæg og repræsentation i DS

4.1 Den undersøgte datamængde

Den samlede datamængde fra DK87-90 for søgeordet *hånd* er ca. 600 kilobytes. Eksempelvis har vi analyseret 120 belæg med ordformen *hånden* og 220 med *hænder*.

DS registrerer ca. 115 ordforbindelser for dette opslagsord, med markering af deres leksikalske og morfosyntaktiske variationsmuligheder. Dette svarer til ca. 160 ordforbindelser i alt, når man har foldet artiklen helt ud, dvs. når alle varianter og alternativer er medregnede.

Efter at vi havde sorteret og typologiseret korpuseksemplerne, samt registreret eventuelle morfosyntaktiske varianter af samme flerordsenhed, foretog vi en sammenligning af de to datasæt. Det, vi i første omgang har interesseret os for, er at få overblik over, hvor stor en del af de autentiske eksempler på nutidig sprogbrug der er medtaget i DS, og hvordan registreringen er udformet.

4.2 Hvad kan vi konkludere af sammenligningen?

4.2.1 Dækningsgraden

Med hensyn til dækningsgraden mellem DS og de undersøgte korpuseksempler har vi kunnet konstatere, at der i DS mangler en del almensproglige flerordsenheder (både kollokationer og fraseoleksemmer), der forekommer flere gange i vort korpus-materiale, fx:

lægge *hånd* på ngt (= beslaglægge ngt.)
 give *én hånden* på at + inf. (= love én ngt.)
 tage sagen i egne *hænder* (= handle/bestemme selv)
 holde *hænderne* fra én/ngt. (fig., = holde sig fra én/ngt.)

Dette kan have flere grunde:

- * Leksikografens personlige skøn om, hvad brugeren hyppigst vil søge i ordbogen, dvs. om brugseksemplet var nødvendigt at medtage i ordbogen eller ej. Da ordbogens maksimale omfang på forhånd var fastlagt af forlaget, kunne det indsamlede materiale ikke medtages i sin helhed.
- * De pågældende udtryk er af leksikografen ikke betragtet som leksikaliserede eller på anden måde specielle ordkombinationer, men som frie ordforbindelser. Dette spørgsmål kan vi ikke komme videre med uden at kende leksikografens klassifikationskriterier.
- * Tilfældigt 'hul' i DS's belægsamling.
- * Udtrykket er en nydannelse, der først er opstået eller leksikaliseret, efter at materialeindsamlingen er afsluttet.

Det skal tilføjes, at der selvfølgelig også er en del af de flerordsenheder, der er registreret i DS, som ikke er belagt i vores korpus, men vi har ikke anset det for relevant at lave en egentlig statistisk sammenligning.

4.2.2 Registrering af leksikalske varianter i DS

I denne henseende er forholdet mellem de to datasæt nogenlunde det samme som under det foregående punkt.

Som nævnt i afsnit 3.2 inddeler vi flerordsenhederne i to hovedgrupper med hensyn til deres leksikalske variabilitet:

Type a	Lukkede enheder (Det variable element har kun ganske få alternativer)
Type b	Åbne enheder (Det variable element har en række alternativer, eksempelvis over 3 eller 4. Tallet er i princippet ikke afgørende; man kan justere det efter bearbejdelsen af en større datamængde, hvis man konstaterer, at et andet tal er mere hensigtsmæssigt.)

DS oplyser ikke noget om, hvor mange og hvilke udfyldninger der er mulige for obligatoriske led, som er angivet i rund parentes, fx

DS: (*gå*) fra hånd til hånd.

DS: (*børnene/sagen*) er i gode hænder

I det første eksempel viser korpusbelæggene, at listen bør udvides til: '*gå/vandre/blive rakt*', men efter vores skøn er der næppe flere udfyldningsmuligheder. Udtrykket hører derfor til de lukkede kollokationer (type a) og burde have været anført som alternerende komponenter (uden parentes).

I det andet eksempel kan eksempelvis også (... *dokumenterne / ejendommen / bedstemor*...) være alternativer. Udtrykket er altså en åben kollokation (type b).

DS dækker ikke - og kan af pladshensyn heller ikke dække alle individuelle kombinatoriske egenskaber for hvert opslagsord. Men vi har konstateret en vis mangel på konsekvens i den anvendte præsentationsmetode. Det kan illustreres med et enkelt eksempel fra DS, der indeholder begge ovenfor nævnte typer af leksikalsk variabilitet.

DS: *have/holde* (en *stok/en økse*) i hånden

Det verbale led er anført med to varianter som alternerende komponenter (type a), hvorimod det obligatoriske objekt markeres og præsenteres som en liste med to eksempler (type b).

For den første hovedgruppe (type a) gælder det, at det er muligt og ønskeligt at præsentere alle leksikalske alternativer for et udskifteligt element. Udover '*have/holde*' har vi et ret stort antal korpuseksempler med '*tage*' som verballed, men der optræder ikke andre verber på denne plads. Man kan derfor sige, at det verbale leds alternativer danner en endelig (og meget begrænset) liste, men at et af alternativerne mangler i DS.

For den anden hovedgruppe (type b) gælder det, at det udskiftelige element har et utal af alternativer. I den ovenfor viste flerordsenhed danner objektets mulige udfyldninger en åben liste, der kan omfatte meget forskellige substantiver, hvilket også er tilfældet i vore korpusbelæg. Alternativerne har dog visse fælles overordnede semantiske træk, fx '*genstand af en for et menneske håndterlig størrelse*'.

Konklusionen må være, at skellet mellem disse to listetyper bør opretholdes i præsentationen i ordbogen. Selv om ordbogens markeringssystem tillader en vis skelnen og systematik, bliver principperne ikke altid overholdt og kan dermed ikke tages for fulde pålydende.

Begge slags lister bør så vidt muligt baseres på et bredt sammensat stort korpus, dels for at undgå huller i materialet, dels for at udelukke, at okkasionelle ordforbindelser bliver håndteret som standardsproglige.

Den lukkede liste affattes eksplicit og udtømmende på grundlag af belæggene. Den åbne liste kan ikke være udtømmende på samme måde, selv ikke med basis i meget store datamængder. En sådan liste baseres derfor på en beskrivelse af alternativernes semantiske egenskaber, dvs. listens elementer bliver ikke leksikalske individer, men semantiske typer. Listen konstrueres ved at uddrage alternativerne fra korpusbelæggene for det variable led og undersøge, hvilke fællestræk der er karakteristiske.

4.2.3 Morfologisk og syntaktisk variation

DS markerer generelt de vigtigste variationsmuligheder, fx om substantivet kan anvendes i både ental/flertal, ubestemt/bestemt form, eller om der kan optræde alternerende præpositioner på samme plads.

Ved sammenligningen mellem de to datasæt viste det sig, at registreringen heller ikke på dette punkt er helt systematisk og entydig og heller ikke udtømmende. Det gælder fx numerusbrugen:

DS: stikke *hånden/hænderne* i lommen vs.
slå ud med *hånden*.

I vore belæg er formen '*hænderne*' hyppigere end entalsformen i den sidstnævnte konstruktion, følgelig burde denne alternering også være anført.

Det er også noget uigennemskueligt, hvilke kriterier der er afgørende for, om DS bringer det verbale led i en flerordsenhed i grundform (infinitiv) eller i bøjet form (præsens eller præteritum), altså hvad der fx er grunden til den forskellige behandling af:

DS: (børnene *går*) hånd i hånd vs.
(*gå*) fra hånd til hånd, eller
(*stå*) med hænderne i siden.

I nogle tilfælde er valget af en speciel finit form velbegrundet. Det er således naturligt at bruge perfektum i det følgende eksempel, selv om andre former er mulige:

DS: (brevet) *er kommet* mig i hænde.

Problemet er bl.a., at der ikke er foretaget en systematisk skelnen mellem, om de morfologiske og/eller syntaktiske restriktioner for en given flerordsenhed gør en "normalisering" til en (sub)lemmaform urimelig, eller om flerordsenheden anføres som et aktualiseret brugs-eksempel. Denne skavank, som DS har til fælles med mange andre ordbøger¹³, er fatal for NLP og er med til at gøre livet besværligt for den, der oversætter til et fremmedsprog.

Konklusionen for dette punkt er, at registreringen af morfologisk og syntaktisk variation bør være ensartet og udtømmende. Dermed menes, at variation altid skal markeres, hvis den er restringeret eller opfattes som afvigende i forhold til de generelt gældende grammatiske regler. Hvis der ikke er restriktioner for et ledes variabilitet i forhold til frie ordforbindelser af tilsvarende struktur, skal leddet registreres i grundform (verber i infinitiv, substantiver i ental/nominativ osv).

5. Projektets næste fase

Det er planen nu at foretage en sammenligning af et større datamateriale fra de to datasæt, for at sikre os et solidt arbejdsgrundlag for den afsluttende fase, hvor arbejdsindsatsen vil blive koncentreret om at udarbejde selve beskrivelsessystemet.

Projektet vil også kunne give en vurdering af, hvorvidt materialet i DS vil kunne bearbejdes og udbygges til at kunne opfylde kravet om en moderne "stil- og konstruktionsordbog".

Perspektiverne i korpusbaseret ordbogsarbejde er lovende også til dette formål: de enkelte ords muligheder for at indgå i større meningsenheder eller brugsmæssige enheder kan bedst kortlægges ved undersøgelse af en omfattende mængde autentiske brugseksempler.

6. Afsluttende bemærkninger

Med visse forbehold kan det siges, at DS er en god specialordbog: den indeholder det mest omfattende, for mennesker systematiserede materiale om moderne dansk sprogbrug. Når vi i vor undersøgelse er stødt på en del tilfælde af inkonsekvent registrering og præsentation, skyldes det givetvis flere faktorer, som for eksempel:

- * Ordbogen er lavet på grundlag af en personlig seddelsamling; der kan have været tilfældige huller i dette materiale;
- * Af pladshensyn var leksikografen nødt til at frasortere en hel del materiale, dette bevirker også, at der mangler visse almindelige flerordsenheder;
- * Den dengang anvendte, traditionelle leksikografi har væsentlige kvantitative begrænsninger i forhold til de datamatiske metoder, leksikograferne har til deres rådighed i dag, hvilket også har påvirket det kvalitative aspekt.

Vi har i vor undersøgelse haft adgang til relativt store datamængder og avancerede datamatiske hjælpemidler. Desuden stiller vi i dag - sammen med mange andre - anderledes krav til ordbøger, end det var muligt for 30 - 40 år siden.

Sidst men ikke mindst er vore synspunkter præget af, at vi til daglig beskæftiger os med datamatisk behandling af naturlige sprog (NLP). Dette arbejde gør os bevidst om, at der er stort behov for, at sproget beskrives systematisk, eksplicit og entydigt, men også om, at dette ideelle mål bliver meget svært at nå.

Konsistens og konsekvens er nøgleordene, der også bliver fremtidens krav i almen leksikografi, blandt andet fordi materialet i moderne ordbøger skal kunne udnyttes til mange forskellige formål.

Noter

1. Der skal ikke her gås i detaljer med, hvilke resultater opslagene gav. Der er tale om en række af de svagheder, som er påvist i Braasch (1987) og (1988) og Kjær (1986) og (1987): bl.a., at bøgerne giver forskellige, til dels modstridende oplysninger, og at det ofte er uklart, hvor meget der hører med til det faste udtryk, og hvor meget der er en frit udskiftelig del af et brugseksempel.

2. Som typisk eksempel kan citeres fra H. Mattutat, Deutsche Grundsprache (1969) under opslagsordet *der Schuh*: [...] Wdg: jdm die Schuld in die Schuhe schieben, fig (geben) [...]

3. Det ville nærmest betyde "at beskyldte nogen for skylden". (Ud over et "historisk" eksempel i Ordbog over det danske Sprog har vi ganske vist et belæg med "skylden" fra Berl. Tid. fra 20.3.91, men det er karakteristisk nok i en artikel fra bladets Bonn-korrespondent, som efter mange år i Tyskland jævnlig bruger germanismer).

4. DS beskriver altså den gængse brug, men er for kategorisk med formuleringen (under *sko sb*): c. tm. [...]; skyde én i -ene at + sætn. (: beskyldte én for ngt forkert, navnlig meninger, lave motiver)
5. Fx giver Hausmann i sin artikel om "Collocations" i WDD (Hausmann 1989,1010) blot en ekstensionel definition af begrebet uden at begrunde sin afgrænsning.
6. Se fx Skyum-Nielsen (1991).
7. Se Coseriu (1971) og v. Polenz (1973).
8. Et eksempel på en sådan kreativ leg med ordene i vores førnævnte faste udtryk er overskriften "Was sich Große in die Schuhe schieben", refererende til gaver på St. Niklausdag (Citeret fra Wotjak 1989,461). Det ville næppe være muligt - og i hvert fald ikke rimeligt - at udforme en leksikografisk beskrivelse sådan, at den også tager højde for den slags variation.
9. Se Poulsen (ej publ.) og Skyum-Nielsen (1991). Se også Sonja Poulsens artikel i dette bind.
10. Kildesprogets grænseflade er slutresultatet af kildesprogsanalysen og dermed input til den egentlige oversættelse (transfer). Målsprogets grænseflade er outputtet fra transfer og dermed input til genereringen af overfladeteksten på målsproget.
11. Se Bergenholtz (1988) og DK87-90 i litteraturlisten.
12. Se note 11 ovenfor.
13. Se her igen de førnævnte arbejder af Braasch og Kjær.

Litteratur

- BBI = Benson, M. / E. Benson / R. Ilson. 1986. *BBI Combinatory Dictionary of English. A Guide to Word Combinations*. Amsterdam - Philadelphia
- WDD = Hausmann, F.J. / O. Reichmann / H.E. Wiegand / L. Zgusta (Hrsg.). 1989ff. *Wörterbücher - Dictionaries - Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Bd. 1-3. de Gruyter, Berlin - New York
- Bergenholtz, Henning. 1988. DK87: Et korpus med dansk almensprog. I: Henning Bergenholtz / Finn Frandsen (red.) Hermes, Tidsskrift for Sprogforskning nr. 1 - 1988 s. 229-237. Det erhvervsproglige Fakultet, Handelshøjskolen i Århus
- Bergenholtz, Henning. 1989. *Probleme der Selektion im allgemeinen einsprachigen Wörterbuch*. Kap. 68 i WDD Bd. 1 (S. 772-779).
- Bergenholtz, Henning / Joachim Mugdan (Hrsg.). 1985. *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.-30.6.1984*. (Lexicographica. Series Maior Bd. 3). Tübingen
- Bork, Egon. 1990a. *Dansk-tysk Ordbog*. (Gyldendals Røde Ordbøger) 9. udg., 4. opl. Gyldendal, København
- Bork, Egon. 1990b. *Tysk-dansk Ordbog*. (Gyldendals Røde Ordbøger) 12. udg., 4. opl. Gyldendal, København
- Braasch, Anna. 1987. *Zur lexikographischen Kodifizierung von Phrasemen in einsprachigen deutschen Wörterbüchern*. (Unpubl. specialeafhandl., Kbh. Univ.). København
- Braasch, Anna. 1988. *Zur lexikographischen Kodifizierung von Phrasemen in einsprachigen deutschen Wörterbüchern - aus der Sicht eines ausländischen Wörterbuchbenutzers*.

- I: Karl Hyldgaard-Jensen & Arne Zettersten (ed.): *Symposium on Lexicography IV (Lexicographica, Series Maior Bd. 26)*. Niemeyer, Tübingen
- Coseriu, Eugenio. 1971. System, Norm und 'Rede'. I: E. Coseriu: *Sprache, Strukturen und Funktionen*. XII Aufsätze zur allgemeinen und romanischen Sprachwissenschaft. Hrsg. v. Uwe Petersen ... (= Tübinger Beiträge zur Linguistik 2) 2. verbesserte Auflage. Tübingen
- DK87-90 = maskinlæsbart dansk korpus, udg. af H. Bergenholtz, Århus. (Kan rekvireres ved henvendelse til H. Bergenholtz). Se Bergenholtz (1988)
- DS = Bruun, Erik. 1978. *Dansk Sprogbrug. En stil- og konstruktionsordbog*. (Gyldendals Røde Ordbøger) 1. udg. Gyldendal, København
- Fleischer, Wolfgang. 1982. *Phraseologie der deutschen Gegenwartssprache*. 1. Aufl. Bibliographisches Institut, Leipzig
- Hausmann, Franz Josef. 1989. Le dictionnaire de collocations. I: *WDD* Bd. 1 S. 1010-1019.
- Heid, Ulrich / Sybille Raab. 1988. Kollokationen im Wörterbuch für multilinguale Generierung und maschinelle Übersetzung. I: *LDV-Forum* Bd. 5 Nr. 4 -Sept. 1988 S. 72-73 (Sammendrag af artikel tilsendt medl. af Arbeitskreis Lexikographie der GLDV)
- Kjær, Anne Lise. 1986. *Fraseoleksemmer i tyske étsprogordsbøger*. ARK 29 (Sproginstitutternes Arbejdsrapporter, Handelshøjskolen i København - bearbejdet cand.ling.merc.-specialeafhandling). København
- Kjær, Anne Lise. 1987. Zur Darbietung von Phraseologismen in einsprachigen Wörterbüchern des Deutschen aus der Sicht ausländischer Textproduzenten. I: Jarmo Korhonen (Hrsg.): *Beiträge zur allgemeinen und germanistischen Phraseologieforschung* S. 165-181. Oulu
- Kjær, Anne Lise. 1990a. *Normbetingede ordforbindelser i tysk juridisk sprog*. ARK 56 (Sproginstitutternes Arbejdsrapporter, Handelshøjskolen i København - Ph.D.-afhandling). København
- Kjær, Anne Lise. 1990b. Phraseology Research - State-of-the-Art. Methods of describing word combinations in language for specific purposes. I: *Journal of the International Institute for Terminology Research IITF*, Vol. 1 (1990), no. 1-2. Wien
- Kromann, Hans-Peder. 1989. Zur funktionalen Beschreibung von Kollokationen und Phraseologismen in Übersetzungswörterbüchern. I: Gertrud Gréciano (Ed.): *Europhras* 88. *Phraséologie Contrastive*. s. 265-272. Strasbourg
- Pfeffer, J. Alan. 1979. Drei deutsche Korpora im Institut für Grunddeutsch. Ziel und Stand ihrer Auswertung. I: Bergenholtz H. / B. Schaefer (Hrsg.): *Empirische Textwissenschaft* S. 307-323. Scriptor, Königstein/Ts
- Pilz, K.D. 1978. *Phraseologie. Versuch einer interdisziplinären Abgrenzung, Begriffsbestimmung und Systematisierung unter besonderer Berücksichtigung der deutschen Gegenwartssprache*. Göppingen
- v. Polenz. 1973. Sprachkritik und Sprachnormenkritik. I: G. Nickel (Hrsg.): *Angewandte Sprachwissenschaft und Deutschunterricht*. München
- Poulsen, Sonja. 1991. *Word Combinations as Lexical Building Blocks*. (Prisopgave indleveret til Odense Universitet 1991 - endnu ikke publiceret).
- Skyum-Nielsen, Peder. 1991. *Fyndord. Studier i kortformernes retorik*. Bd.1-3.(Udkommer 1991 ell. evt. 1992 hos Hans Reitzel). København
- Svensén, Bo. 1987. *Handbok i lexikografi. Principer och metoder i ordboksarbetet*. (TNC 85) (Især s. 92-105: Kollokationer och idiom). Stockholm

Wotjak, Barbara. 1989. Ansatz eines modular-intergrativen [sic!] Beschreibungsmodells für verbale Phraseolexeme (PL). I: Gertrud Gréciano (Ed.); *Europhras 88. Phraséologie Contrastive* S. 459-467. Strasbourg