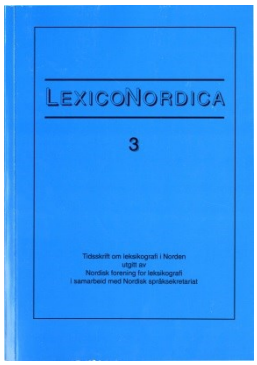
LexicoNordica

Titel:	Korpusbaseret leksikografi
Forfatter:	Henning Bergenholtz
Kilde:	LexicoNordica 3, 1996, s. 5-18
URL:	http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



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 LexicoNordica (1-16) 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.

Henning Bergenholtz

Korpusbaseret leksikografi

The first computerised corpora (Brown, LOB and Limas) can be described as polyfunctional corpora of partial texts. They consisted of a certain number of small, brutally cut parts of texts, all of the same length, and they were intended as a linguistic basis for all linguistic research fields one could think of. But they could not fulfil this aim. Not only lexicography, but especially this science requires corpora with whole texts that have been collected for a special purpose. Whole texts here understood as texts taken in their entirety or at least parts of texts that may be seen as a whole, for instance a complete chapter.

1. Tilbageblik

I 1977 kunne man med god ret skrive om en ny udvikling inden for leksikografien: "Lexikographen legen neuerdings verstärkten wert auf satzübergreifende textkorpora" (Wiegand 1977:129). Udtrykket "tekstkorpora" var så nyt, at Wiegand følte, at det måtte forklares, idet han tilføjede attributtet "satzübergreifend" (går ud over sætningsgrænsen). Hvis Wiegand i stedet for **tekstkorpus** havde brugt udtrykket **korpus**, havde tilføjelsen været mere nødvendig, idet korpus traditionelt ikke bare er blevet brugt om en samling af tekster, men også om en samling af belæg, desuden om en samling af selvdannede eksempler, ja endda om en samling af kollokationer eller ord. Der er i virkeligheden tale om mindst tre forskellige måder at brugen termen korpus på.

For det første har vi dem, der ved et korpus forstår en samling af sproglige enheder under sætningsniveau, dvs. af ord og kollokationer. Denne gruppe er ikke relevant for korpusdiskussionen i denne sammenhæng. For det andet kan man udskille de leksikografer og lingvister, som enten antager samme status for selvlavede eksempler og tekstbelæg (Greenbaum 1984:194) eller advarer mod brugen af et tekstkorpus eller af belæg (Zöfgen 1986 og Pasch 1992:284). Diskussionen med sådanne belæg- og korpusskeptikere er blevet ført i anden sammenhæng (se Bergenholtz/Mugdan 1989 og 1990) og vil ikke blive gentaget her. Endelig er der de leksikografer, der excerperer belæg fra tekster til en seddelsamling eller anvender et tekstkorpus. Det er de to sidste gruppe, som står i centrum i dette bidrag.

I denne sammenhæng er Wiegands udtryk "neuerdings" (i den sidste tid) og "verstärkten wert" (større vægt) særligt interessante. Det var selvfølgelig ikke noget helt nyt, at leksikografer helt eller delvis brugte

foreliggende tekster som empirisk basis. I lingvistikkens historie har det i større eller mindre grad været usus for bl.a. retorikker og grammatikker; og belægsamlinger har i flere hundrede år været det væsentlige empiriske grundlag for store monolingvale betydningsordbøger. Men det er først i dele af den unggrammatiske skole, at denne fremgangsmåde bliver fremlagt som et metodisk ideal for specielt den grammatiske og i videre forstand al lingvistisk forskning. Behaghel beskriver i forordet til sin firebinds tyske syntaks netop modsætningen til de grammatikere, som har samlet så mange belæg som muligt og så vidt muligt tager højde for dem alle. På denne måde, siger Behaghel, inddrager man i virkeligheden først og fremmest det påfaldende, det sjældne og undtagelserne. Alt det regelmæssige i sproget bliver der derimod taget relativt mindre hensyn til. Og én ting kan man overhovedet ikke sige noget om, nemlig at formodede forekomster slet ikke kan findes. Eksempler på det, Behaghel taler om, findes der mange af. Et af de særlig tydelige er problemet kasuskongruens i tyske nominalfraser. Her har en lang række grammatikere af sådanne eksempelsamlinger ladet sig forlede til at udtale sig både om middelhøjtysk, tidligt nyhøjtysk og den nuværende sprogbrug, at det må anses for at være meget udbredt med en manglende kasuskongruens, sml. ***mit*** *Herrn Müller,* ***dem*** *berühmten Maler* og ***mit*** *Herrn Müller,* ***des*** *berühmten Malers*. Den sidste eksempeltype uden kasuskongruens kan findes, men ret sjældent, ret nøjagtig i ca. 3,6% af tilfældene (Bergenholtz 1985).

2. Empirisk basis i leksikografien

I den leksikografiske verden kender vi tilsvarende resultater af belægsamlinger. Hvert belæg er ganske vist et virkeligt eksempel på konkret sprogbrug. Men de, der har søgt og skrevet belæggene af, har i høj grad ladet sig fange af det særlige, det overraskende. Om fx seddelsamlingerne for Duden-ordbøgerne eller ved det Danske Sprognævn er der derfor også med et gran af sandhed blevet sagt, at de er præget af sproglige perversiteter. I stedet for en ukontrolleret belægindsamling foreslår Behaghel følgende metode for den grammatiske forskning:

> Bei diesen Untersuchungen habe ich das Verfahren beobachtet, das sich bei meiner Arbeit über die Zeitfolge bewährt hat, das Verfahren der Stichproben, das gewisse Stücke gewisser Denkmäler vollständig auszubeuten sucht. Wer danach andere Stücke und andere Quellen durchmustert, wird vielleicht wertvolle Ergänzungen bieten können, aber das von mir Gefundene kaum gänzlich umwerfen. (Behaghel 1923:VIII).

Der er to afgørende udtryk i dette citat, som er relevante for den leksikografiske diskussion. Det er for det første kravet om en fuldstændig analyse af hele stikprøven. Hermed adskiller fremgangsmåden sig positivt fra metoden med den myreflittige samlen, som resulterer i et tilfældigt og ikke-kontrollerbart udvalg af alle potentielle belæg, og hvor man især ikke kan forestille sig, at to forskellige excerptorer skulle kunne komme frem til en bare nogenlunde ensartet selektion. Behaghels krav er dog metodisk ikke uden problemer for leksikografisk arbejde. Det kan opfyldes i mange grammatiske undersøgelser, fx ved eksemplet appositioners kasuskongruens. Det tog omkring 400 timer at gennemlæse og afmærke alle appositioner i et korpus med to mio. tekstord, men det drejede sig trods alt om ikke mere end godt 6.000 belæg. I leksikografien vil et sådant korpus ved mange problemstillinger være alt for lille. Men alligevel vil man selv ved et så beskedent korpus have så mange belæg til særlig hyppige funktionsord, at arbejdets omfang med at tage nøjagtig højde for alle forekomster vil udgøre en unødvendig stor arbejdsbelastning. Med andre ord vil vi have det problem, at vi til forskellige leksikografiske problemstillinger vil være tvunget til at gøre brug af en større eller mindre del af det sammenstillede korpus eller evt. have forskellige korpora til forskellige problemstillinger. I virkeligheden svarer dette præcis til Behaghels fremgangsmåde. Han har i praksis alt efter problemstilling haft forskellige "Denkmäler", men det anser han for noget så selvfølgeligt, at der ikke bliver redegjort nøjagtigt herfor.

Det andet afgørende, men samtidig noget uklare udtryk i ovennævnte citat er "visse dele af visse tekster": Hvor mange tekster drejer det sig om? Hvordan er dette tekstudvalg truffet? Hvor store dele af teksterne er medtaget? Og er det hele tiden den samme del af teksterne, fx det første afsnit eller kapitel? Også herom siger Behaghel intet.

3. Teoretiske og metodiske overvejelser

De nævnte problemer udgør nogle af de centrale problemer, som korpussammensætningen og korpusanalysen også i dag står over for. En af de første bøger, som udelukkende tematiserer et tekstkorpus' status, dets sammensætning og analysemetoder, er "Empirische Textwissenschaft, Aufbau und Auswertung von Text-Corpora". Denne bog har ikke fundet videre genklang, sandsynligvis fordi bogen har en tysk titel og de fleste af bidragene er skrevet på tysk. Som særligt centrale anser jeg bidragene af van de Velde (1979), Bausch (1979), Bungarten (1979) og Rieger (1979). Men disse og de fleste andre artikler i denne bog har i

den videnskabelige korpusdiskussion, som i særlig grad har været præget af anglister, ikke spillet den centrale rolle, som de havde fortjent og som sandsynligvis kunne have haft påvirket et til dels problematisk diskussionsforløb i positiv retning. Det drejer sig dels om den statistisk uholdbare tale om 'repræsentative korpora', dels om et korpus' videnskabsteoretiske status i lingvistikken og i andre fag.

Særligt misvisende er den udbredte betegnelse "korpuslingvistik", som efterhånden har bredt sig fra anglisternes rækker til resten af lingvistikken og i øvrigt har givet navn til et nyt tidsskrift. Hermed lader man en særlig form for empirisk basis udgøre en særlig lingvistisk disciplin. Med samme tankegang kunne man også opfinde en intuitionslingvistik, spørgeskemalingvistik eller belægsamlingslingvistik. Tilsvarende kunne man have en korpusleksikografi, en intuitionsleksikografi eller måske i særlig grad en skrive-af-leksikografi. Betegnelsen **korpuslingvistik** er i øvrigt dobbelt misvisende. Ikke alene hylder man en betænkelig habeas-corpus-holdning, som fører til, at det at sammenstille korpus skulle være et særligt sprogvidenskabeligt mål i sig selv. Men man ser også bort fra, at tekstkorpora desuden sammensættes og bruges af andre end lingvister, sml. fx Mergenthaler (1979).

På sin vis er denne habeas-corpus-holdning ikke spor bedre end den tilsvarende radikale afvisning af brugen af korpora, sml. fx: "That is a complete waste of your time and the government's money. You are a native speaker of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text.". (Robert Lees 1962, mundtlig diskussion, citeret i Francis 1979:110). Francis kan i sin modreplik have ret i, at Chomsky hverken dengang eller senere har forelagt en holdbar definition af competence eller performance, og at han ikke kan forstå, hvordan competence kan undersøges uden hensyntagen til virkelig performance, dvs. tekster. Men Francis' glæde over korpusskeptikernes nederlag var mere en foregriben af en senere udvikling end en beskrivelse af en afsluttet debat, idet diskussionen om værdien af tekstkorpusbaserede undersøgelser vedblev at være noget følelsesladet, sml. fx beskrivelsen af korpusanalyser som "eine überflüssige Zeremonie" (Itkonen 1976:65).

4. Sammensætning af korpus

Rent historisk kan man fastslå, at de første korpora, begyndende med det amerikanske Brown-Corpus, over det engelske LOB-Corpus og det tyske Limas-Korpus er totalsproglige korpora, dvs. korpora bestående tekster fra både fagsprog og almensprog. Hvert af disse korpora tager

hensyn til en sagklassifikation, som fx ved Limas-Korpus går ud fra opdelingen i Frankfurter Nationalbibliographie og fra en vægtning efter mængden af udkomne tekster i de der brugte 33 "sagområder". Hver deltekst har en størrelse på 2000 tekstord. Korpus, som er på en mio. tekstord, omfatter således i alt 500 enkelttekster, fx 17 medicinske, 6 matematiske og 64 skønlitterære tekster. En sådan spredning gør det fx muligt at foretage frekvenssammenligninger mellem de hyppigst forekommende ord og fraser inden for de forskellige sagområder. Således kan Johansson (1979) i en undersøgelse af Brown-Corpus påvise tre gange så mange definite nominalfraser med efterfølgende *of*-konstruktioner i naturvidenskabelige og tekniske tekster som i fiktionale tekster.

Med Brown-Corpus og de andre små korpora på en mio. tekstord foreligger der helt sikkert det, som Wiegand kaldte "korpora, som går ud over sætningsgrænsen". Nøjagtigere leksikalske undersøgelser af sprogbrugen i enkelte fag- eller sagområder er selvfølgelig slet ikke mulige i et korpus med en så svag dækning af de enkelte områder. Hverken omfanget eller sammensætningen gør det muligt at foretage bare nogenlunde sikre undersøgelser med henblik på en total- eller en almensproglig ordbog. Den undersøgelse, Wiegand refererede til, antog et omfang på omkring 50 mio. tekstord som et tilstrækkeligt stort korpus til en planlagt stor, ny totalsproglig ordbog, sml. hertil også Mentrup (1979:185).

En yderligere indskrænkning af mulige tekstlingvistiske og også leksikalske undersøgelser ligger i de tekstuelt brutale snit i udgangsteksterne, når hver korpustekst skæres ud i en ganske bestemt fast størrelse. Herved bliver tekstuelle sammenhænge i begyndelsen eller slutningen af en korpustekst af og til så uklare, at en sikker interpretation og dermed leksikografisk brug af disse tekstdele bliver umulig.

De hidtil nævnte korpora kan henregnes til typen **deltekstkorpus**. Disse korpora danner en færdig enhed, som ikke mere kan og skal ændres. En helt anden type er **heltekstkorpus,** som ikke kan skelnes skarpt fra et tekstbibliotek hhv. **tekstbank**. En særlig type heltekstkorpus er det såkaldte **monitor corpus**, som er blevet oprettet og brugt i forbindelse med Cobuild-projektet, sml. Renouf (1991). Et monitorkorpus optager så mange tekster som muligt, de modtages normalt direkte fra forlag o.l. og indgår i deres helhed i korpus, som kan "rulle" hen over skærmen. Alt efter behov bruges større eller mindre dele af det til enhver tid foreliggende tekstmateriale, som hele tiden bliver udvidet, ligesom tekster, som ikke længere er aktuelle, slettes eller overføres til et reservekorpus. På mange måde minder argumentationen om den, vi indledningsvis fandt hos Behaghel.

De førstnævnte korpora (Brown, Limas og LOB) blev af deres oprettere forstået og beskrevet som **polyfunktionale** korpora til brug i alle eller en lang række lingvistiske forskningsområder. Sådanne potentielt alsidige brugsmuligheder kan man utvivlsomt også forudsætte, når der ikke angives specifikke enkeltproblemer, fx hos Glas (1975). I praksis er den forudsete polyfunktionalitet hverken realistisk eller hensigtsmæssig. Til et bestemt problem må man sørge for fremskaffelsen af præcis den form for empirisk basis, som må anses for hensigtsmæssig. Et sådant **monofunktionalt** korpus foreligger fx med Maegaard-Ruus-korporaene, som ganske vist ikke uden store indskrænkninger kan bruges til andet end sprogstatistiske undersøgelser, men som på den anden side er sammensat under særlig hensyntagen til i det mindste ét formål og også brugbare hertil. Derimod er en lang række polyfunktionale deltekstkorpora kun virkeligt brugbare til grammatiske undersøgelser af meget almen karakter, men ikke til opnåelse af nøjagtigere indsigt i grammatiske problemområder og slet ikke til leksikografiske undersøgelser.

5. Korpora som empirisk basis for almensproglige ordbøger

Mange, måske de fleste nyere europæiske og amerikanske ordbøger, beskriver sig selv direkte som totalsproglige ordbøger, dvs. ordbøger, der både tager hensyn til almensproglig og i større grad også til fagsproglig brug. Nogle går endda så vidt, at de som et af deres ordbogs formål ser en hjælp til forståelse mellem fagområderne, dvs. fagfolk imellem og mellem fagfolk og lægfolk, sml. "Durch die starke Berücksichtigung der Fachsprachen wird es auch eine sichere Basis für die Verständigung zwischen den Fachbereichen schaffen." (DUDEN-GWB, Vorwort, side 2). Man kan ganske vist godt forestille sig en stor interdisciplinær ordbog, som både tager virkelig hensyn til fagsprogene og til almensproget, men i det tilfælde vil man ikke kunne nøjes med inddragelse af leksikografer med en lingvistisk baggrund. For alle de angiveligt totalsproglige ordbøger, som er udkommet indtil nu, gælder det, som Paolo Beni i en anmeldelse af Accademia della Crusca's ordbog skrev i 1612 (se herom Hausmann 1989). Man bør i en optimal alment brugbar ordbog, mener Beni, foretage et meget bredt udvalg af stikord, som bruges i almensproget. Derimod frarådes optagelse af stikord fra de faglige ordforråd. Denne tankegang kan føres tilbage til den empiriske basis, som optimalt bør lægges til grund for en almen ordbog. Sådanne ordbøger udarbejdes af lingvistiske leksikografer, som ganske vist kan rådføre sig med fageksperter, men dog ikke har forudsætning for at udvælge de relevante fagtekster og endnu mindre

for at analysere alle mulige fagtekster. Det betyder ikke, at fagtermer helt skal udelukkes fra lemmalisten til en almensproglig ordbog. En lang række fagtermer indgår i det almene sprog. Og sådanne termer bør principielt også optages i lemmalisten. Det betyder heller ikke, at den leksikografiske definition ikke skal være fagligt korrekt, som fx COBUILD anser som en nødvendighed ud fra deres deskriptive argumentation, sml. "Hence we have explained the technical words according to the way we use them in ordinary English." (COBUILD, Introduction, p. XX). Men det betyder, at man bør indskrænke sig til et udvalg af sådanne tekster, som formodes hørt eller læst af en stor del af et bestemt sprogs modersmålsbrugere.

I forbindelse med konceptionen af det danske polyfunktionale deltekstkorpus DK87–90 (sml. Bergenholtz 1988) blev der således opstillet følgende udgangspunkt, som i nogen grad kan overføres til leksikografiske sammenhæng:

1. der skal ikke medtages fagtekster
2. korpus skal kun indeholde førstegangsudgivelser
3. der skal ikke medtages oversættelser
4. tekstudvalget skal være på forkant med sprogudviklingen
5. tekstrecipienterne skal udgøre en væsentlig del af befolkningen

Disse principper medfører (1), at der ikke medtages tekster af typen fagmand –> fagmand eller fagmand –> semifagmand. Der medtages almensproglige tekster, som muligvis er skrevet af en fagmand eller semifagmand, men hvor den forudsete tekstrecipient er en lægmand. Endvidere gælder (2), at kun udgivelser fra det pågældende år medtages, dvs. fx ikke H.C.Andersen-tekster. Sådanne tekster læses ganske vist i betydelig grad, men de genspejler ikke uden indskrænkninger den nutidige sprogbrug. Det samme gælder (3) for oversættelser, som ofte bærer tydeligt præg af deres oprindelige sprog. Mængden af aviser, dagblade, reklamer osv. er ganske vist betydeligt større end mængden af fiktionale tekster, sidstnævnte anses dog (4) for at være på forkant med udviklingen. Endelig (5) udelukkes tekster, som kun henvender sig til børn eller unge. Disse forudsætninger førte til følgende repræsentation af tekstarter:

1. romaner og noveller (50% af alle tekster)
2. aviser (25% af alle tekster)
3. ugeblade (25% af alle tekster)

Hver af disse tre tekstarter læses af mere end 50% af befolkningen, de er i modsætning til ikke-medtagne børnebøger og fagtekster ikke skrevet for en bestemt, forholdsvis begrænset del af den danske befolkning. Der er således gode grunde til at anse disse tre tekstarter for en væsentlig del af det almindelige danske skriftsprog. Derimod er procentopdelingen ikke uproblematisk; en opdeling mellem de tre tekstarter i forholdet 1:1:1 havde også været mulig. Jeg valgte at give romaner og noveller en særlig stor vægt, fordi de (på grund af dialogerne) i højere grad end aviser og ugeblade bærer præg af udviklingen i det talte sprog.

6. Korpora som empirisk basis til fagordbøger

De fleste foreliggende fagsproglige korpora er polyfunktionale heltekstkorpora. Et af dem, det såkaldte **Limas-Kfz-Korpus**, indeholdende tekster om automobiler hentet fra lærebøger, brugsanvisninger og bilfagblade, er aldrig blevet brugt til noget som helst (se Bergenholtz/Pedersen 1994:165). Andre korpora er efter en del udlugning af ikke-faglige, altså almensproglige tekster, blevet brugt under udarbejdelse af en bilingval ordbog (GENTEKNOLOGISK ORDBOG). Men mængden af udarbejdede fagordbøger, der bygger på et gennemanalyseret tekstkorpus som en del af ordbogens empiriske basis, er indtil nu ikke stor (ud over den ovenfor nævnte ordbog kendes kun PUMPETEKNOLOGISK ORDBOG). Både tekstselektion til og omfang af korpus udgør særlige problemer, som her ikke kan behandles, der henvises til den noget mere udførlige diskussion i Bergenholtz/Pedersen (1994).

Man kan skelne mellem et flerfagskorpus (fx Siliakus 1979), et enkeltfagskorpus (fx Bergenholtz/Kaufmann 1991) og et delfagskorpus (fx Dyrberg et al. 1988). Ligesom udarbejdelse af flerfagsfagordbøger vil sammensætning og analyse af flerfagskorpora være forbundet med en lang række problemer og i mange tilfælde ikke give tilfredsstillende resultater. For enkeltfags- og delfagskorpora må der gælde følgende: "korpus bør sammensættes i samråd med fageksperter" (Terminologiafdelingen 1987:9). Ikke bare er lægfolk ikke i stand til at overse et fags systematik og dække hele faget ind, de er heller ikke i stand til at skelne klart mellem tekster for forskellige målgrupper. Fx har Lauridsen/Riiber/Søndergård (1991) medtaget tekster fra fire grupper:

1. fra fageksperter til fageksperter
2. fra fageksperter til lægfolk
3. fra lægfolk til lægfolk
4. fra lægfolk til fageksperter

Til dette korpus har Kaufmann (1993) og Stummann (1993), begge fageksperter (molekylærbiologer) og samtidig deltagere i et fagleksikografisk projekt, haft en lang række indvendinger. Alle tekster tilordnet gruppe 1 må betegnes som fejlklassificerede, en dansk fageksperter skriver på engelsk, hvis målgruppen er fageksperter. En dansk molekylærbiolog skriver evt. på dansk for semifagfolk, dvs. for fagfolk fra nabodiscipliner. De fleste af disse tekster var tilordnet gruppe 1, men visse andre gruppe 2. Til gruppe 3 og 4 indvendte molekylærbiologerne, at det for dem var umuligt at se forskel på disse grupper af tekster. I øvrigt kunne de ikke betegne lægfolks udsagn om genteknologi som fagtekster, men som almensproglige tekster.

Fageksperter bør dog ikke kun medvirke ved sammensætning af korpus, deres medvirken er i lige så høj grad en nødvendighed under arbejdet med korpus, hvis leksikografen ikke selv har en vis grad af fagkompetence. Der må dog her skelnes mellem forskellige problemstillinger. Hvor en lingvist uden større problemer vil kunne foretage en fleksionsmorfologisk analyse af et fagsprogskorpus, vil et vist kendskab til faglige sammenhæng være nødvendigt ved selektion af kollokationer, og et temmeligt stort kendskab til faget under udvælgelse af evt. eksempler. Optimalt vil det her være, at der under hele arbejdet med sammensætning og analyse af korpus til stadighed foregår et tæt samarbejde mellem en leksikograf og en fagekspert, således at intet arbejdstrin udføres alene af en fagekspert, men heller intet alene af en leksikograf (men man kunne dog forestille sig, at leksikografen både er fagsprogsekspert og fagekspert).

7. Analyse af korpus

Under hensyntagen til 60ernes og 70ernes tekniske muligheder var et korpus på en mio. tekstord sikkert et "large corpus" (Francis 1979). Men det er nu næsten helt rørende, når en ordbog påberåber sig sin store nøjagtighed under henvisning til et sådant lille polyfunktionalt deltekstkorpus, fx i forordet til den første udgave af AMERICAN HERITAGE DICTIONARY. Med en sådan basis vil man selvfølgelig kunne få nogen hjælp ved en almensproglig ordbog, men dette korpus kan helt sikkert ikke være en tilstrækkelig empirisk basis i alle problematiske sammenhæng: betydningsangivelser, grammatiske angivelser, kollokationsangivelser, oplysninger om ortografiske varianter osv. osv. Men også ordbøger, som bygger på et større korpus, har tilsyneladende det problem, at korpus ganske vist foreligger, men at de pågældende leksikografer af forskellige grunde ikke har brugt det. Det

er blevet sagt eksplicit i forbindelse med TRÉSOR DE LA LANGUE FRANÇAISE, at det i praksis i de fleste tilfælde kun har været muligt at tage højde for mindre end en procent af tekstmaterialet. Hvor stor en del af korpus, der er blevet brugt af korpusbaserede ordbøger, kan vi ikke vide. Principielt kan man mest naturligt læse forteksten til ordbøger, der henviser til det brugte korpus, sådan, at korpus er blevet inddraget under hele det leksikografiske arbejde. Den leksikografiske virkelighed ser dog ud til at være en anden – i hvert fald, hvis Rösel's (1995) analyser gælder for mere end de undersøgte kollokationer: De korpusbaserede ordbøger indeholder kun få og i forhold til Rösels tekstmateriale tilsyneladende tilfældigt udvalgte og slet ikke alle vigtige kollokationer. Den ordbog, som har de mest relevante og også de fleste kollokationer er den ikke-korpusbaserede BBI-DICTIONARY.

Nu er det selvfølgelig ikke sådan, at meget på alle måder er bedre end lidt. Det afgørende må være en hensyntagen til de forudsete ordbogsfunktioner. En L1 -> L2-oversættelsesordbog skal helt sikkert omfatte flere grammatiske angivelser end det er tilfældet for en L2 -> L1-ordbog. Når det drejer sig om kollokationer, kan man opstille lignende argumenter. Efter min mening og erfaring kommer der et yderligere argument til: Hvis mængden af relevante kollokationsangivelser til et bestemt lemma bliver for stor, fx mere end en eller måske to spalter, må man have et noget snævrere selektionskriterium end ved andre ordbogsartikler. I det sidste tilfælde kunne man fx i en bilingval ordbog i sådanne tilfælde i højere grad bruge det ellers noget problematiske kriterium: er kollokationen umiddelbart forståelig og helt problemløs at oversætte. Der er her brug for yderligere metodiske overvejelser, som i fagleksikografisk sammenhæng vil kunne bygge på Pedersen (1995), hvorimod man i almensproglig leksikografisk sammenhæng snarere vil kunne sige, hvad man ikke skal gå ud fra, nemlig den teoretisk svagt funderede og også praktisk yderst problematiske teori, som den er fremlagt af Hausmann (1985). Her bør man i højere grad gå ud fra Kjellmer (1982) og Sinclair (1991). Dog i modsætning til den fuldstændige optagelse af alle kollokationer i KJELLMER vil en relatering til brugergruppe og ordbogsfunktion kunne være medvirkende til den nødvendige selektion. Den tilsvarende problematik gør sig ligeledes gældende ved valg af eksempler og ved valg hhv. fravalg af bl.a. de ortografiske, grammatiske, dialektale og stilistiske varianter, som forekommer i korpus.

8. Litteratur

8.1 Ordbøger

COBUILD = *Collins COBUILD English Language Dictionary.* Editor in Chief: John Sinclair, Managing Editor: Patrick Hanks. London/ Glasgow: Collins 1987.

DUDEN-GWB = *Duden. Das große Wörterbuch der deutschen Sprache in sechs Bänden.* Hrsg. u. bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion unter Leitung von Günther Drosdowski. Mannheim/Wien/Zürich: Bibliographisches Institut. Bd. 1 A–Ci 1976, Bd. 2 Cl–F 1976, Bd. 3 G–Kal 1977, Bd. 4 Kam–N 1978, Bd. 5 O–So 1980, Bd. 6 Sp–Z 1981.

GENTEKNOLOGISK ORDBOG = Uwe Kaufmann/Henning Bergenholtz: *Genteknologisk ordbog. Dansk-engelsk/engelsk-dansk molekylærbiologi og DNA-teknologi*. København: Gad 1992.

KJELLMER = Göran Kjellmer: *A Dictionary of English Collocations. Basis on the Brown Corpus. In Three Volumes.* Oxford: Clarendon Press 1995.

PUMPETEKNOLOGISK ORDBOG = Jette Pedersen: *A Grundfos Basic Dictionary of Pump Technology and Related Terminology.* Århus/ Bjerringbro 1995.

8.2 Anden litteratur

Bausch, Karl-Heinz 1979: Intuition und Datenerhebung in der Linguistik. Zur pragmatischen Basis linguistischer Methodologie. I: Bergenholtz/Schaeder 1979, 71–88.

Behaghel, Otto 1923: *Deutsche Syntax. Eine geschichtliche Darstellung. Bd. 1: Die Wortklassen und Wortformen. A: Nomen. Pronomen.* Heidelberg: Winter.

Bergenholtz, Henning 1985: Kongruenz der Apposition. I: *Beiträge zur Geschichte der deutschen Sprache und Literatur (Tübingen) 107,* 21–44.

Bergenholtz, Henning 1988: DK87: Et korpus med dansk almensprog. I: *Hermes 1,* 229–237.

Bergenholtz, Henning/Uwe Kaufmann (eds.) 1991: *Gene Technology Corpus.* Århus/København.

Bergenholtz, Henning/Joachim Mugdan 1989: Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität. I: *Computional Linguistics. Computerlinguistik. An International Handbook on Computer Oriented Language Research and Applications. Ein internationales Handbuch zur computerunterstützten Sprachforschung und ihrer Anwendungen,* hrsg. von István

S. Bátori/Winfried Lenders/Wolfgang Putschke. Berlin/New York: de Gruyter, 141–149.

Bergenholtz, Henning/Joachim Mugdan 1990: Formen und Probleme der Datenerhebung II: Gegenwartsbezogene synchronische Wörterbücher. I: *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. Zweiter Teilband,* hrsg. von Franz Josef Hausmann/Oskar Reichmann/Herbert Ernst Wiegand/ Ladislav Zgusta. Berlin/New York: de Gruyter, 1611–1625.

Bergenholtz, Henning/Burkhard Schaeder 1977: Deskriptive Lexikographie. I: *zeitschrift für germanistische linguistik 5,* 2–33.

Bergenholtz, Henning/Burkhard Schaeder (Hrsg.) 1979: *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora.* Königstein/Ts.: Scriptor.

Bergenholtz, Henning/Jette Pedersen 1994: Zusammensetzung von Textkorpora für die Fachlexikographie. I: Burkhard Schaeder/Henning Bergenholtz (Hrsg.): *Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern.* Tübingen: Narr, 161–176.

Bungarten, Theo 1979: Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft. I: Bergenholtz/Schaeder 1979, 28–51.

Dyrberg, Gunhild/Dorrit Faber/Steffen Leo Hansen/Joan Tournay 1988: Etablering af et juridisk tekstkorpus. I: *Hermes 1,* 209–227.

Francis, W. Nelson 1979: Problems of Assembling and Computerizing Large Corpora. I: Bergenholtz/Schaeder 1979, 110–123.

Glas, Reinhold 1975: Das LIMAS-Korpus, ein Textkorpus für die deutsche Gegenwartssprache. I: *Linguistische Berichte 40,* 63–66.

Greenbaum, Sidney 1984: Corpus Analysis and Elicitation Tests. I: *Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research,* ed. by Jan Aarts and Willem Meijs. Amsterdam: Rodopi, 193–201.

Hausmann, Franz Josef 1985: Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. I: Henning Bergenholtz/Joachim Mugdan (Hrsg.): *Lexikographie und Grammatik. Akten des Essener Kolloquiums 1984*. Tübingen: Niemeyer, 118–129.

Hausmann, Franz Josef 1989: Kleine Weltgeschichte der Metalexikographie. I: *Wörterbücher in der Diskussion. Vorträge aus dem Heidelberger Lexikographischen Kolloquium,* hrsg. von Herbert Ernst Wiegand. Tübingen: Niemeyer, 75–109.

Itkonen, Isa 1976: Was für eine Wissenschaft ist die Linguistik eigentlich? I: Dieter Wunderlich (Hrsg.): *Wissenschaftstheorie der Linguistik.* Kronberg: Athenäum, 56–76.

Johansson, Stig 1979: The Use of a Corpus in Register Analysis: The Case of Learned and Scientific English. I: Bergenholtz/Schaeder 1979, 281–293.

Kaufmann, Uwe 1993: Anvendelse af det danske genteknologiske tekstkorpus ved udarbejdelsen af Genteknologisk ordbog, med specielt henblik på udvælgelsen af eksempler. I: Gert Engel (red.): *Proceedings af seminar om korpuslingvistik i fagsprogsforskningen*. *Hindsgavl Slot 26. og 27. nov. 92.* [Kolding], 56–68.

Kjellmer, Göran 1982: Some problems in relation to the study of collocations in the Brown Corpus. I: Stig Johansson (ed.): *Computer Corpora in English Language Research.* Bergen: Norwegian Computing Centre for the Humanities, 25–33.

Lauridsen, Ole/Theis Riiber/Henning Søndergaard 1991: Erstellung eines dänischen und eines deutschen Textkorpus – Fachsprache der Gentechnik. I: *Hermes 6,* 125–137.

Maegaard, Bente/Hanne Ruus 1980: Danske almindelige ord: rangfrekvenslister og deres brug. I: *SAML 1*, 5–22.

Mentrup, Wolfgang 1979: Überlegungen zur Zusammenstellung und Verwendung eines Korpus für ein großes interdisziplinäres Wörterbuch der deutschen Sprache. I: Bergenholtz/Schaeder 1979, 182–203.

Mergenthaler, Erhard 1979: Das Textkorpus in der psychoanalytischen Forschung. I: Bergenholtz/Schaeder 1979, 131–147.

Pasch, Renate 1992: Es lebe das lexikographische Beispiel! (Probleme der lexikographischen Beschreibung wahrheitsfunktionaler Satzverknüpfer mit Kontextbeschränkungen. I: *Lexikontheorie und Wörterbuch. Wege der Verbindung von lexikologischer Forschung und lexikographischer Praxis,* hrsg. von Ursula Brauße/Dieter Viehweger. Tübingen: Niemeyer, 245–293.

Pedersen, Jette 1995: The Identification and Selection of Collocations in Technical Dictionaries. I: *Lexicographica 11,* 50–73.

Quirk, Randolph/Jan Svartvik 1979: A Corpus of Modern English. I: Bergenholtz/Schaeder 1979, 204–218.

Renouf, Antoinette 1991: The Establishment and Use of Text Corpora at Birmingham University. I: *Hermes 7,* 71–80.

Rieger, Burkhard 1979: Repräsentativität. Von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. I: Bergenholtz/Schaeder 1979, 52–70.

Rösel, Petr 1995: Kollokationen und Sublemmabestand. Ist- und Soll-Stand in englischen monolingualen und in bilingualen Wörterbü-

chern der Sprachrichtung Englisch-Deutsch. I: *Lexicographica 11*, 172–195.

Siliakus, Hendricus Johannes 1979: In Search of a Common Vocabulary for the Social Sciences and the Humanities – a Report. I: Bergenholtz/Schaeder 1979, 148–170.

Sinclair, John 1991: *Corpus. Concordance. Collocation.* Oxford: Oxford University Press.

Stumman, Bjarne 1993: Anvendelsesmuligheder og fagligt indhold af det danske genteknologiske tekskorpus. I: Gert Engel (red.): *Proceedings af seminar om korpuslingvistik i fagsprogsforskningen*. *Hindsgavl Slot 26. og 27. nov. 92.* [Kolding], 69–74.

Terminologiafdelingen 1987: *Pilotprojekt vedrørende database til terminologisk information og generering af ordbøger.* København: Handelshøjskolen i København.

van de Velde, Roger G. 1979: Probleme der linguistischen Theorienbildung einer empirischen Textwissenschaft. I: Bergenholtz/ Schaeder 1979, 10–27.

Wiegand, Herbert Ernst 1977: [referat af Bergenholtz/Schaeder 1977]. I: *Zeitschrift für germanistische linguistik 5*, 129.

Wolski, Werner 1986: Partikeln im Wörterbuch. Eine Fallstudie am Beispiel von *doch*. I: *Lexikographica 2*, 244–270.

Zöfgen, Ekkehard 1986: Kollokation – Kontextualisierung – (Beleg-) Satz. Anmerkungen zu Theorie und Praxis des lexikographischen Beispiels. I: A.Barrera–Vidal/H.Kleineidam/M.Raupach (Hrsg.): *Französische Sprachlehre und bon usage. Festschrift für Hans-Wilhelm Klein zum 75. Geburtstag.* München: Hueber, 219–238.