Henning Bergenholtz

DK87: Et korpus med dansk almensprog

Abstract

At the Aarhus School of Business a corpus of standard Danish has been established which contains 1 mio words divided into 200 texts of 5.000 words each. All the texts are original 1987 publications, 25% of them newspapers, 25% magazines and 50% fiction Furthermore, three other corpora are under preparation: a corpus of Danish, French and English within the law of contract. The corpora are distributed free of charge to linguists under the following conditions: 1. The corpus may not be further distributed and 2. the corpus may not be used for commercial purposes. The corpus will be extended to a total of 5 mio words in the course of the coming five years.

1. Diskussionen om tekstkorpora

I lingvistiske deldiscipliner som den historiske lingvistik eller fagsprogsforskningen har det næsten altid været en selvfølge at gå ud fra tekstanalyser. Inden for de områder af lingvistikken, som især beskæftiger sig med almensprogets syntaks, har det siden 60'erne været anderledes. Gang på gang er der blevet ført en næsten rituel kamp mellem dem, der tvivlede på nytten af tekstkorpusundersøgelser og dem, som forsvarede dem. Den ene part var korpusskeptikerne, som betragtede bearbejdelsen af korpora som "en unødvendig ceremoni" (Itkonen 1976,65) eller en beskæftigelse, der er ligeså fornuftig, som det ville være for en fysiker eller en biolog at nøjes med at analysere film om de hændelser, som foregår omkring os i vores dagligdag (Chomsky 1980,199). Den anden part, korporas forsvarere, syntes til dels at hylde habeas-corpus-opfattelsen: Hvis du har et korpus, så kommer de nødvendige teorier af sig selv. Som en fortsættelse af denne teori-naive indstilling ser jeg bestræbelser på at skabe en egen ny bindestregs-lingvistik, en korpus-lingvistik (til de forskellige argumenter i disse kontroverser se Bergenholtz/ Mugdan 1988). Imidlertid synes denne konflikt at have mistet sin skarphed i løbet de sidste tre-fire år. Korporas principielle værdi bliver i mindre grad bestridt, og tekstkorpusbrugere giver klarere

* Jeg vil gerne takke Sven-Olaf Poulsen for konstruktiv kritik og frugtbare diskussioner ved udarbejdelsen af denne artikel.

udtryk for de foreliggende korporas begrænsninger.

Et af de centrale temaer har været, om og hvordan man kunne opbygge et repræsentativt korpus. Efter min mening er denne diskussion blevet afsluttende behandlet af Rieger 1979, som tager afstand fra brugen af prædikatet **repræsentativ** som krav til eller vurdering af et tekstkorpus. En stikprøve kan gælde som repræsentativ, når den med henblik på visse egenskaber stemmer overens med den grundmængde, den er taget fra. Men en sådan overensstemmelse kan kun påvises, når der vides lige så meget om grundmængden som om stikprøven - hvorved en stikprøve bliver overflødig. Der findes ganske vist metoder til udvælgelse af stikprøver, som med en vis sandsynlighed (men ikke med sikkerhed) kan føre til udvælgelse af en repræsentativ stikprøve. Man kan gå ud fra et tilfældighedsprincip, hvor hvert element i grundmængden skal have samme chance for at blive valgt til en stikprøve. Denne betingelse er dog ikke opfyldt, når man fx til et korpus med romaner vælger dem ud, man tilfældigvis kender. Hvis man vil foretage en tilfældig stikprøve, kunne man i stedet tage alle romaner fra et bestemt tidsrum, give hver et nummer og få genereret det påkrævede antal numre efter lotteriprincippet.

Ofte vil en sådan fremgangsmåde ikke være mulig. Man kan så vælge en stikprøve, hvor bestemte egenskaber i stikprøven udviser den samme fordeling som i grundmængden. Denne metode bliver fx brugt ved meningsmålinger: Man spørger en lille del af befolkningen, som mht. alder, køn, erhverv, bopæl osv. er sammensat som hele befolkningen. En forudsætning for en sådan fremgangsmåde er, at egenskaberne i grundmængden er kendt. Det er netop tilfældet ved opinionsundersøgelser, hvor befolkningens sammensætning mht. alder, køn erhverv osv. er kendt; ikke-kendte og ikke-relevante egenskaber som fx hårfarve og vægt spiller her ingen rolle.

En forudsætning for at kunne opbygge et repræsentativt tekstkorpus er, at man kender grundmængden af tekster i det pågældende sprog eller ved så meget om de relevante tekstegenskaber, at en repræsentativ stikprøve kan udvælges. Endvidere skal det afgøres, om en tekst tilhører det pågældende sprog eller ej: Tilhører en 50 år gammel tekst det samme sprog som en tekst fra dette år? Skal dialekttekster tages med? Skal der medtages fagtekster? Der kan på disse og andre spørgsmål ganske vist findes svar, men de kan ikke blive så præcise, at man vil kunne angive grundmængden af tekster på dette sprog (hvis man da ikke tager et klart afgrænset område som fx alle danske romaner fra 1987).

Med andre ord vil det ikke være muligt at opbygge et korpus, som

med en rimelig sandsynlighed er repræsentativt for et bestemt sprog. Det i statistikken brugte udtryk **repræsentativ** kan derfor næppe bruges uden fare for misforståelser. Under henvisning til Bungarten 1979 bruger jeg derfor det noget mindre præcise udtryk **eksemplarisk**. Et korpus kan gælde som eksemplarisk, når man:

1. vælger en veldefineret delmængde af et sprog eller
2. antager en plausibelt lydende, hypotetisk fordeling af bestemte tekstegenskaber og lægger disse til grund for udvælgelsen

Den første metode er efter min mening at foretrække inden for de områder af fagsprogsforskningen, som beskæftiger sig med tekniske fagsprog. Den anden metode mener jeg er bedre egnet, når det gælder udforskningen af alment sprog og ikke-tekniske fagsprog.

I alle tilfælde spiller korpus' omfang og sammensætning, herunder længden af de enkelte tekster, en afgørende rolle. Det har været en udbredt, næsten traditionel opfattelse, at 1 mio tekstord ikke blot var et stort, men endda et repræsentativt udsnit af et bestemt sprog. Der kan i denne sammenhæng henvises til udtalelser om det tyske Limas-Korpus, det amerikanske Brown-Corpus eller det engelske LOB-Corpus, der alle har størrelsen 1 mio tekstord. Meget mindre tal opererer Politov 1987 med, der anser en samlet stikprøve på mellem 20.000 og 35.000 løbende ord for at være tilstrækkelig for opbygningen af fagsprogskorpora. Efter mit skøn vil en størrelsesorden på 1 mio muliggøre et eksemplarisk fagsprogskorpus. For et almensprogligt korpus vil et sådant tal derimod være utilstrækkeligt Jeg kan henvise til Bergenholtz/Mugdan 1985, hvor der bliver argumenteret for, at først et korpus i størrelsesordenen 5 mio tekstord vil være tilstrækkeligt til en leksikografisk beskrivelse af de 2.000 hyppigste lexemer og affixer i det moderne tyske almensprog. Ved planlægningen af den ikke realiserede store interdisciplinære tyske ordbog antog man, at et korpus på 50 mio tekstord (sml. Mentrup 1979) ville blive nødvendig. Hoffman/Piotrowski 1979,79 mener under henvisning til kendte statistikere, at et repræsentativt korpus for et naturligt sprog må have en størrelse på mellem 10^9 og 1.5 x 10^{14} tekstord. Når man ser bort fra den problematiske brug af **repræsentativ** kan der indvendes, at et korpus af denne størrelse ikke kan bearbejdes af nutidens computere. Denne indvending kan også gøres gældende over for Bahr 1987, som - under henvisning til Hoffman/Piotrowski - anser et tekstkorpus på ca. 500 mio tekstord for at være nødvendigt for udarbejdelsen af en ny historisk tysk ordbog.

Mht. delstikprøvernes størrelse har fx Limas-korpus og Brown-

Corpus tekster med et omfang på præcis 2.000 ord (dvs. fem til seks trykte sider pr. korpustekst). Politov foreslår, at hver delstik-prøve skal være på 200 tekstord. Hermed står man over for yderligere et problemområde: Hvis omfanget af de enkelte tekster i et korpus af en given størrelse bliver meget stor, så må antallet af tekster blive tilsvarende mindre. Faren for at få et korpus, som næppe kan få prædikatet "eksemplarisk for et bestemt sprog", vil øges tilsvarende. Hvis der derimod indgår meget små deltekster i et korpus, vil man kun kunne bruge det til ordstatistiske undersøgelser. Tekstanalyser vil man næppe kunne udføre, også til lexikografiske formål vil tekster med en længde på fx 200 tekstord kun være betinget brugbar (mange begyndelses- eller slutpassager af delteksterne ville være svært forståelige).

Indtil for nogle få år siden blev tekstkorpora indtastet rent manuelt. Den efterfølgende korrektur var meget tidskrævende, og selv efter en dobbelt korrektur måtte man regne med en betydelig fejlprocent (op til 1%). Med indsatsen af optisk udstyr er processen blevet væsentligt nemmere. Endvidere kan man v.hj.a. korrekturprogrammer finde en stor del af de fejl, som opstår ved den optiske indlæsning.

I det nu færdiggjorte danske tekstkorpus var fejlprocenten efter scanningen, men før den automatiske korrektur mindre end en promille. Der bør dog tilføjes, at denne forholdsvis lille fejlprocent kun gælder for de afleverede tekster, dvs. at den ikke afslører en anden fejlkilde, som uden tvivl er forårsaget af det monotone arbejde: Ved mere end 10% af avis- og ugebladsteksterne var overskriften "oversprunget" eller slutningen "glemt". Sandsynligvis er disse mangler (som korrekturprogrammet ikke finder) opstået i forbindelse med afbrydelser under indlæsningen, men disse fejl forårsager dog en tidskrævende yderligere arbejdsindsats.

2. Hidtidige danske korpora

Foruden de mange små, men ikke alment tilgængelige danske korpora findes der de korpora, som Bente Maegaard og Hanne Ruus har sammenstillet for hver af de følgende tekstarter: børnebøger, romaner, aviser, ugeblade og populære fagblade. Ordfrekvensen i de enkelte korpora, men ikke i den samlede tekstmængde, foreligger i bogform. Hvert korpus består af 1.000 deltekster med 250 løbende ord hver, dvs. knap en trykt side pr. korpustekst og i alt 250.000 tekstord pr. korpus. Ved længere deltekster vil der være en fare for, at temaet i en bestemt tekst kunne føre til tilfældige hyppighedsfore-

komster (Maegaaard/Ruus 1981,7). Udvalget af tekster er i princippet blevet fortaget ud fra et læsesociologisk synspunkt. Ved aviser og ugeblade er man gået ud fra de forskellige oplagstal: jo højere oplag, des større blev antallet af stikprøver fra en avis eller et ugeblad (Maegaard/Ruus 1986,II,6-9). Ved børnebøger gik man (formentlig p.gr.a. manglende læsesociologiske undersøgelser) ud fra følgende princip: Kun de danske forfattere, som i tiden 1970-1974 har fået udgivet mere end tre værker, blev taget med i udvalget, som er sammensat af genoptryk og nyudgivelser af billedbøger for de mindste, læsebøger for skolens førsteklasser og romaner for de større børn. Ved udvalget af romantekstprøver finder man tekster af "de mest læste danske forfattere" fra årene 1970-1974 (Maegaard/Ruus 1981b,5f). Listen over disse forfattere omfatter 20 navne, som udgør et bredt spektrum fra Steen Steensen Blicher (1782-1848), Hans Christian Andersen (1805-1875), Herman Bang (1857-1912), Martin A. Hansen (1909-1955) til nutidige forfattere som Klaus Rifbjerg og Anders Bodelsen.

Sammensætningen af dette korpus baserer i væsentlig grad på tekstreceptionsprincippet, som dog blev fraveget i nogen grad ved udvalget af roman- og børnebogstekstprøver, fordi selektionen ikke blev styret af oplagets størrelse. Det er desuden ikke helt overbevisende, at oplagets størrelse alene bør være udslaggivende for et receptions-orienteret udvalg. Efter min erfaring bliver visse ugeblade og aviser i højere grad end andre givet videre til venner og bekendte.

Endnu vigtigere end disse overvejelser er følgende mere principielle problemstillinger:

1. korpus' homogenitet
2. den tidsmæssige fordeling af teksterne
3. omfanget af delteksterne

Homogenitetsproblemet udtaler Maegard/Ruus 1980,8 sig meget bestemt om. De går ud fra det krav, at et tekstkorpus skal være homogent, ellers ville de statistiske resultater kun kunne gælde for de udvalgte tekster og ikke for tekstarten i sin helhed. Problemet er her, hvad der skal forstås ved "homogen". Hvor ensartet skal elementerne af en mængde være for at kunne betragtes som homogene? Vi står her overfor følgende dilemma: Jo præcisere tekstarterne er definerede, des mere ensartet vil tekstsamlingen blive, men udsagnene, som undersøgelser på denne tekstsamling vil give, vil også blive tilsvarende mindre almengyldige. Hvis man på den anden side efter receptionsprincippet går ud fra et udvalg af tekster, som læses af et flertal af den voksne befolkning, vil samlingen blive mindre homogen, men til gengæld i højere grad give mulighed for udsagn om

det pågældende almensprog.

Jeg anser de enkelte tekstarter i de omtalte fem delkorpora for at være kun betinget homogene: I romankorpus er der fx store forskelle mellem Blicher fra den første halvdel af det 19. århundrede og Rifbjerg fra den anden halvdel af det 20. århundrede - også set ud fra læserens synspunkt. Det samme gælder for Maegaard/Ruus' børnebogskorpus, hvor der er meget store forskelle mellem de yngstes første læsebog og en roman for større skoleelever.

Hvad angår antallet af tekster og delteksternes størrelse, ville det være hensigtsmæssigt med et stort antal af hele tekster eller i det mindste sammenhængende deltekster, der i sig selv danner en helhed. Små deltekster med 250 løbende ord er som sagt kun brugbare i ordstatistiske undersøgelser, men ikke velegnede til alle former for tekstanalyse. Den af Hoffman/Piotrowski anførte størrelse for et statistisk set tilstrækkeligt stort korpus vil sikkert kunne tilfredsstille kravene både mht. en bred tekstspredning og længden af tekster. Men med den nutidige tekniske udvikling kan ønsker om et korpus i den størrelsesorden ikke opfyldes.

3. Et korpus med dansk almensprog

Danlex-gruppen foreslår i Engel et al. 1987,184f, at der bliver oprettet et "stort fælles tekstkorpus i form af en database på 10 mio ord" til brug for den en- og tosprogede leksikografi i Danmark. De påpeger, at der her foreligger en national opgave, som ideelt bør støttes i form af et fast beløb på finansloven. Som et bidrag til denne opgave har vi på Handelshøjskolen i Århus sammenstillet et korpus med dansk almensprog. Principperne for udvalget og sammenstillingen blev opstillet i samarbejde med Finn Frandsen, Karen M. Lauridsen og Ole Lauridsen. Det konkrete tekstudvalg er efter de udarbejdede retningslinjer blevet udført af Lisbeth Boel, Louise Uggerhøj og Richard Almind. Tekstscanningen blev i løbet af en måned klaret af et privat firma, Henning Bruun Databaser i Frederikssund.

Der er blevet sammenstillet et korpus, kaldet DK87, der indeholder 1 mio tekstord med originaltekster fra året 1987, dvs. at hverken oversættelser eller genoptryk er medtaget. I henhold til argumenterne i kap.2 kan et sådant forholdsvis lille korpus på ingen måde betegnes som repræsentativt for dansk sprog. Derimod mener jeg nok, at man kan kalde det et **eksemplarisk** korpus for det skrevne danske **almensprog** 1987. Da de enkelte korpustekster enten er hele tekster eller sammenhængende forholdsvis store tekstdele, vil DK87 kunne blive en frugtbar materialebasis for et bredt spektrum af lingvistiske projekter, hvor tekstuelle sammenhæng spiller en væsentlig rolle - fx ved syntaktiske, tekstanalytiske eller leksikografiske projekter.

Anderledes end ved Maegaard/Ruus's korpora har vi lagt tekstproduktionen (og ikke tekstreceptionen) til grund for udvælgelsen. DK87 består af repræsentantenter for tre tekstarter:

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

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. Vi 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. Mens de 50 korpustekster med aviser hhv. med ugeblade kun udgør en meget lille del af den samlede mængde fra 1987, omfatter de 100 korpustekster med romaner og noveller mere end halvdelen af nyudgivelserne 1987. Blandt disse nyudgivelser er der ikke blevet skelnet mellem "triviallitteratur" og "skønlitteratur", da en sådan opdeling efter vores mening er teoretisk og praktisk uigennemførlig.

Mht. romaner og noveller består hver korpustekst af mindst et kapitel eller, hvis der ingen kapitelinddeling findes, af en tekstdel, som udgør en helhed. Der var tilstræbt en tekstlængde på ca. 5.000 løbende ord, som nogle gange ikke er nået og nogle gang er blevet overskredet, men i gennemsnit har teksterne en længde på 5.000 ord. De fleste tekster ligger mellem 4.000 og 5.500 ord, men en enkelt tekst har et omfang på kun 2.000 ord og fem andre på mellem 7.000 og 9.000 ord.Valget af et kapitel i en roman hhv. novelle blev foretaget efter et tilfældighedsprincip. Af de mest kendte ugeblade og et bredt udsnit af landsdækkende og lokale aviser blev tre-fem eksemplarer pr. avis hhv. ugeblad udtaget. Fra hver af disse aviser eller ugeblade blev der udvalgt hele tekster, der tilsammen udgør ca. 5.000 løbende ord.

DK87 foreligger i form af 3.5" disketter til Apple Macintosh eller 5.25" 360 MB hhv. 3.5" 1.4 MB disketter til IBM og dermed kompatible systemer. Forskere ved forskningsinstitutioner i Dan-

mark, og i begrænset omfang også i udlandet, kan få det stillet til rådighed, såfremt følgende betingelser overholdes:

1. det må ikke udnyttes kommercielt
2. det må ikke uden tilladelse gives videre til andre brugere

Disketterne indeholder ud over teksterne et konkordansprogram (kun for IBM-brugere), som kan generere en højre- eller venstrealfabetisk sætningskonkordans. Desuden indeholder den første diskette en beskrivelse af dette program såvel som en bibliografisk oversigt over korpusteksterne og henvisninger til arbejder, der behandler DK87 i sig selv, hhv. til undersøgelser, der baserer derpå. Konkordansprogrammet er skrevet af John Bergenholtz, den øvrige programmering af Jørgen Albretsen.

For årene 1988, 1989, 1990 og 1991 vil der blive sammenstillet yderligere korpora med dansk almensprog, som skal foreligge som et samlet korpus med 5 mio tekstord for årene 1987-1991. Der er (endnu) ikke planlagt korpora for de efterfølgende år, men et sådant projekt ville på længere sigt kunne blive interessant for udforskningen af udviklingen i dansk sprog.

Det her beskrevne danske korpus, DK87, men også tre korpora med engelske, franske og danske tekster inden for aftaleret (se Dyrberg et al. 1988) kan rekvireres gratis hos:

Henning Bergenholtz
Handelshøjskolen i Århus
Fuglesangs Allé 4
DK-8210 Århus V

DK87 vil foreligge fra august 1988, de tre øvrige korpora fra begyndelsen af 1989.

Litteratur

Bahr, Joachim (1987): Entwurf eines historischen Wortschatzarchivs, i: *Zeitschrift für germanistische Linguistik 15*, 1987, 141-168.

Bergenholtz, Henning/Mugdan, Joachim (1985): Grammatik im Wörterbuch: von *Ja* bis *Jux*, i: *Studien zur Neuhochdeutschen Lexikographie V*. Hrsg. von Herbert Ernst Wiegand. Hildesheim/Zürich/New York: Olms 1985, 47-102.

Bergenholtz, Henning/Mugdan, Joachim (1988): Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität, i: *Computational Linguistics. Ein internationales Handbuch computerunterstützter Sprachforschung und ihrer Anwendung*. Hrsg. von István Bátori et al. Berlin/New York: de Gruyter (i trykken).

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

Bungarten, Theo (1979): Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft, i: Bergenholtz/Schaeder 1979, 28-51.

Chomsky, Noam (1980): Rules and Representations. New York: Columbia University Press 1980.

Dyrberg, Gunhild/Faber, Dorrit/Hansen, Steffen Leo/Turnay, Joan (1988): Etablering af et juridisk tekstkorpus, i: *Hermes 1*, 1988.

Engel, Gert /Jacobsen, Jane Rosenkilde/Madsen, Bodil Nistrup/Hjorth, Ebba/Norling-Christensen, Ole/Ruus, Hanne (1987): *Ordbøger i Danmark. En oversigt*. København 1987.

Hoffmann, L./Piotrowski, R. G. (1979): *Beiträge zur Sprachstatistik*. Leipzig: VEB Verlag Enzyklopädie 1979.

Itkonen, Esa (1976): Was für eine Wissenschaft ist die Linguistik eigentlich?, i: *Wissenschaftstheorie der Linguistik*. Hrsg. von Dieter Wunderlich. Kronberg: Athenäum 1976, 56-76.

Maegaard, Bente/Ruus, Hanne (1978): DANwORD. Hyppighedsundersøgelser i moderne dansk: Baggrund og materiale, i: *danske studier 1978*, 42-70.

Maegaard, Bente/Ruus, Hanne (1980): Danske almindelige ord: rangfrekvenslister og deres brug, i: *SALM 1*, 1980, 5-22.

Maegaard, Bente/Ruus, Hanne (1981a): *Hyppige ord i Danske Børnebøger. København*: Gyldendal 1981.

Maegaard, Bente/Ruus, Hanne (1981b): *Hyppige ord i Danske Romaner*. København: Gyldendal 1981.

Maegaard, Bente/Ruus, Hanne (1986a): *Hyppige ord i Danske Aviser, Ugeblade og Fagblade*. 2 bd. København: Gyldendal 1981.

Mentrup, Wolfgang (1979): Überlegungen zur Zusammenstellung und Verwendung eines Korpus für ein großes interdisziplinäres Wörterbuch der deutschen Sprache, i: Bergenholtz/Schaeder 1979, 128-203.

Politov, Stefan (1987): Zur Entwicklung der statistischen Fachsprachenlexikforschung. *Special Language, i: Fachsprache 9*, 1987, 149-166.

Rieger, Burkhard (1979): Repräsentativität. Von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung, i: Bergenholtz/Schaeder 1979, 52-70.