

## DK87-DK90: DANSK KORPUS MED ALMENSPROGLIGE TEKSTER

Af Henning Bergenholtz (Handelshøjskolen i Århus)

### 1. Habeas Corpus<sup>1</sup>

I lingvistiske deldiscipliner, såsom 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. Altså på den ene side korpuskeptikerne, 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). På den anden side stod korporas forsvarere, som syntes til dels at hylde habeas-corpus-opfattelsen: Hvis du har et korpus, så kommer de nødvendige teorier af sig selv. Som en indirekte støtte til denne teori-naivisme betragter jeg bestræbelserne på at skabe en egen ny bindestrengs-lingvistik, **corpus-linguistics**, en betegnelse, som især bliver brugt i forbindelse med analysen af engelske korpora (til de forskellige argumenter i disse kontroverser se Bergenholtz/Mugdan 1989). Imidlertid synes denne konflikt at have mistet sin skarphed i løbet af det sidste årti. Korporas principielle værdi bliver i mindre grad bestridt, og tekstkorpusbrugere giver klarere udtryk for de foreliggende korporas begrænsninger.

### 2. Udvalg af tekster til et korpus

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. Det er ganske vist muligt at foretage en stikprøve, som med en vis sandsynlighed kan føre til udvælgelse af en repræsentativ stikprøve, idet man går ud fra et tilfældighedsprincip, hvor hvert element i grundmængden skal have samme chance for at blive valgt. 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 teksteegenskaber, 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 i 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 teksteegenskaber og lægger disse til grund for udvælgelsen.

I alle tilfælde spiller korpus' omfang og sammensætning, herunder længden af de enkelte tekster, en afgørende rolle. Det har længe været en udbredt, næsten traditionel opfattelse, at 1 mio tekstord ikke blot var ét 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 omfatter netop 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. Selv et korpus på 1 mio løbende ord vil efter min foreløbige erfaring med det bioteknologiske tekstkorpus (se O.Lauridsen i dette bind) ikke kunne udgøre et tilstrækkeligt bredt materiale til brug i fagleksikografien<sup>2</sup>. For et almensprogligt korpus vil et sådant omfang i alle tilfælde være utilstrækkeligt. Jeg kan henvise til Bergenholtz/Mugdan (1985), hvor der bliver argumenteret for, at først et korpus i størrelsesordenen 5 mio tekstord danner tilstrækkeligt grundlag for en videnskabeligt holdbar leksikografisk beskrivelse af de 2.000 hyppigste lexemer og affixer i moderne tyske almensprog. Ved planlægningen af den ikke realiserede store interdisciplinære tyske ordbog antog man, at et korpus på 50 mio tekstord ville blive nødvendigt (sml Mentrup 1979). Korpora af denne størrelse er planlagt af de tre store engelske ordbogsforlag. 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  $1.5 \times 10^{14}$  tekstord. Når man ser bort fra den problematiske brug af repræsentativ kan dertil indvendes, at et korpus af denne størrelse ikke kan bearbejdes af nutidens computere. Denne indvending kan også gøres gældende overfor 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 indeholder fx Limas-korpus og Brown-Corpus tekster med et omfang på præcis 2.000 ord (dvs 5-6 trykte sider pr korpustekst). Politov foreslår, at hver delstikprøve 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 stort, 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 leksikografiske formål vil tekster med en længde på fx 200 tekstord kun være betinget brugbar (mange begyndelses- eller slutpassager af delteksterne ville være svært forståelige).

### 3. Hittidige 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 hver 250 løbende ord, dvs knap en trykt side pr korpustekst og i alt 250.000 tekstord pr korpus. Ved længere deltekster vil der opstå en fare for, at temaet i en bestemt tekst kan føre til tilfældige hyppighedsforekomster (Maegaard/Ruus 1981a,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, der er sammensat af genoptryk og nyudgivelser af billedbøger for de mindste, læsebøger for skolens førsteklasse og romaner for de større børn. Under romantekstprøverne 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 udslagsgivende for et receptionsorienteret 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 Maegaard/Ruus (1980,8) sig meget bestemt om. De kræver, at et tekstkorpus skal være homogent, fordi de statistiske resultater ellers kun ville gælde for de udvalgte tekster og ikke for teksten i sin helhed. Problemet er her, hvad der skal forstås ved "homogen". Hvor ensartede 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 almenlydige. 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 deres 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 tilstrækkelig stort korpus vil sikkert kunne tilfredsstille kravene både mht en bred tekstspredning og til længden af tekster. Men med den nutidige tekniske udvikling kan ønsket om et korpus i den størrelsesorden ikke opfyldes.

### 4. Tekstudvalg til DK87-90

\* For hvert år er der udvalgt en tekstmængde på 1 mio tekstord, som består af tekster hhv deltekster fra 100 nyudkomne romaner, 50 ugeblade og 50 aviser. Fra hver roman er der udvalgt et kapitel på ca 5000 ord, fra hver avis og ugeblad er der medtaget en række større og mindre bidrag ligeledes i et omfang på ca 5000 ord. Forholdet mellem de tre tekstarter udgør således

romaner og noveller (50% af tekstmængden)  
aviser (25% af tekstmængden)  
ugeblade (25% af tekstmængden)

Hver af disse tre tekstarter læses af en betydelig og bred del 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. Jeg holder derfor - i modsætning til fx Frandsen (1990), men på lignende vis som Mikkelsen (1990) - fast ved betegnelsen **almensprogligt** korpus, evt. ville **fællessprogligt** korpus være mere træffende. Den for tiden herskende tendens indenfor fagsprogsforskningen, som under indflydelse af det uheldige udtryk LSP (Language for Specific Purposes) anser så godt som alle tekster for at være fagsproglige, er ikke bare kontraintuitiv, den er efter min mening ikke frugtbar for den virkelige fagsprogsforskning.

Derimod er den procentuelle opdeling mellem fiktionale tekster, aviser og ugeblade 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 korpusstekster med aviser hhv med ugeblade kun udgør en meget lille del af den samlede mængde fra hvert af årene 1987-90, omfatter de 100 korpusstekster med romaner og noveller mere end halvdelen af nyudgivelserne i de pågældende år. Blandt disse nyudgivelser er der ikke blevet skelnet mellem "trivallitteratur" og "skønlitteratur", da en sådan opdeling efter vores mening er teoretisk og praktisk uigennemførlig.

Mht romaner og noveller består hver korpus tekst 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; her er dog naturligvis tale om en gennemsnitslængde. 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<sup>3</sup>.

##### 5. Første hyppighedsundersøgelser på grundlag af DK87-90

En første bearbejdelse af korpus foreligger i form af en frekvensordbog (Bergenholtz 1989), som indeholder en alfabetisk sorteret liste af samtlige ordformer med angivelse af hyppigheden i DK87 og DK88 og i de enkelte tekstarter. Derefter følger en liste med en sortering efter hyppighed, hvor der er medtaget alle ordformer med en absolut hyppighed på mere end 30 forekomster. Mit arbejde med denne frekvensordbog viser, at omfanget af de enkelte tekster (5.000 løbende ord) i for høj grad påvirker hyppigheden af alle ord med en frekvens på under 50, til en sådan statistisk undersøgelse ville en større mængde af forskellige, men kortere tekster - som ved Maegaard/Ruus' korpus - have været fordelagtigere. Med det dobbelte tekstgrundlag som udgangspunkt vil denne skævhed til dels kunne opfanges. Frekvensordbogen vil i derfor i løbet af 1991 foreligge i en ny bearbejdelse, der også medtager materialet fra DK89 og DK90.

Den anlagte fremgangsmåde med et korpus for hvert år implicerer på noget længere sigt muligheden for diakrone undersøgelser, det gælder selvfølgelig i særlig grad, hvis sådanne korpora bliver samlet over en lang årrække. Ved en sammenligning af DK87 med DK88 og DK89 har jeg forsøgt at finde frem til ord, som er nye i dansk sprog eller som er blevet brugt i klart større omfang. Resultatet er ikke overvældende interessant, på så kort tid kan man med så forholdsvis små tekstmængder ikke finde tydelige ændringer.

Tanken bag et lag af tekstkorpora med 1 mio tekstord fra hvert år var bl.a., at disse tekster skulle danne grundlaget for en senere undersøgelse af udviklingen af dansk sprog. Denne tanke mener jeg stadig at kunne opretholde. Det vil dog gælde i særlig grad, hvis man kunne sammenligne fx DK87 med et DK 97 eller et DK 77. Man vil sandsynligvis også kunne få interessante resultater ud af en sammenligning af hyppigheder i Maegaard/Ruus-korpus med mine korpora. Den vil fx vise, at en

række engelske ord først de sidste ti år er trængt ind i danske romaner. Men i første omgang er der foretaget sammenligninger mellem de enkelte DK-korpora. Jeg har ladet udarbejde et program, hvor man kan indgive en frekvensdifference mellem to korpora enten som et absolut tal eller i form af en faktor i stigningen i hyppigheden. For at undgå at skulle tage hensyn til tilfældige sving i den nederste del af frekvensskalaen ansættes en mindste stigning i hyppighed med en faktor 3 og samtidig en mindste stigning på 20 forekomster som minimale krav til signifikans. Ved gennemgang af materialet på basis af dette program er resultaterne skuffende; de viser, at sproget kun ændrer sig i ubetydelig grad fra et år til et andet.

| Ord          | 88  | 87 | 89  | 88 | 89 | 87 | Maegaard & Ruus |
|--------------|-----|----|-----|----|----|----|-----------------|
| KM/T         | 98  | 3  |     |    |    |    | 39              |
| KONCERNEN    |     |    |     |    | 24 | 3  |                 |
| KRIMINELLE   | 23  | 3  |     |    |    |    | 5               |
| KVADRATMETER |     |    |     |    | 33 | 10 | 19              |
| KÆRLIG       | 33  | 7  |     |    |    |    | 4               |
| LEJE         |     |    | 42  | 14 |    |    | 42              |
| LILLA        | 112 | 10 |     |    |    |    |                 |
| LP           | 44  | 12 |     |    |    |    | 4               |
| LUFTHAVN     | 30  | 6  |     |    |    |    | 25              |
| MENIGHEDEN   |     |    | 26  | 5  | 26 | 5  |                 |
| MESTER       | 47  | 12 |     |    | 44 | 12 | 31              |
| MIO.         |     |    | 137 | 35 |    |    | 5               |
| MODE         | 44  | 13 |     |    |    |    | 44              |
| MODEL        |     |    |     |    | 57 | 18 | 130             |
| MONSIEUR     | 31  | 10 |     |    |    |    |                 |
| MORD         | 75  | 25 |     |    |    |    | 16              |
| MOSEN        | 26  | 4  |     |    |    |    | 5               |
| MUS*         |     |    | 78  | 15 | 78 | 19 | 17              |
| OKTAN        |     |    | 23  | 3  | 23 | 0  |                 |
| OSSE         | 141 | 45 |     |    |    |    | 44              |

Kun ved manuel gennemgang af materialet kan man finde nye ord som *glasnost* eller *perestrojka*. Og kun ved navnes hyppigheder kan man måske se visse tendenser:

| Ord          | 88 | 87 | 89  | 88 | 89  | 87 |
|--------------|----|----|-----|----|-----|----|
| BERIT        | 34 | 4  |     |    |     |    |
| BERTEL       |    |    | 39  | 11 | 39  | 13 |
| BIRGER       |    |    | 42  | 5  |     |    |
| BJØRN        |    |    | 141 | 40 | 141 | 41 |
| BONNIE       | 34 | 5  |     |    |     |    |
| BRYAN        |    |    | 54  | 0  | 54  | 3  |
| BØRGE        |    |    |     |    | 46  | 9  |
| CAROLINE     | 36 | 4  |     |    |     |    |
| CHARLES      | 65 | 8  |     |    |     |    |
| CHRISTIANS   |    |    | 26  | 2  | 26  | 3  |
| CHRISTIANSEN | 25 | 3  |     |    |     |    |

|         |    |    |    |   |
|---------|----|----|----|---|
| CLAUS   | 84 | 25 |    |   |
| CLAUSEN | 29 | 9  |    |   |
| DANIEL  |    |    | 29 | 4 |
| DANNY   | 66 | 3  |    |   |
| DIANA   | 58 | 5  | 27 | 5 |
| DITTE   | 57 | 9  | 31 | 9 |
| DORTHE  | 33 | 6  |    |   |

Virkelig interessante resultater vil som antydnet først kunne opnås ved sammenligning af korpora, der genspejler udviklingen over en periode på mindst 10 år. Mere givtige her og nu har undersøgelser af sproglige tvivlstilfælde været. Sådanne undersøgelser kunne, hvis de blev foretaget systematisk, føre til klarere og sikrere oplysninger i håndbøger om dansk sprog, især med hensyn til ortografien:

| Ord         | Håndbog i<br>Nudansk | Dansk<br>Sprogbrug | DK 87 | DK 88 | DK 89 | Total DK |
|-------------|----------------------|--------------------|-------|-------|-------|----------|
| tillykke    | +                    |                    | 3     | 2     | 0     | 5        |
| til lykke   | +                    |                    | 7     | 8     | 11    | 26       |
| ingenting   | +                    | +                  | 137   | 128   | 100   | 365      |
| ingen ting  | +                    |                    | 4     | 2     | 3     | 9        |
| nogensinde  | +                    |                    | 76    | 86    | 72    | 234      |
| nogen sinde | +                    |                    | 58    | 71    | 55    | 184      |
| tilfælles   | +                    |                    | 1     | 4     | 3     | 8        |
| til fælles  | +                    |                    | 9     | 11    | 4     | 24       |

Særligt spændende er resultaterne af en undersøgelse af sætningslængde, udtrykt i antallet af ord, hvor jeg er gået ud fra en definition af en sætning hhv periode, der afsluttes med et punktum, udråbstegn, spørgsmålstegn eller kolon (forkortelsespunktum er der taget hensyn til ved hjælp af en liste over forkortelser). Resultaterne er meget stabile fra år til år:

|                 | DK87  | DK88  | DK89  |
|-----------------|-------|-------|-------|
| Totalgennemsnit | 14,79 | 14,53 | 15,10 |
| Aviser          | 19,68 | 18,26 | 18,56 |
| Blade           | 15,49 | 15,10 | 18,56 |
| Bøger           | 12,91 | 12,95 | 13,49 |

| Avis              | DK87  | DK88  | DK89  |
|-------------------|-------|-------|-------|
| B.T.              | 17,50 | 15,59 | 16,45 |
| Ekstra Bladet     | 16,82 | 15,61 | 15,76 |
| Jyllands-Posten   | 20,72 | 19,50 | 19,80 |
| Midtjyllands Avis | 20,25 | 18,95 | 20,20 |
| Poliitken         | 19,76 | 20,10 | 20,55 |

| Ugeblad         | DK87  | DK88  | DK89  |
|-----------------|-------|-------|-------|
| Alt for Damerne | 19,55 | 19,31 | 17,99 |
| Billed Bladet   | 13,35 | 14,46 | 13,69 |
| Hjemmet         | 15,53 | 13,22 | 14,00 |
| Se og Hør       | 15,64 | 17,55 | 17,90 |

Ud fra disse tal kan der ikke aflæses nogen form for læsbarhed, som fx: jo kortere sætninger, jo lettere læsbar er de. Men konstansen i sætningslængderne hos de forskellige medier er så påfaldende, at der bør søges efter forklaringer. I modsætning til aviser og ugeblade er der en meget stor spredning hos romanforfatterne. De foreliggende data kunne blive meget interessante for den tekstlingvistiske analyse af de nyeste danske forfattere, sml:

## DK87 Romaner

| ≤ 8,99 |                                                       | ≥ 20,00 |                                                                    |       |
|--------|-------------------------------------------------------|---------|--------------------------------------------------------------------|-------|
| 107    | Andersen, Vita<br><i>Hva'for en hånd vil du ha'.</i>  | 8,46    | 135 Enemark, Viggo<br><i>Bobler i Mudderet. Fortællinger.</i>      | 20,50 |
| 109    | Bang, Carl<br><i>Hvem lyser paven i band? .</i>       | 8,23    | 143 Johansen, Anders<br><i>Støjen fra den herskende stilhed. .</i> | 26,61 |
| 139    | Jensen, Bjarne<br><i>Kærlighed er altings hjerte.</i> | 7,59    | 146 Larsen, John Fellow<br><i>Odensefortælling</i>                 | 73,70 |
| 162    | Nielsen, Niels E<br><i>Awang Dybets vulkaner.</i>     | 8,32    | 151 Mathiasen, Jørgen<br><i>Corpus delicti.</i>                    | 20,40 |
| 164    | Nielsen, Peter<br><i>Stik.</i>                        | 8,26    | 152 Metz, Georg<br><i>De tider.</i>                                | 21,79 |
| 182    | Sandquist, Britta<br><i>Stratocaster.</i>             | 8,55    | 174 Pouplier, Erik<br><i>Machiavelli.</i>                          | 22,25 |
| 183    | Sanne, Inge<br><i>Paradisaldalen.</i>                 | 7,44    | 175 Preisler, Juliane<br><i>Du lille.</i>                          | 27,90 |

## DK88 Romaner

| ≤ 8,99 |                                               | ≥ 20,00 |                                                                                                      |        |
|--------|-----------------------------------------------|---------|------------------------------------------------------------------------------------------------------|--------|
| 108    | Berg, Anneli<br><i>Det ene med det andet.</i> | 7,15    | 115 Brovst, Bjarne Nielsen<br><i>Jeppe-Marie.</i>                                                    | 22,15  |
| 151    | Lassen, Cæcilie<br><i>Lulu:</i>               | 8,20    | 121 Ericsson, Marie-Louise<br><i>Farmors Hotel.</i>                                                  | 21,27  |
| 160    | Mørch, Dea Trier<br><i>Skibet i flasken.</i>  | 8,65    | 123 Eriksen, Jens-Martin<br><i>Rejse under mørket.</i>                                               | 22,48  |
| 163    | Nielsen, Niels E<br><i>Raslen med nøgler</i>  | 7,50    | 129 Hansen, Gynther<br><i>Hitler, min far og mig.</i>                                                | 21,80  |
| 180*   | Roulund, Grete<br><i>Dødens horisont.</i>     | 8,44    | 130 Hansen, Jørgen Christian<br><i>Dværgene.</i>                                                     | 21,20  |
| 197    | Vestergaard, Per<br><i>Huset.</i>             | 8,20    | 134 Helleberg, Maria<br><i>Så mangen sti vild.</i>                                                   | 24,25  |
|        |                                               |         | 136 Heyde, Aase<br><i>Majbørnene.</i>                                                                | 20,91  |
|        |                                               |         | 138 Hove, Peder<br><i>En dag i marts - Shellhusbombardementet.</i>                                   | 20,97  |
|        |                                               |         | 139 Hultberg, Peer<br><i>Slagne veje - Fortællinger.</i>                                             | 4385,0 |
|        |                                               |         | 140 Inuk, Lotte<br><i>Søskendesjæle.</i>                                                             | 27,75  |
|        |                                               |         | 152 Lund, Niels<br><i>Magnetbjerget.</i>                                                             | 30,10  |
|        |                                               |         | 186 Septemberfortællingerne 3<br><i>7 Fortællinger om mødet. Udvalgt novelle af Svend Age Madsen</i> | 23,12  |
|        |                                               |         | 193 Thomsen, Christian Braad<br><i>Den fortabte søns hjemkomst.</i>                                  | 21,34  |

## DK89 Romaner

| ≤ 8,99 |                                                                                      | ≥ 20,00 |                                                                                                                                                                                           |       |
|--------|--------------------------------------------------------------------------------------|---------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
| 104    | Andersen, Dortea Birkedal<br>Sallys Bizniz 17 op.                                    | 8,49    | 106 Arbejderliv<br>Noveller fra SiD-familien.<br>Udvalgte noveller af Jensen,<br>Johannes Thorwald: Han så<br>Lazarus langt borte og Hattens,<br>Ole Sand: Rapport fra en<br>arbejdsplads | 29,26 |
| 105    | Andreasen, Jan<br>Vi græder og vi længes efter kærlighed.-<br>historien om en dreng. | 8,65    | 110 Bjelke, Henrik<br>Nattens Budapest og andre<br>topografier.                                                                                                                           | 29,61 |
| 107    | Asmussen, Peter<br>Stemme.                                                           | 7,21    | 123 Eriksen, Jens-Martin<br>Jim og jeg.                                                                                                                                                   | 35,18 |
| 147    | Jørgensen, Jens C<br>På livstid.                                                     | 8,64    | 129 Hansen, Gynther<br>Soldaterne.                                                                                                                                                        | 20,91 |
| 158    | Larsen, Marianne<br>En skønne dag.                                                   | 8,80    | 138 Hultberg, Peer<br>Præludier.                                                                                                                                                          | 24,71 |
| 163    | Madsen, Ove<br>Bønder. En truet art.                                                 | 8,86    | 146 Jørgensen, Erik Lau<br>På lykke og fromme.                                                                                                                                            | 20,21 |
| 197    | Vizki, Morti<br>Forleden hos stjernekastere.                                         | 8,94    | 155 Larsen, Asger Nørgaard<br>Rejsen til verdens ende.                                                                                                                                    | 25,17 |
|        |                                                                                      |         | 175 Riffbjerg, Klaus<br>Det svage køn.                                                                                                                                                    | 21,46 |
|        |                                                                                      |         | 183 Skov, Christian<br>Ingers hvide ben.                                                                                                                                                  | 21,82 |
|        |                                                                                      |         | 189 Stangerup, Henrik<br>Den kvarte sandhed. Fra et tiår.                                                                                                                                 | 20,87 |
|        |                                                                                      |         | 199 Welner, Mikael<br>Øjet på ham.                                                                                                                                                        | 22,29 |

Påfaldende er John Fellow Larsens "Odensefortællinger" fra DK87 og i endnu højere grad Peer Hultbergs "Slagne veje" fra DK88; her skyldes resultatet dog computerprogrammet, som forudsætter brug af sætningstegn som punktum, udråbstegn eller spørgsmålstegn. Overraskende er måske SiD's novellesamling fra DK89 med en påfaldende lang gennemsnitssætning, forudsat at den formodede læsergruppe skal være SiD's medlemmer. Først en nøjagtigere analyse, der også tager hensyn til delsætnings længde og indbyrdes relationer vil kunne føre til relevante interpretationer. I dette bidrag lægges nogle af de foreliggende data frem, som også skal sammenholdes med de tekster, som mht sætningslængde ligger indenfor romaners og novellers "normale" ramme:

## DK89 Romaner

| ≥ 9,00 < 17,25 |                                                                                                     | ≥ 17,25 < 20,0 |                                                                      |       |
|----------------|-----------------------------------------------------------------------------------------------------|----------------|----------------------------------------------------------------------|-------|
| 101            | Aamund, Jane<br><i>Klinkevals.</i>                                                                  | 13,57          | 111 Blædel, Leif<br><i>Et forår i Minefeltet. Erindringsstumper.</i> | 19,49 |
| 102            | Agger, Ragnhild<br><i>Den halve sol.</i>                                                            | 11,56          | 119 Christensen, Martha<br><i>Manden som, ville ingen ondt...</i>    | 19,39 |
| 103            | Amdrup, Erik<br><i>Renters rente.</i>                                                               | 10,30          | 124 Fenigsen, Alicia<br><i>Og på den anden side.</i>                 | 18,40 |
| 108            | Bang, Carl<br><i>Omvæltninger.</i>                                                                  | 9,21           | 128 Hansen, Eva Brink<br><i>Mit eventyrlige Skranefjell.</i>         | 17,51 |
| 115            | Bødker, Cecil<br><i>Tavs.</i>                                                                       | 12,70          | 130 Hansen, Jørgen Christian<br><i>Den Skrigende Buddha. Essays.</i> | 17,66 |
| 116            | Bødker, Cecil<br><i>Vandgården.</i>                                                                 | 14,52          | 132 Helleberg, Maria<br><i>Som en vredens plov.</i>                  | 19,34 |
| 117            | Borum, Poul<br><i>Glæden ved at vågne om morgenen og<br/>glæden ved at falde i søvn om aftenen.</i> | 16,37          | 134 Holger-Nielsen, Jesper<br><i>Skyggedans.</i>                     | 17,25 |

|     |                                                                  |       |                                                                                                                                                      |       |
|-----|------------------------------------------------------------------|-------|------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
| 118 | Brovst, Bjarne Nielsen<br><i>Lise Munk. En pige fra Vedersø.</i> | 14,46 | 143 Jepsen, Hans Lyngby<br><i>Til Andrea.</i>                                                                                                        | 18,64 |
| 120 | Christensen, Ole Frøsløv<br><i>Tornen i øjet.</i>                | 11,60 | 145 Jonshøj, Ole<br><i>Rindende vand. En huslæge fortæller.</i>                                                                                      | 17,41 |
| 121 | Dyrberg, Gunnar<br><i>Kom maj....</i>                            | 13,33 | 153 Krasilnikoff, Arthur<br><i>Alting flyder eller Panta Rai. Af en ung sømands breve til sin elskede.</i>                                           | 18,77 |
| 136 | Holst, Kirsten<br><i>Sov Dukke Lise.</i>                         | 10,29 | 164 Madsen, Svend Åge<br><i>At fortælle menneskene.</i>                                                                                              | 19,90 |
| 137 | Hove, Peder<br><i>Det store spring.</i>                          | 13,39 | 177 Rosen, Marianne<br><i>Den utrolige og sandfærdige historie im den billedskønne Donatella og hendes forfærdelige familie fortalt af en moder.</i> | 17,76 |
| 139 | Jakobsen, Søren<br><i>Den italienske forbindelse.</i>            | 12,21 | 180 Schønberg, Bent<br><i>Hans Brenaa. Multimenneske.</i>                                                                                            | 18,50 |
| 140 | Jarlskov, Flemming<br><i>Skjult kamera.</i>                      | 13,95 | 198 Clausen, Torben Bjerg<br><i>Regina og Bruno Wechseltmann Slår bro fra kyst til kyst.</i>                                                         | 18,56 |

## 6. DK91ff

I særlig grad vil DK87-90 være velegnet til brug i leksikografisk sammenhæng og til forskellige former for tekstanalyser, herunder selvfølgelig også grammatiske analyser. De midler, jeg har til rådighed, kan kun dække udvælgelse og indlæsning af de nævnte fire korpora. Det vil, som Engel et al. (1987) skriver, være en national opgave at videreføre de tiltag, som blev påbegyndt af Maegaard/Ruus og genoptaget med oprettelsen af DK87-90. Jeg vil hermed opfordre et af Institutterne for Nordisk Sprog til at påtage sig dette arbejde, som måske kunne finde en yderligere begrundelse i et kollokvium med foredrag om foreliggende analyser, som brugerne af de nævnte korpora indtil nu har gennemført.

## 7. Litteratur:

- Bahr, Joachim (1987): Entwurf eines historischen Wortschatzarchivs; in: *Zeitschrift für germanistische Linguistik* 15, 1987, 141-168.
- Bergenholtz, Henning (1988): DK87: Et korpus med dansk almensprog; in: *Hermes* 1, 1988, 229-237.
- Bergenholtz, Henning (1989): *Frekvensordbog. Baseret på tekster fra danske romaner, ugeblade og aviser fra 1987-1989.* Århus: Århus Handelshøjskole 1989.
- Bergenholtz, Henning/Mugdan, Joachim (1985): Grammatik im Wörterbuch: von *Ja* bis *Jux*; in: *Studien zur Neuhochdeutschen Lexikographie* V. Hrsg. von Herbert Ernst Wiegand. Hildesheim/Zürich/New York: Olms 1985, 47-102.
- Bergenholtz, Henning/Mugdan, Joachim (1989): Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität; in: *Computational 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. Batori, Winfried Lenders, Wolfgang Putschke. Berlin/New York: de Gruyter 1989, 141-149.
- Bergenholtz, Henning/Schæder, Burkhard (Hrsg.) (1979): *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Korpora.* Königstein/Ts.: Scriptor 1979.
- Bungarten, Theo (1979): Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft; in: *Bergenholtz/Schæder* 1979, 28-51.
- Chomsky, Noam (1980): *Rules and Representations.* New York: Columbia University Press 1980.
- Dansk Sprogbrug = Erik Brun: *Dansk Sprogbrug. En stil- og konstruktionsordbog.* København: Gyldendal 1980.
- Engel, Gert/Jacobsen, Jane Rosenkilde/Madsen, Bodil Nistrup/Hjorth, Ebba/Norling-Christensen, Ole/Ruus, Hanne (1987): *Ordbøger i Danmark. En oversigt.* København 1987.
- Frandsen, Finn (1990): "Men kan man overhovedet forske og undervise i almensprog?"; in: *Netværk LSP SSP. Nyhedsbrev* 1, 1990, 12-15.
- Håndbog i Nudansk = Henrik Galberg Jacobsen/Peter Stray Jørgensen: *Håndbog i nudansk.* København: Politikens Forlag.
- 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?; in: *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; in: *danske studier* 1978, 42-70.
- Maegaard, Bente/Ruus, Hanne (1980): Danske almindelige ord: rangfrekvenslister og deres brug; in: *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 (1986): *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; in: *Bergenholtz/Schaeder 1979*, 128-203.
- Mikkelsen, Hans-Kristian (1989): Fagsprog og almensprog: særsprog og fællessprog; in: *Netværk LSP SSP. Nyhedsbrev* 2, 1990, 1-6.
- Politov, Stefan (1987): Zur Entwicklung der statistischen Fachsprachenlexikforschung. *Special Language*; in: *Fachsprache* 9, 1987, 149-166.
- Rieger, Burkhard (1979): Repräsentativität. Von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung; in: *Bergenholtz/Schaeder 1979*, 52-70.

#### Noter:

- 1) I dette bidrag gentages noget af argumentationen fra Bergenholtz-1988; som dog modificeres på enkelte punkter.
- 2) Hermed modificeres formodningen i Bergenholtz 1988, 231, hvor et omfang på 1 mio tekstord antoges at være tilstrækkelig for prædikatet **eksemplarisk** fagtekstkorpus.
- 3) De her beskrevne korpora stilles gratis til rådighed; interesserede forskere bedes henvende sig til Handelshøjskolen i Århus.

Mette Kunøe og Erik Vive Larsen (udg.):  
3. Møde om Udforskningen af Dansk Sprog  
Århus 1990

## De refleksive pronomener i dansk

Af Søren Brandt

### 1 Indledende bemærkninger

Danske grammatikker beskriver brugen af de refleksive pronomener dels ved *subjektreglen* hvorefter 'sig og sin henviser til sætningens grundled' og dels ved en række undtagelser hertil, idet 'I nogle tilfælde henviser sig og sin til et andet ord end sætningens grundled' (Mikkelsen 1911, 258). I modsætning hertil benytter vi den af (Diderichsen 1939) beskrevne regel som man kunne kalde *omsagnsreglen*, hvorefter et refleksiv i omsagnsdelen af en prædikation skal have dennes subjekt eller *basis* som korrelat. I en prædikation på sætningsform er subjektet *basis* og resten omsagn, så subjektreglen følger af omsagnsreglen; men den dækker også de øvrige tilfælde hvor refleksiv kan benyttes.

Nærværende beskrivelse, der er en komprimeret udgave af (Brandt 1990), er baseret på det sprog jeg mener jeg selv benytter, idet det bemærkes at jeg nok er mere liberal med hensyn til at acceptere refleksiv reference end det er sædvanligt blandt akademikere. Denne liberalitet afspejler derimod en sprogbrug, man ofte finder i almindeligt utvunget tale- og skriftsprog, og under alle omstændigheder må man som sprogbeskriver forlade sig på sin egen sprogforfølelse når man vurderer en lang række mere eller mindre hypotetiske og ofte af andre grunde lidet mundrette konstruktioner.

Vi følger Chomskys skelnen mellem anaforske og pronominal elementer, men betegner her disse elementer som henholdsvis *heteronome* og *autonome* pronomener. Disse betegnelser afspejler den regel at heteronome pronomener skal have et *korrelat*, mens dette ikke gælder de autonome pronomener. I de efterfølgende eksempler er korrelater ofte angivet ved hjælp af indekser.

### 2 Projektiv syntaks

Ved udformningen af beskrivelsen benytter vi hvad vi skal kalde *projektiv syntaks*. Dens grundlæggende konvention er at alle syntaktiske konstruktioner, når der ses bort fra ledfølgen, kan beskrives ved genskrivningsreglen:

$$X' \Rightarrow X' Y'$$

Reglen symboliserer at et led af en bestemt type og niveau (bar), her angivet som  $X'$ , opbygges af et *kerneled* af samme type samt ingen, et eller flere *adled* af vilkårlig type. Kerneleddet har enten samme niveau som hele leddet eller det næste lavere niveau. Ledtypen har en af værdierne  $N$ [ominall],  $V$ [erball],  $A$ [djektiv] eller  $P$ [ræposition], og niveauet antages at ligge i intervallet fra 0 til 2, hvor 0 er det *leksikalske* niveau, så  $X^0$  symboliserer et leksikalsk element som er *kernen* i konstruktionen. De højere niveauer  $X^1$  og  $X^2$  betegnes som *projektioner* af  $X^0$ . En projektion på det højeste mulige