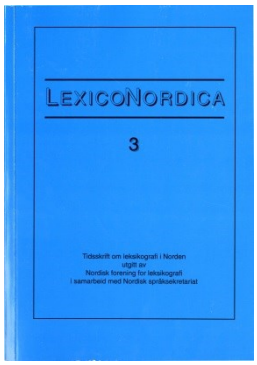
LexicoNordica

Titel:	Klassifikation af korpustekster, og kvantitative mål for sammensætningen af et almensprogligt korpus
Forfatter:	Ole Norling-Christensen
Kilde:	LexicoNordica 3, 1996, s. 121-129
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.

Ole Norling-Christensen

Klassifikation af korpustekster, og kvantitative mål for sammensætningen af et almensprogligt korpus

The paper gives an introduction to the design and composition of general language corpora, and the problem of statistical representativeness is considered. In order to determine, check and document their composition, a classificatory scheme for text types is needed. The building of a corpus is seen as an iterative process: The original design may set up overall quantitative measures for a few easily distinguishable and rather general text types; after the collection of some texts or text samples, which should be annotated according to a more fine-grained classification scheme, the distribution by more specific classes (e.g. topic under a general class of non-fiction) is measured, and the design criteria are adjusted accordingly. The design of the corpus of The Danish Dictionary and of the corpus of Danish which constitutes part of the European PAROLE project, are used as examples.

Indledning

Det største og bredest sammensatte danske korpus er opbygget til brug for Den Danske Ordbog (DDO). Det omfatter i alt 40 millioner tekstord fordelt på henved 44.000 forskellige tekstprøver fra perioden 1983–92, herunder samtlige tekster (fire millioner tekstord) fra Henning Bergenholtz' DK87–90. Erfaringerne fra opbygningen af DDO's korpus, og også en del af dets tekster, indgår i det europæiske korpus- og ordbogsprojekt PAROLE, som er en fortsættelse af NERC og desuden bygger på bl.a. standardiseringsforslag fra EAGLES.

De tre nævnte akronymer er navne på EU-støttede projekter inden for sprogforskning og sprogteknologi. **NERC** (**N**etwork of **E**uropean **R**eference **C**orpora) søgte ca. 1991–93 at sammenfatte den daværende viden om korpora og korpusarbejde med henblik på at rådgive EU (Generaldirektorat XIII i Luxembourg). Hovedresultaterne er samlet i NERC (1995), som desuden rummer en detaljeret plan for et europæisk korpussamarbejde. Inden for **EAGLES** (**E**uropean **A**dvisory **G**roup on **L**anguage **E**ngineering **S**tandards) har en række ekspertgrupper 1993–96 udarbejdet forslag til fælleseuropæiske standarder for opbygning og dokumentation af korpora og for den leksikografiske beskrivelse af EU-sprogene; de fleste af dem kræver dog yderligere afpudsning, og en

række delområder mangler endnu at blive behandlet. Første fase af **PAROLE** (**P**reparatory **A**ction for Linguistic **R**esources **O**rganization for **L**anguage **E**ngineering) måtte derfor bl.a. omfatte udarbejdelsen af en fælles norm for sammensætning og opbygning af de europæiske korpora (Norling-Christensen 1996), samt en norm for deres tekniske udformning (Ridings 1996). Under anden fase (1996–98) opbygger de enkelte sproggrupper (14 vesteuropæiske sprog) nu offentligt tilgængelige korpora, som følger disse normer. Endvidere konstrueres for hvert sprog et **leksikon**, dvs. en ordbog der kan anvendes af datamaskiner.

I det følgende gives en introduktion til planlægningen og opbygningen af almensproglige korpora med eksempler fra DDO's korpus, det danske PAROLE-korpus og et subkorpus heraf.

Definitioner

Med udgangspunkt i en klassifikation opstillet af Bernard Quémada skelner Atkins/Clear/Ostler (1992:1) mellem fire slags (elektroniske) tekstsamlinger: Et **arkiv** er et lager af elektroniske tekster uden særlig sammenhæng og ofte i forskellige tekniske formater, fx hidrørende fra forskellige tekstbehandlingssystemer. Et **elektronisk tekstbibliotek** (engelsk: electronic text library, ETL; fransk: textothèque) er en samling af elektroniske tekster i standardiseret format; sammensætningen af biblioteket kan følge visse retningslinier med hensyn til fx indhold, men uden at der er lagt strenge begrænsninger på udvælgelsen. Et **korpus** er da en delmængde af et bibliotek, udvalgt efter eksplicitte kriterier og med et bestemt formål for øje. Endelig forstås ved **sub-korpus** en delmængde af korpuset, fx fremkommet som mellemresultat under elektronisk korpusanalyse.

Denne præcisering af korpusbegrebet, som jeg vil søge at fastholde i det følgende, synes i dag at være almindeligt accepteret i fagkredse; men endnu i NERC (1995:63) tales der om

> a multifunctional corpus.. [which] .. should contain full texts rather than samples .. [and] .. should be extensively documented, so as to facilitate the selection of specific subcorpora, which can be expanded in order to construct large specialized or task-oriented corpora,

altså snarere et tekstbibliotek. Også Atkins/Clear/Ostler (l.c.) vælger i øvrigt, efter at have givet de nævnte definitioner, for kortheds skyld at

bruge *corpus* som et fælles overbegreb for bibliotek, korpus og subkorpus.

Som eksempler på de "bestemte formål", der har været retningsgivende for sammensætningen af nogle danske korpora, kan nævnes, at Maegaard og Ruus opbyggede (jf. fx Ruus 1995,1:19) DANwORD med henblik på undersøgelse af ordhyppigheder, DK87-90 skulle især tjene som en almensproglig reference for fagsproglige korpusstudier, mens DDO's korpus er beregnet til ordbogsredigering.

Repræsentativitet

Statistiske undersøgelser af et repræsentativt udsnit af befolkningen, eller af en population som statistikerne siger, kan give meget præcise oplysninger om populationen som helhed. Det gælder, hvad enten populationen er Danmarks befolkning på et givet tidspunkt, ti årgange af et bestemt dagblad, jf. Manual (1994:95), eller fx den del af et uddødt sprog som er overleveret. Stikprøveteori er den del af den statistiske teori, som omhandler metoder til udvælgelse og analyse af stikprøver fra en population, der består af et endeligt og kendt antal enheder. Enhederne skal i princippet kunne opregnes på en liste (som for den danske befolkning fx kan være folkeregisteret), fra hvilken man efter visse regler kan udvælge en mindre del til nærmere undersøgelse. Hvis enheden ikke kan defineres eksakt, eller hvis populationens størrelse ikke kendes, kan stikprøveteori ikke give eksakt vejledning, men højst tjene som en kvalitativ rettesnor. Korpusbyggeren må se i øjnene at korpus ikke bliver repræsentativt i statistisk forstand; i stedet må han stræbe efter at korpus, med mindre præcise termer, bliver *eksemplarisk* (Manual l.c.) eller *(af)balanceret* (Atkins/Clear/Ostler 1992:6) med hensyn til det sproglige materiale det skal være en model af. Dog søger Biber (1993) at fastholde ideen om statistisk baseret repræsentativitet, jf. kapitlet "Udvælgelse af tekstprøver" nedenfor.

Det almensproglige korpus

Formålet med DDO's korpus er at det skal udgøre en væsentlig del af det empiriske grundlag for en *deskriptiv ordbog over almensproget*, og det egentlige problem er i første omgang ikke af statistisk natur, men derimod at fastslå, hvad der skal forstås ved almensprog. Først når der foreligger en operational definition af det sprog der skal beskrives, vil statistisk teori eventuelt kunne vejlede med hensyn til hensigtsmæssig udvælgelse af stikprøver.

Skulle ordbogen have dækket et nøjere afgrænset område af sproget, fx ét forfatterskab eller ét bestemt fagområdes sprog, ville problemet om korpus' sammensætning være noget enklere. Ved enkelt-forfatterskaber kan man overveje at medtage alle tekster fremfor stikprøver. Ved fagsprog, hvor en vis normering ofte vil være ønskelig, kan fagfolk hjælpe med til at udpege fagets centrale og typiske tekster, og man kan supplere med lærebøger el.l. med gode registre. Havde målet endelig været en normativ ordbog over (fx) det danske almensprog, ville man vel have nedsat en komité af "gode danske mænd" til at udvælge "gode danske tekster".

Måske afgrænses almensproget bedst ved udelukkelsesmetoden. DDO valgte at udelukke det sprog, som fagfolk bruger til at kommunikere med andre fagfolk om deres fælles fag. Sproget i videnskabelige afhandlinger, eller i mekanikerens reparations-manualer, hører således efter vor opfattelse ikke til i en almensproglig ordbog. De danske fagsproglige korpora karakteriserede deres tekster ved en simpel sender modtager-model:

lægmand –> lægmand
lægmand –> fagmand
fagmand –> lægmand
fagmand –> fagmand

Det er den sidstnævnte kategori, fagsprog i snæver forstand, som vi har udelukket.

Modellen er nok for simpel. Manual (1994:16) indfører endnu en rolle: "halvfagmanden" (the semi-expert), som fx kan være under uddannelse i faget, være fagmand inden for en beslægtet disciplin, eller være popularisator ([fag]jounalist, marketingmedarbejder). Anlægger vi et receptionssynspunkt, ser på, hvilket sprog folk udsættes for (når de ikke er beskæftiget med deres fag), indtager halvfagmandens sprog en helt central rolle. Det skrives og tales af de journalister og andre formidlere, der i presse, radio og tv formidler politik, økonomi, økologi, sport, musik, bilteknik, osv. Sådanne tekster er i DDO's terminologi "(alment) fagsprog", og det hører klart med til grundlaget for en almensproglig ordbog.

Hermed har vi opridset konturerne af et *almensprogligt* ***korpus***; det kan være en samling af prøver på så mange og så forskellige tekster som muligt; de må ikke være fag-interne og de skal være rettet mod og/eller produceret af brede dele af befolkningen. Men vi har ikke hermed sagt, hvad *almensprog* er. Man kunne definere det snævert, som den del af sproget alle danskere (/svenskere etc.) er fælles om – den

kerne af ord, som vi alle forstår. Ruus (1995) anviser i sin disputats en korpusbaseret metode til at bestemme det centrale ordforråd, anvender den på sproget dansk, og konstruerer en ordbog over 1117 kerneord i form af et semantisk netværk. Dette er spændende lingvistisk grundforskning; men selv hvis denne ordbog blev omsat til en form, som gjorde det muligt for den almindelige ordbogsbruger at slå op i den, ville brugeren nok blive skuffet. For man slår jo ikke op for at finde ud af det, alle ved i forvejen, men det, man kan være i tvivl om. Sat på spidsen, kunne man sige, at når vi i vores ordbøger definerer det velkendte, det indlysende, så gør vi det mere for vor egen skyld end for brugerens. Dog bør det tilføjes, at ikke al sproglig kompetence er bevidst; én af ordbogens opgaver er at minde om det, man egentlig godt vidste.

Kerneordene er, groft sagt, ord der forekommer meget hyppigt i mange forskellige slags tekster; de vil derfor også optræde hyppigt i det almensproglige korpus. Men derudover vil de mange forskellige tekstarter, omhandlende mange forskellige emner, bidrage med *deres* specifikt hyppige ord. Den bredt anvendelige almensproglige ordbog får dermed ikke kun 1117 opslagsord, men snarere fem eller ti gange så mange, og dermed måske 50.000 eller flere, som ikke hører til kerneordene. De er tværtimod i ordbogen, fordi mange møder dem, men ikke alle er sikre på hvad de betyder eller hvordan de bruges.

Det sidste, hvordan de bruges, er vigtige oplysninger i en ordbog, men er ikke med i "Kerneordene". Noget af det, statistik på et godt korpus kan belyse, hvis korpus er stort nok, er syntaktiske mønstre (valens), samt ordkombinatorikken, altså hvilke ord der typisk forekommer sammen. *Blussende rød* og *strålende hvid* er godt og typisk dansk; *blussende gul* og *strålende sort* er syntaktisk korrekt, men ganske umuligt. Kontaminerede idiomer kan vise sig at være at være hyppigere end de oprindelige: *Gammel rotte* og *garvet politimand* er godt gammelt dansk. Men det meningsløse *garvet rotte* udviser overraskende højt forekomsttal, og må registreres i ordbogen. Sådanne forhold afspejles i korpus og kan eftervises med statistiske metoder, hvis korpus er tilstrækkelig omfattende.

Udvælgelse af tekstprøver

Stikprøveteorien skelner mellem simpel tilfældig udvælgelse og stratificeret udvælgelse.

Ved **simpel tilfældig udvælgelse** har alle individer samme sandsynlighed for at blive udvalgt; sandsynligheden for, at en bestemt slags

individ bliver udvalgt, er simpelthen antallet af denne slags individer divideret med det totale antal individer i populationen. For udvælgelsen af tekstprøver til et korpus rejser denne definition straks tre problemer: Hvilke individer er der tale om (ord, sætninger, afsnit, hele tekster)? – Hvad er det totale antal individer i populationen, som jo er hele det sprog der ønskes en model af? – Og er samme sandsynlighed for ethvert individ overhovedet ønskelig?

Biber (1993:247) anslår, at et korpus, der skulle være repræsentativt i demografisk forstand, altså rumme de sproggenrer som folk typisk bruger i samme forhold som de bruger dem, ville bestå af ca. 90% samtale og 3% breve og notater, mens de resterende 7% måtte deles af alle øvrige genrer,

> such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing. (Very few people *ever* produce published written texts, or unpublished written and spoken text for a large audience). Such a corpus would permit summary descriptive statistics for the entire language represented by the corpus. These kinds of generalizations, however, are typically not of interest for linguistic research.

I stedet for den demografiske repræsentativitet foreslår han derfor, at korpus sammensættes ved stratificeret udvælgelse på en sådan måde, at det bedst muligt belyser den sproglige variation, og han angiver statistiske metoder til at måle om givne sproglige træk er tilstrækkeligt repræsenteret; de vil fx vise, at 90% samtale i et stort korpus er langt mere end nødvendigt for at give et pålideligt billede af denne tekstart.

Ved **stratificeret udvælgelse** inddeles populationen i forskellige lag, **strata**, iht. kriterier, som på forhånd er kendt. Den tilfældige udvælgelse foretages separat for hvert stratum med henblik på, at man efterfølgende skal kunne analysere stikprøverne separat for hvert stratum. Ved befolkningsundersøgelser kan kriterierne fx være alder og køn, for korpusopbygning er fx genre eller teksttype relevante kriterier, og man vil stræbe efter tilstrækkelig repræsentation af hver type og altså tildele sjældent forekommende teksttyper forholdsvis større vægt end de hyppige. Yderligere foreslår Biber, at korpus opbygges ad flere omgange, så udvælgelseskriterierne kan justeres undervejs. Jeg kan ikke her gå yderlige ind på Bibers argumentation eller på hans skelnen mellem klassifikation/stratificering efter henholdsvis tekst-eksterne kriterier ("genre" eller "register") og tekst-interne kriterier ("teksttype"), men må henvise interesserede læsere til den pågældende artikel.

Den Danske Ordbogs korpus

Uden at kende Bibers arbejder fulgte vi faktisk i stor udstrækning hans anbefalinger, da vi i 1991 planlagde DDO's korpus. Dog blev størrelsen af den enkelte tekstprøve, det ovenfor omtalte individ, ikke nøjere præciseret; kortere tekster er medtaget i deres helhed, mens der fra længere skrevne tekter, typisk hele bøger, er udvalgt et eller nogle få kapitler, i alt højst 10.000 tekstord. På basis af dikotomierne skriftsprog / talesprog, reception ("offentligt sprog") / produktion ("privat sprog"), og almensprog / (alment) fagsprog, opdeltes korpus i otte strata, og for hvert fastsattes skønsmæssige mål for deres andel af de 40 millioner tekstord. Hver tekstprøve blev annoteret med oplysning om stratum, men desuden en række andre oplysninger, bl.a. om medie, genre, emne og produktions- eller trykår, samt om sprogbrugerens alder, køn m.v. Hermed var indledt en iterativ proces, hvor den første udvælgelse skete efter ret grove kriterier, de otte strata, hvorefter teksternes fordeling med hensyn til de øvrige kriterier løbende kunne kontrolleres og den videre tekstakkvisition justeres derefter. Denne metode bidrog fx væsentligt til, at mange forskellige emner blev repræsenteret blandt fagteksterne. En fyldig beskrivelse af korpus, arbejdet med at opbygge det, og hvad det bruges til, gives i Norling-Christensen & Asmussen (i trykken); en fuldstændig oversigt over klassifikationen efter medie, genre og emne findes i Norling-Christensen (1996).

Parole-korporaene

De korpora, der produceres inden for PAROLE-projektet skal kun dække skriftsprog, er primært til brug inden for sprogteknologi, og skal være offentligt tilgængelige. Allerede disse krav udelukker en hel del af DDO-materialet fra at blive genbrugt, først og fremmest talesproget; men også visse andre teksttyper må anses for mindre relevante i denne sammenhæng. Desuden udelukkes tekster, som af forlag m.fl. er stillet til rådighed med den klausul, at de alene må anvendes til ordbogsarbejdet, samt andre tekster som af ophavsretsmæssige grunde ikke kan mangfoldiggøres og distribueres.

Budgettet for det enkelte sprog er begrænset, og bl.a. af denne grund er de fælles krav til korpussammensætning ret løse. Kun for fordelingen på medier (hvor der bl.a. sættes en overgrænse for andelen af avistekster) og for perioden (højst 20% fra 1970'erne, og intet ældre) er der fastsat bindende retningslinier.

Mindst 250.000 ord skal være grammatisk annoteret (såkaldt **tagging**), minimalt med ordklasse; taggingen skal ikke blot være maskinel, men kontrolleret af mennesker, så den sidenhen kan bruges til træning og/eller afprøvning af taggere, dvs. programmel der automatisk tilordner ordklasse og eventuelt andre morfosyntaktiske oplysninger til tekstord. Til brug for opbygningen af det regelsæt, som skal anvendes ved automatisk tagging af det danske PAROLE-korpus, er der ved stratificeret udvælgelse udtrukket et subkorpus på 100.000 tekstord, som er ved at blive manuelt tagget. De er fordelt på ca. 700 tekstprøver, taget fra hver sin tekst; en prøve består af hele afsnit op til et maksimum af 160 tekstord. Dette design forventes at sikre maksimal spredning mht. sådanne sproglige træk som er af betydning for automatisk tagging; og som en sidegevinst opnås formentlig, at de 100.000 ord kan betragtes som en citatsamling og distribueres uden ophavsretlige problemer.

Afslutning

Som det især er fremgået af omtalen af PAROLE-korporaene, kan korpusbyggeren ikke nøjes med teoretiske overvejelse af lingvistisk, statistisk og sociologisk art. Også praktiske og økonomiske overvejelser hører med: Hvilke tekster kan overhovedet fremskaffes? Hvad koster det at scanne eller indtaste dem, hvis de ikke allerede foreligger i et egnet elektronisk format? Hvor megen annotation er der resurser til at udføre? Hvad med ophavsretten? – Til en lang række formål vil mange korte tekstprøver være at foretrække frem for fulde tekster. Men med fulde tekster (hele bøger) når man langt hurtigere, nemmere og billigere op på den ønskede korpusstørrelse. Det gælder her, som i næsten alle forhold, at det praktisk mulige må afvejes mod det ideelt øns-kelige, og at kompromiser ikke kan undgås. Lad det være en trøst, at

> In our ten years' experience of analysing corpus material for lexicographical purposes, we have found any corpus – however 'unbalanced' – to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts. It would be shortsighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as 'unreliable' or 'irrelevant' simple because the corpus used cannot be proved to be 'balanced'. (Atkins/Clear/Ostler 1992:6).

Litteratur

Atkins, Sue / Jeremy Clear / Nicholas Ostler 1992: Corpus Design Criteria. In *Literary and Linguistic Computing* 7.1, 1–16.

Biber, Douglas 1993: Representativeness in Corpus Design. In *Literary and Linguistic Computing* 8.4, 243–257.

Manual 1994 = *Manual i Fagleksikografi*. Red. Henning Bergenholtz og Sven Tarp. Herning: Systime.

NERC 1995 = Calzolari, N. / M. Baker / P.G. Kruyt (eds) 1995: *Towards a Network of European Reference Corpora.* Report of the NERC Consortium, feasibility study, coordinated by A. Zampolli. Pisa: Giardini.

Norling-Christensen, Ole (ed.) 1996: *Design and Composition of Reusable Harmonized Written Language Reference Corpora for European Languages*. PAROLE (MLAP: 63-386) WP 4 – Task 1.1. København: Det Danske Sprog- og Litteraturselskab.

Norling-Christensen, Ole / Jørg Asmussen, i trykken: The DDO Corpus of Contemporary Danish. In *Computers and the Humanities* (forthcoming).

Ridings, Daniel 1996: *Text Representation in PAROLE*. Göteborg.

Ruus, Hanne 1995: *Danske Kerneord. Centrale dele af den danske leksikalske norm 1–2*. København: Museum Tusculanums Forlag.