NyS

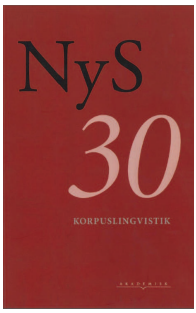
Titel: Korpus 2000. *Et overblik over projektets baggrund, fremgangsmåder og perspektiver*

Forfatter: Jørg Asmussen

Kilde: *NyS – Nydanske Sprogstudier* 30. *Korpuslingvistik,* 2002, s. 27-38

Udgivet af: Akademisk Forlag A/S

URL: www.nys.dk



Søgbarhed

Artiklerne i de ældre NyS-numre (NyS 1-36) 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.

Korpus 2000

Et overblik over projektets baggrund, fremgangsmåder og perspektiver

JØRG ASMUSSEN

INDLEDNING – DE YDRE RAMMER FOR KORPUS 2000

Korpus 2000 er navnet på et projekt hos Det Danske Sprog- og Litteraturselskab, DSL, og blev igangsat i april 2000. Formålet med projektet er at opbygge et tekstkorpus, der skal afspejle alment dansk skriftsprog fra årene omkring år 2000, nærmere betegnet 1998-2002. Dette korpus, der ligesom projektet kommer til at hedde Korpus 2000, skal være rimelig stort: det skal bestå af mindst 20 millioner løbende tekstord fordelt på mindst 20.000 tekstprøver. Og det skal være offentligt tilgængeligt forstået på den måde, at alle skal kunne udføre sproglige undersøgelser på dette korpus, både over internettet og vha. et cd-rom-baseret system.

Projektet støttes med 2,8 mio. kr. af *År 2000 Fonden*, "som skal fremme folkelige initiativer og debat, der viser veje til en fortsat positiv udvikling af det danske samfund og demokrati" – som det hed på dens i mellemtiden nedlagte hjemmeside. Fonden administreres af Forskningsministeriet. En væsentlig betingelse for Fondens bevilling var, at projektet formidles til en så bred offentlighed som muligt, at det altså ikke blot kommer en snæver kreds af fagfolk til gode.

Dette er de ydre, faste rammer for projektet.[1] I det følgende vil jeg dels gøre rede for baggrunden for projektet og dels for, hvordan man på projektet søger at udfylde de skitserede rammer, så Korpus 2000 – både projektet og produktet – kan være til nytte for så mange som muligt. De formidlingsmæssige aspekter ved projektet skal jeg ikke komme ind på her. Ledetråden i redegørelsen vil være de følgende fire spørgsmål:

1. Hvad er baggrunden for Korpus 2000?
2. Efter hvilke principper opbygges Korpus 2000?

3. Hvad forstås ved offentlig tilgængelighed?
4. Hvad skal Korpus 2000 bruges til? – og hvordan?

1. HVAD ER BAGGRUNDEN FOR KORPUS 2000?

Jeg skal under behandlingen af spørgsmål 2 komme ind på, hvad jeg forstår ved begrebet *korpus*; lige nu vil jeg nøjes med den lidt løse definition *større tekstsamling i elektronisk form*. Prøver man at skaffe sig et overblik over, hvilke større tekstsamlinger over nyere dansk sprog der findes, tegner der sig sandsynligvis et ufuldstændigt billede. Korpus 2000 har prøvet at danne sig sådan et overblik gennem en officiel henvendelse til samtlige ca. 50 institutioner her i landet, som enten selv beskæftiger sig med sprogforskning eller må formodes at have en interesse heri, ikke mindst på et så omfattende materialegrundlag, som et tekstkorpus kan udgøre. Vi skrev i denne henvendelse: "(vi vil) opfordre jer til at kontakte os, hvis I skulle være bekendt med danske korpora, f.eks. af mindre omfang, af særlige teksttyper eller om særlige emner"[2]. Der indløb to tilkendegivelser: én vedrørende VISL's Korpus og én vedrørende Dansk Talesprogskorpus, BySoc. To korpora i øvrigt, som vi i forvejen var bekendt med eksisterede. Så vidt jeg er orienteret, findes der således følgende rimeligt tilgængelige danske korpora:

- DANwORD
- DK 87-90
- Den Danske Ordbogs Korpus
- Parole Dansk Talesprogskorpus
- VISL's Korpus

Det er tænkeligt, at der herudover findes nogle fagsprogsorienterede tekstsamlinger i elektronisk form, evt. findes der også samlinger med oversat sprog samt originaltekster. Forlagsverdenen råder velsagtens også over et eller andet mindre lettilgængeligt. Endvidere findes der formodentlig en række ad hoc-samlinger, som typisk mere eller mindre legalt kan være tappet fra internettet eller fra andre (elektroniske) kilder. Fælles for disse formodede samlinger er, at de på grund af deres utilgængelighed ikke kan formodes at bidrage væsentligt til lingvistisk forskning.

DANwORD består af 1,25 millioner løbende tekstord i 1000 tekstprøver på hver 250 ord, som er hentet fra fem forskellige medietyper hhv. tekstgenrer fra perioden 1970-74. Der er her tale om et pionerarbejde inden for opbygningen af dansksprogede tekstkorpora, men det er desværre aldrig rigtigt blevet ført videre. En nærmere beskrivelse af dette korpus' opbygning samt flere referencer findes fx i Ruus 1995.

DK 87-90 består af af 4 millioner løbende tekstord, 1 million for hvert af årene 1987-90. Hver årgang består af 500 tekstprøver på hver 2000 ord fordelt på medierne bøger, blade og aviser. DK87-90 er dermed stort set opbygget efter de samme principper som det amerikanske Brown Corpus fra tresserne og det tyske LIMAS-Korpus fra halvfjerdserne. Til forskel fra de to forbilleder inddrager DK 87-90 konsekvent den diakrone dimension. Heller ikke DK 87-90 blev umiddelbart ført videre, men slutter netop i året 1990. Siden 1993 administreres DK 87-90 af DSL, som har suppleret tekstprøverne med yderligere oplysninger om blandt andet forfatterens alder og køn, tekstens emne og genre. DK 87-90 indgår desuden som en del i Den Danske Ordbogs Korpus. DSL har siden 1993 udlånt DK 87-90 på disketter eller cd-rom til interesserede og under visse betingelser. Dette korpus kan som noget nyt nu også downloades fra DSL's hjemmeside, dog ikke i form af læsbare tekstdokumenter (som på disketterne og cd-rom'en), men som binært korpusmodul til konkordanssystemet Semaskop, som vi i forbindelse med Korpus 2000-projektet har under udvikling, og som ligeledes kan downloades i en testversion fra DSL's hjemmeside.

Den Danske Ordbogs Korpus (DDO's Korpus) består af 40 millioner løbende tekstord fra perioden 1983-92 fordelt på godt 43.000 tekstprøver. Hver tekstprøve er udstyret med en række nøje fastlagte oplysninger om selve tekstprøven og dens forfatter. En detaljeret beskrivelse af disse oplysninger og af opbygningen af DDO's Korpus i det hele taget finder man i Norling-Christensen&Asmussen 1998. DDO's Korpus blev til som led i DSL-projektet *Den Danske Ordbog* og blev færdig i 1993. Det har siden tjent som primær kilde til DDO og en særlig version af det på 36 mio. lbd. ord, som ikke indeholder tekstprøver, der af tekstleverandørerne er belagt med brugsrestriktioner, er blevet brugt af adskillige, både forskere, studerende og andre[3]. Ulempen for andre brugere end redaktørerne på DDO har indtil for nylig været, at dette korpus af især tekniske årsager kun har været tilgængeligt hos DSL, hvor man dog uden større proble-

mer altid har kunnet udføre de undersøgelser, man ønskede. I mellemtiden kan DDO's Korpus dog downloades fra DSL's hjemmeside som korpusmodul til konkordanssystemet Semaskop.

PAROLE-Korpus består af 250.000 løbende tekstord fordelt på 1500 tekstprøver. Hvert enkelt ord i dette korpus er morfosyntaktisk annoteret, altså forsynet med oplysninger om både ordklasse og bøjning samt grundform. Den morfosyntaktiske opmærkning af dette korpus er kontrolleret og rettet til, så det bl.a. kan tjene som udgangspunkt for træningen af automatiske taggere, men ikke mindst også kan tjene til udviklingen af annotationssensitive konkordans- og statistikværktøjer. PAROLE-Korpuset blev udarbejdet hos DSL under det EU-støttede projekt PAROLE (Preparatory Action for linguistic Resources Organization for Language Engineering) i perioden 1996-1998. Selvom projektet er slutevalueret og godkendt af EU-Kommissionen, er det på mange måder ufuldstændigt. For den danske del af projektet mangler der først og fremmest et distribuerbart – dog ikke morfosyntaktisk annoteret – korpus på 2,75 millioner løbende tekstord. PAROLE-Korpus kan downloades som SGML-dokument fra DSL's hjemmeside sammen med en udførlig dokumentation (Keson 1998). Det er desuden aftalt med ELDA (European Language resources Distribution Agency), at de ligeledes vil tage sig af distrubutionen.

Dansk Talesprogskorpus – eller *BySoc* – er på i alt 1,3 millioner løbende tekstord og består af transskriptioner af ca. 80 samtaler. Det blev indsamlet af projektet *Bysociolingvistik* omkring 1987. Korpuset er offentligt tilgængeligt over internettet: man kan få opstillet konkordanser efter forskellige kriterier[4]. Dele heraf indgår desuden i DDO's Korpus, dog ikke i den alment tilgængelige version.

VISL-Korpus består i øjeblikket af 10 millioner løbende tekstord. Korpuset er tilgængeligt på internettet: man kan her få opstillet konkordanser efter forskellige kriterier[5]. Der er ingen umiddelbart tilgængelige eksplicitte oplysninger om sammensætningen af dette korpus, og der er heller ikke knyttet oplysninger til de enkelte brugte teksteksempler i dette korpus. Korpuset synes hovedsagelig at bestå af materiale fra ganske få (ca. 2) nyhedsmedier. Dele af dette korpus er morfosyntaktisk taggede. VISL står for *Visual Interactive Syntax Learning* – et projekt ved Syddansk Universitet i Odense.

Af det nævnte fremgår det, at DSL turde være ganske markant repræsenteret i det mindste i den synlige del af det danske korpuslandskab. Dette

må især føres tilbage til DDO's Korpus, som både med hensyn til den meget varierede og tilstræbt balancerede sammensætning (både tale- og skriftsprog, både almensprog og alment fagsprog, både produktions- og receptionssprog, mange medier, mange genrer repræsenteret), med hensyn til annotationsgraden på tekstniveau og med hensyn til omfanget stadig er enestående blandt de danskprogede korpora. Opbygningen af Den Danske Ordbogs Korpus har givet DSL mange erfaringer med opbygningen af store korpora, ikke blot hvad angår de principielle konceptuelle overvejelser, der ligger til grund for sådan et projekt, men især med hensyn til den praktiske afvikling af arbejdet; herunder ikke mindst selve tekstakvisitionen. Hertil kom hyppige henvendelser til DSL vedr. adgang til Den Danske Ordbogs Korpus eller vedr. mulige – nyere – alternativer. På denne baggrund anså vi det hos DSL som både nærliggende og tiltrængt at videreføre korpusaktiviteterne ved at igangsætte et nyt korpusprojekt, nemlig Korpus 2000.

2. EFTER HVILKE PRINCIPPER OPBYGGES KORPUS 2000?

Inden jeg kommer til selve principperne, skal det afklares, hvad jeg her vil forstå ved et korpus. Indtil nu har jeg blot anvendt betegnelsen *korpus*, som om der herskede bred enighed om, hvad begrebet egentlig dækker over. Men det gør der næppe: Er fx en ordbog eller ordseddelsamling et korpus? En stak aviser? En radioudsendelse, optaget på bånd? Det bør afklares, hvilke kvalitative kriterier der egentlig konstituerer et korpus, og det skal desuden fastlægges, hvilke kvantitative mål der skal gælde, før man kan tale om et korpus.

John Sinclair 1996 påpeger, at elektronisk lagrede korpora, altså korpora i digital form, udgør et ret nyt fænomen, og at der derfor endnu ikke har udkrystalliseret sig en konsensus om, hvad der egentlig er et korpus – og om hvordan man i det hele taget kan klassificere korpora. Han opstiller derfor i sin rapport en definitions- og klassifikationsramme, som ikke skal diskuteres her, men som den følgende definition vil være delvis inspireret af. Et korpus betragtes herefter som en meget stor digitaliseret samling af prøver på skrevet eller nedskrevet løbende autentisk tekst, der med hensyn til forskellige teksttypologiske kriterier er struktureret efter et eksplicit princip med henblik på at muliggøre bestemte sprogvidenskabelige undersøgelser. *Meget stor* betyder, at samlingen kun

kan overskues vha. edb-baserede værktøjer. *Meget stor* betyder også så stor, at de sprogvidenskabelige undersøgelser, der kan udføres på baggrund af korpuset, vil kunne antage empirisk karakter – og derfor bl.a. vil kunne gentages med samme resultater på et andet korpus af tilsvarende teksttypologisk beskaffenhed[6]. På baggrund af denne definition er hverken ordbøger, seddelsamlinger, stakke af aviser eller båndoptagede radioudsendelser korpora.

Definitionen indebærer, at et korpus ikke kan betragtes uafhængigt af det lingvistiske formål, det siden skal bruges til. Det er fx kun vanskeligt muligt at undersøge teksttypen *erhvervskorrespondance* på baggrund af et korpus, der udelukkende indeholder avisartikler. En leksikograf på jagt efter denotationer er ikke nødvendigvis tjent med et korpus bestående af deiktisk talesprog. En ordklasseopmærkning ville derimod måske være ønskelig for leksikografen. I modsætning måske til sociolingvisten, der både kan være interesseret i at have talesprog – produktionssprog i det hele taget – og oplysninger om sprogbrugeren; oplysninger, som måske igen er ganske irrelevante for den, der vil undersøge syntaksen i avisernes nyhedsartikler, og som måske ville have gavn af en eller anden form for syntaktisk opmærkning. Spørgsmålet er, om et korpus med universel anvendelighed overhovedet er et realistisk mål. Efter vores opfattelse er det det ikke. Alligevel ville det være det mest rationelle, om man kunne opnå en vis grad af generalitet, når nu der skal udarbejdes et nyt korpus, således at vores, dvs. Korpus 2000-projektets, bestræbelser kommer så mange til gode som muligt – ikke mindst også med henblik på opbygningen af fremtidige korpora.

For at opnå denne generalitet har vi besluttet ikke at lade os binde af den konkrete korpusopgave, som Korpus 2000-projektet udgør, men at tilstræbe en fremgangsmåde, der kan genbruges i forbindelse med udarbejdelsen af andre korpora eller måske alternative Korpus 2000-korpora. Fælles for samtlige hidtidige danske korpusprojekter var netop, at man kun havde det planlagte korpus for øje, allerede under selve tekstindsamlingsfasen: da Den Danske Ordbogs Korpus blev opbygget, modtog man typisk tekster på diskette, valgte herfra de tekstprøver, der skulle bruges i korpus, fx ti sider fra en roman, udstyrede dem med de vedtagne tekst- og sprogbrugeroplysninger og konverterede det hele så, tekst og oplysninger, til det format, som er gældende for dette korpus. Mange af disse gamle disketter med råtekster til DDO's korpus eksisterer ganske

vist den dag i dag – mange af dem dog i forældede fysiske formater – og de indeholder dermed råstof til adskillige potentielle korpora. Men der er ikke knyttet tekstuelle oplysninger til dem, for disse blev jo kun klistret på den prøve, den udvalgte del, der nu befinder sig i DDO's korpus: disse forhold gør det yderst resursekrævende, at udnytte det korpus-råstof, der jo faktisk ligger på disse gamle medier, og som ikke blev brugt i DDO's korpus.

Det særlige led, som vi så indskyder for at generalisere korpusopbygningen, er en database, som kommer til at indeholde de tekster, vi får, samt oplysninger om dem. Databasen er tilrettelagt på en sådan måde, at den principielt kan optage tekster af en hvilken som helst art, dvs. også tekster, der ikke umiddelbart er relevante for selve Korpus 2000. Det viser sig nemlig, at mange af de tekster eller teksttyper, som adskillige tekstleverandører tilbyder os, ikke umiddelbart er relevante for Korpus 2000, som oftest fordi vi har rigeligt af denne type i forvejen; men teksterne kan meget vel blive relevante i en anden sammenhæng, i et andet korpus. Som nævnt skal Korpus 2000 mindst indeholde 20 millioner løbende tekstord, men ifølge et foreløbigt forsigtigt skøn råder vi allerede på nuværende tidspunkt over ca. 150 millioner ord. Derfor ville det være et alvorligt spild af resurser, hvis vi forkastede alle de tekster, vi ikke umiddelbart får brug for i forbindelse med selve produktet Korpus 2000. I stedet for at forkaste 'overflødige' tekster optager vi alt tekstmaterialet, som vi får tilbudt, i databasen. Hver enkelt tekst, som vi indlemmer i databasen, forsynes dernæst med systematiske oplysninger om selve teksten og dens afsender. Vi anvender i øvrigt en lettere modificeret version af de oplysningskategorier, som også bruges i DDO's Korpus[7]. De mulige værdier, der skal kunne knyttes til oplysningskategorierne, hvilke genrebetegnelser, emner osv., vi skal operere med, er under løbende udarbejdelse.

Vi kalder denne database bestående af tekster samt systematiserede tekstoplysninger for en *tekstbank*. Processen at indlemme en tekst i denne tekstbank og at forsyne den med de rette oplysninger, kalder vi tekstregistreringen. Denne proces skal der ikke redegøres for i detaljer her, men grundlæggende vil selve teksten blive opbevaret i to versioner i tekstbanken, et forholdsvis oprindeligt, typografi-orienteret format (RTF, som dtp- eller tekstbehandlingstekster konverteres til eller, for internetteksters vedkommende, HTML; herudover er ren tekst også tilladt som oprindeligt format dér, hvor typografien ikke har spillet nogen

rolle). Herudover vil teksterne foreligge i et forholdsvis simpelt struktureret SGML- eller XML-format, nemlig det samme, som PAROLE (uden morfosyntaktisk opmærkning) foreligger i. Alle andre former for opmærkning, først og fremmest af morfosyntaktisk art, vil først blive tilføjet, når der udtrækkes et konkret korpus fra tekstbanken, idet denne opmærkning i høj grad må antages at være bestemt af rekvirentens præferencer. For Korpus 2000 har vi indtil videre besluttet en morfosyntaktisk opmærkning, som vil svare til den, der kendes fra PAROLE-Korpus. Dog vil søgesystemerne, som udvikles sideløbende med Korpus 2000, være i stand til at afbilde PAROLE-taggingen på en simplificeret måde over for brugeren, hvis han ønsker det.

Fra tekstbanken kan der siden udtrækkes skræddersyede korpora efter forskellige rekvirenters forskellige ønsker. Den første rekvirent er os selv, og vi vil udtrække Korpus 2000, der – med DDO's Korpus som delvist forbillede – skal indeholde så mange forskellige teksttyper som muligt i et så afbalanceret forhold som muligt, primært fra årene 1999-2001. Men efter at selve Korpus 2000 er tilvejebragt, skal tekstbanken principielt kunne servicere en hvilken som helst rekvirent og kunne sammenstille korpora til vedkommende efter vedkommendes særlige specifikationer. Det kunne være korpora, som fx udelukkende består af avisartikler, eller af tekster vedrørende sundhed og helbred eller tekster forfattet af kvinder eller af unge – eller korpora efter alle mulige andre kriterier. Og jo mere materiale tekstbanken indeholder, jo mere specifikke ønsker kan efterkommes. Fremgangsmåden, først at etablere en tekstbank og sidenhen at udtrække det egentlige korpus herfra, har efter vores mening den styrke, at man under indsamlingen af tekster ikke tager stilling til, hvordan et senere korpus skal komme til at se ud, men overlader dette spørgsmål til en senere rekvirent.

Ved at anvende dette tekstbankprincip håber vi, at vores arbejde også vil være til gavn for andre, der måtte have en interesse i at opbygge eller få adgang til særlige korpora. Disse vil nemlig med fordel kunne benytte sig af den tekstbank, der nu etableres i forbindelse med Korpus 2000-projektet og måske selv bidrage med tekstmateriale til banken.

Det er vores tanke, at tekstbanken også efter afslutningen af projektet Korpus 2000 i sommeren 2002 fortsat bør udbygges og vedligeholdes, så den kan servicere alle korpusinteresserede. DSL er ved at undersøge mulighederne for et sådant fortsat arbejde.

3. HVAD FORSTÅS VED OFFENTLIG TILGÆNGELIGHED?

Det krav, at Korpus 2000 skal gøres offentligt tilgængeligt, indebærer en række ophavsretlige problemstillinger. Det turde være en selvfølge, at man med et korpusprojekt forsøger at etablere et arbejdsgrundlag – en materialesamling – for lingvister og andre sproginteresserede. Det er hverken hensigten at udgive et informations- eller tekstsøgesystem, ej heller at give korpusbrugerne adgang til fuldtekster. Alligevel skal teksterne jo opbevares i tekstbanken, som derfor er administrativt tilgængeligt for en gruppe af DSL's medarbejdere. Og det er da også med rette, at der blandt tekstleverandørerne er en del, der bekymret overvejer, om de fx virkelig vil have deres oprindeligt håndskrevne dagbogsoptegnelser deponeret i en sådan tekstbank. Derfor betragter vi det som en nødvendighed dels at oplyse tekstleverandørerne om, hvad vi gør med deres tekster, herunder oplyse dem om, hvor tilgængelige teksterne bliver for tredjepart. Først når leverandøren giver sin tilladelse til, at vi må bruge teksterne, bliver de lagt ind i tekstbanken – uanset om leverandøren er et forlag, en festsangdigter eller indehaveren af en hjemmeside. En egentlig distribution af heltekster fra vores tekstbank, fx som XML-formateret korpus, vil der derfor næppe kunne blive tale om. En meget stor del af vores tekstleverandører lægger overordentlig stor vægt på, at deres tekster ikke videregives til andre, at andre end ikke må få hele tekster at se. Men leverandørerne accepterer dog generelt, at mindre tekstpassager godt må vises. Adgang til Korpus 2000 vil derfor blive givet gennem et konkordans- og statistiksystem på internettet. Dette system etableres omkring CQP-kernen, som blev udviklet ved Institut für Maschinelle Sprachverarbeitung ved universitetet i Stuttgart[8], som vi – i delvis samarbejde med udviklerne fra IMS – udvikler et passende web-interface til. Desuden vil cd-rom-versionen af Korpus 2000 være indkapslet i det førnævnte konkordanssystem, som allerede nu kan downloades i en prototypeversion fra DSL's hjemmeside.

En anden, supplerende, løsning på tilgængelighedsproblemet er muligvis et *citatkorpus*, som ikke består af lange sammenhængende tekstpassager, men i stedet for af mindre enheder fra tekster; enheder, hvis længde ikke overstiger et rimeligt citats længde. Til disse tekster kan der stadigvæk knyttes tekstuelle attributter og langt de fleste lingvistiske undersøgelser vil sandsynligvis kunne udføres ligeså godt på et sådant ci-

tatkorpus, som på et korpus bestående af større, sammenhængende tekster eller tekstuddrag.

4. HVAD SKAL KORPUS 2000 BRUGES TIL – OG HVORDAN?

Da Korpus 2000's målgruppe ikke begrænser sig til fagfolk, men først og fremmest tænkes at bestå af lægfolk, vil der blive lagt vægt på, at konkordans- og statistiksystemet omkring selve korpuset dels vil kunne tjene til, at brugeren *hurtigt og forholdsvis simpelt* kan få be- eller afkræftet sine antagelser om sprogbrug – som denne tog sig ud omkring år 2000 og er afspejlet i Korpus 2000 -, og dels også vil kunne tjene til at gøre brugeren opmærksom på bestemte sprogbrugsfænomener. Derfor vil søgesystemerne omkring Korpus 2000 ikke blot give indblik i korpus i form af traditionelle KWIC-konkordanser, som det så er op til brugeren at fortolke, men de vil også i form af blandt andet delvis automatisk fortolkede konkordanser og statistisk isolerede sprogbrugsfænomer i hele korpus kunne give et indblik i, hvad der i korpus er hyppigt eller sjældent, typisk eller utypisk, påfaldende eller mindre påfaldende sprogbrug. Dette vil evt. blive kombineret med en diakron perspektivering på baggrund af materiale fra DDO's Korpus.

Alle undersøgelsesfaciliteter skal som nævnt være nemme at gå til – også for fagfolk. En skelnen mellem lægfolk og fagfolk i *anvendelsen* af korpora er sandsynligvis kun af underordnet betydning. Hvis man antager, at eksperten vil have fuldstændig frie hænder til at kunne udføre alle tænkelige undersøgelser, er det nødvendigt, at han sætter sig ind i en kompleks formalisme, hvori han generelt kan udtrykke, hvad han specifikt søger efter i korpus. Det må dog anses for overordentlig usandsynligt, at en lingvist, der er stødt på et bestemt sprogbrugsfænomen og nu gerne vil have flere autentiske eksempler på dette eller lignende fænomener, vil være villig til først at skulle tilegne sig en formalisme, hvori kan formuleres en søgeforespørgsel til korpus, bare for denne ene undersøgelses skyld. Derfor vil Korpus 2000 prioritere søgemetoder, der nedtoner de formelle søgetekniske aspekter så vidt muligt. Dette betyder muligvis, at der med hensyn til søgemulighederne vil forekomme begrænsninger i forhold til sådanne søgninger, som udføres på baggrund af en søgeformalisme; derimod vil der være langt færre praktiske barrierer at overvinde for brugeren, før han ser et resultat.

For at kunne udbrede brugen af korpora inden for lingvistisk forskning bør disse derfor være *lettilgængelige*, hvilket dels selvfølgelig vil sige, at brugeren nemt skal kunne få adgang til dem, men i lige så høj grad, at undersøgelser skal kunne udføres umiddelbart på dem. Lettilgængelighed kan opnås ved at standardisere undersøgelserne, ved at opstille en række veldefinerede undersøgelsestyper.

For at få et indtryk af, hvilke søgemuligheder brugerne ønsker, har vi allerede nu gjort en prototype af vores cd-rom-søgeværktøj Semaskop tilgængelig til download fra vores hjemmeside. I løbet af sommeren 2001 følger så en internetversion til afprøvning direkte på vores hjemmeside. Det er Korpus 2000's håb, at disse tiltag vil kunne afføde en dialog med de fremtidige brugere om, hvad de mener at kunne bruge et tekstkorpus til, og hvad de forventer at kunne søge og finde i det. En sådan prøvede vi også at lægge op til i vores allerede nævnte henvendelse til de sprogforskningsinteresserede institutioner i Danmark: (vi opfordrer) alle, der måtte have interesse i projektet, til at kontakte os og fortælle os nærmere om jeres særlige behov, skrev vi. De reaktioner vi fik herpå, kan endnu tælles på én hånd. Min fortolkning af dette er, at man er afventende interesseret, men man ved ikke rigtig, hvad man dog skal bruge et korpus til. Grundlæggende mangler der stadig en bestemmelse af, i hvilke sammenhænge og under hvilke betingelser korpora bliver til nyttige redskaber for blandt andet lingvistisk forskning. Det er Korpus 2000's håb at kunne bidrage til en sådan bestemmelse.

Jørg Asmussen
Det Danske Sprog og Litteraturselskab,
email: ja@dsl.dk

NOTER

1. den officielle projektbeskrivelse for Korpus 2000 findes under http://korpus.dsl.dk/korpus2000/beskrivelse.html
2. en fortegnelse over de institutioner med relation til sprogforskning, som Korpus 2000 har rettet henvendelse til vedr. evt. brug af og ønsker til danske korpora findes under http://korpus.dsl.dk/korpus2000/inst-fortegnelse.htm; henvendelsens fulde ordlyd findes under http://korpus.dsl.dk/korpus2000/inst-henvendelse.html
3. VISL's korpus er tilgængeligt under http://corp.hum.ou.dk/corpuseye.html Oplysninger om dette korpus på hjemmesiden er yderst sparsomme
4. Dansk Talesprogskorpus, BySoc, er tilgængelig under http://www.cphling.dk/BySoc/index.html – på hjemmesiden findes der desuden flere oplysninger om dette korpus
5. En oversigt over nogle af brugerne og deres projekter findes under http://korpus.dsl.dk/korpus2000/ddo-brugere
6. Det skal her påpeges, at kravet om eksplicitte teksttypologiske kriterier, der i sidste ende vil kunne bidrage til en nøjagtig, formel, deklaration af korpora og dermed gøre korpora formelt sammenlignelige, indebærer en lang række hidtil kun sporadisk formulerede problemstillinger, hvis acceptable løsning næppe ligger lige om hjørnet Men kravet er en grundlæggende forudsætning for at kunne udføre empirisk korpusbaseret lingvistisk forskning. Oplysningskategorierne, som Korpus 2000 anvender i forbindelse med tekstregistreringen, findes under http://korpus.dsl.dk/korpus2000/etb-tekstkategorier
8. CQP: Corpus Query Processor. Flere oplysninger under http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

LITTERATUR

Keson, B. (1998): *Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus.* DSL.

Norling-Christensen, O. & Asmussen, J. (1998): The Corpus of the Danish Dictionary. *Lexikos* 8. Stellenbosch.

Ruus, H. (1995): *Danske kerneord* I-II. København.

Sinclair, J. (1996): *Preliminary recommendations on Corpus Typology.* EAGLES-rapport 1996