

Øvrige kommentarer: En bibliografisk record, skrevet i IBMs Arkivassistent, leverer oplysninger om gruppetilhør, forfatter, tekstomfang, tryksted, -år, mv. En generel beskrivelse af først og fremmest den danske og det tyske korpus findes i Hauschildt, Bent og Ole Lauridsen "Fagsproglige korpora og terminologi" i Picht, Heribert (red.) Terminologi, edb og vidensteknik. Nordterm-symposium 1989. Varde 1990, samt Lauridsen, Ole, Theis Riiber og Henning Søndergaard "Erstellung eines dänischen und eines deutschen Textkorpus - Fachsprache der Gentechnik" i Hermes 6/1991. Det erhvervssproglige Fakultet, Handelshøjskolen i Århus.

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

DET DANSKE MASKINLESBARE TEKSTKORPUS
INDEN FOR OMRADET GENTEKNIK

Af Ole Lauridsen (Handelshøjskolen i Århus)

1. Oprettelsen af det danske gentekniske tekstkorpus indledtes i november 1988 som et delvis forskningsrådsfinansieret samarbejde mellem handelshøjskolerne i Århus og København; det udvalgte forskerteam havde i opdrag udover det danske tekstkorpus også at oprette et tysk og skulle desforuden indgå i et snævert samarbejde med forskere fra Handelshøjskolen i København og Handelshøjskole Syd, der stod bag et tilsvarende spansk korpus. Ideen bag denne konstruktion var, at de tre korpora i videst muligt omfang skulle opbygges parallelt, så at de kunne danne basis for ikke kun sprogimmanente, men også kontrastive studier af forskellig art.

2. Ifølge Forskningsrådets opslag skulle korporaene rumme tekster, der med hensyn til oprindelse og funktion "viser fagsproget i sin anvendelse på videnskabeligt, teknisk, parlamentarisk og journalistisk plan". Denne præmis kræver en vertikal tilgang til det fagsproglige område og betegner således et brud med den praksis, der tidligere er blevet fulgt ved organiseringen af danske fagsproglige korpora (jur-korporaene samt de aftaleretlige korpora); disse har - som det tidligere er blevet sagt - udelukkende optaget tekster fra ekspertniveau og er altså opbygget ud fra et horisontalt grundkoncept. Den vertikale indfaldsvinkel skulle muliggøre undersøgelser af de ændringer, som - og jeg citerer her tidligere medlem af Forskningsrådet Henning Fonsmark - "fagsproget nødvendigvis må undergå, efterhånden som dets budskab skal kommunikeres til en stadig bredere, stadig mindre indforstået kreds af interesserede, og dermed trinvist skal gennemløbe en stadig mindre specialiseringsgrad uden nødvendigvis helt at miste sin karakter af fagsprog." Gruppen valgte endvidere at optage tekster skrevet af almenmand, i praksis læserbreve og lignende debatindlæg, for derigennem at åbne mulighed for undersøgelser af den

almindelige borgers håndtering af fagsproget - et område, der normalt ikke er blevet tilgodeset i tilstrækkeligt omfang.

3. Korporaene indeholder i overensstemmelse med Forskningsrådets opslag hver godt 1 mio. tekstord, som det efterhånden er blevet tradition i den angelsaksiske, danske og tyske korpusverden; jf. blot de to forudgående indlæg. Og der er her blot tale om tradition; millionen er ikke videnskabelig begrundet, men fuldstændig arbitrær. Det har dog i praksis vist sig, at korpora af dette omfang er umiddelbart håndterlige i forbindelse med pc-baserede analyser og ydermere danner et passende grundlag for de fleste typer af lingvistiske undersøgelser.

Ved oprettelsen af korpora i Danmark og efterhånden også i udlandet har man generelt stræbt efter at muliggøre diskursanalyser såvel som grammatisk-syntaktiske, leksikografiske samt terminologiske undersøgelser. Dette aldeles rimelige grundprincip fulgte vi, idet vi dog erkender, at omfanget næppe tillader fuldstændige leksikografiske bearbejdelser af området, ligesom tekstkorpora heller ikke bidrager essentielt til terminologiarbejdet.

4. En første forudsætning for overhovedet at kunne påbegynde litteratursøgningen var at søge begrebet genteknik defineret og afgrænset. Vi stødte her på det overraskende forhold, at der hverken på nationalt eller internationalt plan er udarbejdet nogen generelt anerkendt områdesystematik med dertil hørende definitioner. Og da hverken danske officielle love og akter eller den udenlandske litteratur bidrog til en afklaring, valgte vi som overordnet princip ved al litteratursøgning at lade teksten selv, i praksis afsenderen være afgørende faktor: skriver afsenderen under en eller anden synsvinkel om emnet genteknik, falder teksten inden for vores interessesfære, også selv om afsenders forståelse af begrebet genteknik eventuelt ikke måtte stå for en tilbundsående videnskabelig prøvelse. Vi traf altså alt i alt en helt principiel afgørelse om i videst muligt omfang at

indtage en absolut objektiv holdning til korpus, dvs. at lade teksten og dens forfatter styre tekstselektionen - herom mere detaljeret senere.

I mangel på definitionsbundne nøglebegreber indkredsede vi så ved læsning af tekster, der var blevet os anbefalet af fagfolk fra inden for genteknik, en række søgetermer, hvis anvendelse i litteratursøgningen efter vores bedste skøn ville garantere en optimal emnedækning. Ved litteratursøgningen blev især databaser, danske såvel som udenlandske, inddraget, men også direkte henvendelser til offentlige institutioner bar frugt.

5. I forbindelse med korpus' overordnede struktur var det som kort omtalt et grundlæggende princip for os at undgå subjektivitet; årsagen er åbenbar: en organisation, der bygger på en enkelt persons eller en snæver personkreds' egne, måske ligefrem stærkt specialiserede lingvistiske analyser af de inddragne tekster, kan hindre eller umuliggøre andres brug af korpus - netop af den grund valgte vi i øvrigt at afstå fra en grammatisk kodering ("tagging") af vores korpora.

For så at leve op til kravet om minimal subjektivitet i den overordnede korpusorganisation valgte vi meget enkelt princip, inspireret af den del af den Hoffmann/von Hahnske taksonomi, der tager udgangspunkt i begrebet kommunikationspartner, idet svært håndterlige, subjektivt betingede størrelser som "videnskabsmand" : "videnskabsmand (tekniker)" : "videnskabelig og teknisk leder af produktionen" erstattedes af færre og mere håndfaste og antallet af kommunikationsniveauer i medfør heraf nedsattes fra fem til tre. Vores princip muliggør en konsekvent kategorisering af teksterne efter afsender/modtager-forholdet; der er ikke tale om et indre kriterium, men om et ydre objektivt kriterium, der netop ikke forudsætter en forudgående lingvistisk analyse af teksterne - selve valget af dette inddelingskriterium er så til gengæld subjektivt, det skal ikke bestrides.

Med udgangspunkt i afsenders status som lægmand eller fagekspert, dette sidste forstået som enhver person, der på

et eller andet niveau arbejder professionelt med genteknik og/eller er uddannet inden for dette område, samt afsenders fastlæggelse af modtagers status opstilledes 3 kommunikationsniveauer:

I	fagekspert	<-->	fagekspert
II	fagekspert	<-->	lægmand
III	lægmand	<-->	lægmand

Ud fra disse 3 niveauer afgrænsedes 4 kommunikationsretninger, som så blev afgørende for inddelingen:

I	fagekspert	-->	fagekspert
II a	fagekspert	-->	lægmand
II b	lægmand	-->	fagekspert
III	lægmand	-->	lægmand

Som det vil fremgå, medfører den valgte struktur en ophævelse af den klassiske sondring mellem fagsprog og almensprog, idet sagområdet - eller domænet - her ikke behandles udelukkende af fagfolk, hvilket er den normale begrundelse for overhovedet at tale om fagsprog. Da korporaene rummer emnespecifikke tekster affattet af såvel fageksperters som lægmænd, foreligger snarere, hvad vi vil betegne *sagsprog*, og dette begreb bliver således en fællesmængde for fagsprog i klassisk og fagsprog i bredeste forstand; kun af praktiske grunde har jeg i det forudgående anvendt termen fagsprog.

Fastlæggelsen af afsenders og modtagers status skete på grundlag af oplysninger på bogomslag, i forord, bibliografier etc. I tvivlstilfælde henvendte vi os til afsenderen, til forlagene eller til redaktionerne; ikke signerede artikler medtoges kun, hvis udgiveren kunne opgive afsenders status. Hvis et kollektiv af fageksperters og lægmænd stod som afsender, kategoriseredes teksten efter afsenderstatus fagekspert, kun ikke, hvis en tekst faldt i entydige ekspert- og lægmændsdele (som tilfældet ofte er ved fx interviews); da foretog vi en dobbeltkategorisering, og i den

bibliografiske record er det så angivet i en note, hvilke dele der tilhører hvilken kommunikationsretning.

Også afsenders fastlæggelse af modtagers status har været relevant for kategoriseringen. Angiver afsender, at modtagergruppen består af både fageksperters og lægmænd, kategoriseredes teksten i overensstemmelse hermed - vores vurdering var irrelevant; der foretoges altså her en dobbeltkategorisering på samme måde som ved afsendere med forskellig status. Ved denne kategorisering har vi altså givet fuldstændig afkald på en subjektiv vurdering af teksterne og mener derved at opnå optimal objektivitet.

Der er tilstræbt en kvantitativt ligelig repræsentation af tekstord inden for hver af de fire kommunikationsretninger, altså 25% (250.000 ord) til hver. Imidlertid viste det sig, at dette mål ikke i alle henseender kunne nås, bl.a. fordi engelsk internationalt er dominerende på ekspertniveau, og når en given kommunikationsretning så ikke kunne opnå en repræsentation på 25%, blev den tiloversblevne ordplads fordelt ligeligt på de andre retninger.

Det maksimale omfang af et tekstexcerpt er fastlagt til 5000 ord i løbende tekst, men kan overstige denne grænse med 10%, hvis excerptet derved udgør et samlet hele. Valget af dette omfang er - som valget af korpus' samlede omfang - ikke videnskabeligt begrundet, men skyldes dels hensynet til diverse tekstlingvistiske undersøgelser, der kræver tekster af en vis længde, dels hensynet til ønsket om et stort antal tekster, der skulle sikre, at man ikke blot indfangede et mindre antal personers individuelle sprogbrug, men snarere kunne fremlægge et relativt bredt snit gennem sproget.

I én sammenhæng har vi imidlertid afvejet fra grænsen på 5000 ord + evt. 10%: vi måtte konstatere, at kategori IIb (lægmand (i henseende til genteknik oftest politikere) --> ekspert) var kraftigt underrepræsenteret, og besluttede derfor at indlæse fx loven om miljø og genteknologi samt lovforslag hertil in toto for at skabe et så rimeligt forskningsgrundlag som muligt.

Udvælgelsen af excerpter fra bøger er sket skiftevis forfra, midtfra og bagfra for at undgå en overvægt af mere

introducerende eller afrundende afsnit og for at kompensere for en eventuel stigende kompleksitet. Teksterne stammer fra årene 1978-1989 og placerer sig dermed i perioden fra genteknologiens introduktion i en bredere offentlighed frem til korpus' oprettelse.

6. Mens vi i forbindelse med den basale korpusorganisation har tilstræbt os på at arbejde objektivt, har vi i anden sammenhæng givet os subjektive vurderinger eller kategoriseringer i vold, nemlig ved oplysninger om aspekt, dvs. den eller de synsvinkler, som afsenderen anlægger, og oplysninger om publikationsform. Disse områder kan umiddelbart forekomme lige så objektive som kommunikationsretningerne, men i praksis har vi ud fra egne vurderinger måttet fastlægge et sæt af aspekter og publikationsformer, som andre muligvis ikke ville kunne anerkende: fx henføres tekster vedrørende dambrug til aspektet dyr og tekster vedrørende skovbrug til aspektet planter, idet vores læsning tilsagde os, at de som helhed passede bedre til disse områder end til fx aspektet industri; yderligere måtte vi overhovedet begrænse antallet af aspekter for at det ikke skulle udarte i overdetaljeret uoverskuelighed. Tilsvarende problemer rejser sig i forbindelse med publikationsform.

Information om aspekt og publikationsform er tillige med angivelse af forfatter, trykkeår, trykkested, titel osv. samlet som orienteringshjælp for brugeren i den bibliografiske record, der er baseret på ISO 690 Documentation - Bibliographic references - Content, form and structure og lagret i IBM's Arkivassistent.

7. Teksterne ligger - som teksterne i DK- og aftaleretkorporaene - i ASCII-filer, for øvrigt med en fejlpromille på under 1. Antallet af billeder, grafer etc. markeres ved tal kombineret med særlige tegn mellem registreringsnummer og tekst.

Kun i de tilfælde, hvor det er skønnet væsentligt for forståelsen af en tekst, er tegnet placeret inde i teksten på egen linje; men stadig gives der så en samlet oversigt mellem nummer og tekst (transparent 2).

8. Væsentlig for udnyttelsen af korpora er naturligvis en viden om, hvilket sprogteoretisk grundsyn der har dannet basis for hele organisationen; her derfor ganske kort nogle principielle betragtninger over begreberne korpus, sprog og repræsentativitet.

Vi antager ikke, at begrebet sprog er identisk med en eller anden mængde af faktiske ytringer, men kan på den anden side heller ikke acceptere tanken om, at faktiske ytringer blot er et bevis på eksistensen af menneskets evne til sprog; en sådan opfattelse vil i sin yderste konsekvens betyde, at et korpus' organisation reelt er ligegyldig - at et korpus udelukkende kan benyttes som problemgenerator ved afprøvning af hypoteser ofte skabt på grundlag af konstruerede sætninger, som i øvrigt meget vel kan være præget af den enkelte lingvists langvarige arbejde med et givet fænomen og dermed - vi kender det alle - måske være noget uldne. Vi mener heroverfor, at man ideelt set kun kan støde frem mod en forståelse af sprogsystemet og ideelt set kun skabe sine teorier ud fra analysen af en større mængde faktiske ytringer, og herved bliver korpus' status i forhold til det omgivende sprog for os afgørende for vægten af analyseresultaterne og dermed teorierne. Men hvad afspejler da et korpus?

Man kan i streng statistisk forstand som bekendt kun tale om en delmængdes repræsentativitet, hvis man kan definere den grundmængde, hvorefter delmængden tages. Et korpus må naturligvis betragtes som en delmængde, nemlig som delmængden af grundmængden "alle ytringer på et givet sprog", og denne kan af banale grunde, som jeg skal spare forsamlingen for, naturligvis ikke indfanges - uanset om sprog betragtes diakront eller synkront. Nok hævdes det, at der synkront faktisk er mulighed for at etablere en grundmængde "sprog", fx "avissprog 1990" eller "skønlitterært sprog 1985-1990" osv., og ud fra en sådan grundmængde kan der naturligvis skabes et repræsentativt korpus. Men: man har ved en sådan fremgangsmåde ikke taget udgangspunkt i selve sproget, men i medier, genrer osv. i fuldstændig vilkårlige tidsafsnit - og det er snarere sprogbrugen her, man så beskæftiger sig med.

Intet tekstkorpus kan ud fra denne synsvinkel gøre krav på betegnelsen repræsentativt, og de resultater, der nås via korpusanalyser, har i sidste instans da også kun gyldighed for sproget, som dette repræsenteres i det givne korpus - med andre ord: man når via korpus ikke endegyldigt frem til selve sprogsystemet; de tidligere ofte alt for få, generelt disparate og tilfældigt excerperede tekstbelæg er reelt blot erstattet af anderledes store og sammenhængende tekstmængder.

Men selv om repræsentativitetstanken altså afvises, vil jeg ikke dermed postulere, at vores og de tidligere oprettede fagsproglige korpora er uden udsagnskraft om de respektive fag- og sagsprog. Allerede de intense og vidtfavnende litteratursøgninger, for vores vedkommende også den meget brede niveaudækning, den trods alt ikke ubetydelige mængde af enkelttekster og enkeltafsendere og selvfølgelig også det ganske store samlede omfang - eller sagt i få ord: forstudierne og organisationen sandsynliggør, at tekstsamlingerne giver en pålidelig genspejling af sagsproget, en genspejling, der så åbner adgang for videre teoretiske overvejelser, for teoridannelser, som alt i alt må siges at stå på et sikrere grundlag end tidligere tiders kartotekskasse-lingvistik, for nu at føje nok en bindestrekslingvistik til rækken.

9. I vores opdrag lå udelukkende oprettelsen af de to korpora, ingen egentlig forskningsmæssig udnyttelse; for dog at få en idé om tekstsamlingernes ordsammensætning og for dermed at åbne mulighed for leksikografiske, orddannelsesmæssige samt i en vis udstrækning terminologiske pilotundersøgelser, har vi i Afdelingen for Datalingvistik igangsat frekvens- og dispersationsanalyser dels af hvert korpus samlet, dels af enkeltniveauerne i hvert korpus. Yderligere er det danske korpus blevet udnyttet af Carol Henriksen, RUC, som baggrund for videregående kommunikationsanalyser, der inddrager illustrationsmaterialet, ligesom Uwe Kaufmann, Landbohøjskolen, på længere sigt vil benytte det som eksem-

plifikationsmateriale ved udarbejdelsen af et genteknologisk opslagsværk.

10. Alle vil mod betaling af de faktiske omkostninger kunne rekvirere kopier af korpora og records under forudsætning af, at materialet udelukkende udnyttes i forskningsøjemed og ikke videregives til tredjemand.

Henvendelse til: Ole Lauridsen, Tysk Institut, HHÅ, Fuglesangs Allé 4, 8210 Århus V.

11. En oversigt over danske maskinlæsbare korpora følger på K.M. Lauridsens artikel om det aftaleretlige korpus.