

**ET DANSK TEKSTKORPUS INDEN FOR EMNEOMRÅDET AFTALERET
FAGSPROG OG KORPUSLINGVISTIK**
Af Karen M. Lauridsen (Handelshøjskolen i Århus)

Statens Humanistiske Forskningsråd etablerede i 1987 initiativområdet Fagsprog og fagsproglig kommunikation i erhvervslivet. Et af de første projekter under initiativområdet har været etableringen af tre parallelle korpora inden for emneområdet aftaleret: 3 x 1 mio. løbende ord i dansk, engelsk, og fransk. I det følgende vil tre delemler i forbindelse hermed blive berørt:

1. de generelle principper for etableringen af de aftaleretlige korpora
2. de maskinlæsbare korpora
3. nogle teoretiske overvejelser vedrørende fagsprog / fagsprogsforskning og korpuslingvistik

1. De generelle principper for etableringen af de aftaleretlige korpora

Emnet aftaleret er begrænset til den generelle del af aftaleretten. For hvert sprog er der udarbejdet en bibliografi, som er struktureret teksttypologisk i følgende seks sektioner:

1. love og andre regler
2. lovforarbejder
3. domme
4. kontrakter
5. juridiske lærebøger
6. artikler i juridiske tidsskrifter

Samtidig hermed opereres med en tematisk klassifikation af emneområdet. Denne er gjort temmelig fleksibel for at kunne dække de tre retssystemer på trods af deres forskellighed, således at det er muligt ikke blot at arbejde med de tre tekstkorpora enkeltvis, men også at kontrastere dem i deres helhed eller inden for de enkelte teksttyper.

Teksterne til de maskinlæsbare korpora er herefter udvalgt på basis af disse to klassifikationer: så vidt

muligt en ligelig fordeling mellem de seks teksttyper samt en tematisk baseret fordelingsnøgle. Denne er fastsat i samarbejde med danske jurister således, at de temaer, der rent faktisk giver anledning til den største tekstproduktion, også er blevet vægtet mest i de maskinlæsbare korpora.

Det har ved udvælgelsen af teksterne været et grundlæggende princip, at de skulle være så "repræsentative" som muligt. Repræsentativitet er dog en meget vanskelig størrelse at operere med i denne sammenhæng, og begrebet kan her ikke opfattes i streng, statistisk forstand. Man kan af indlysende årsager ikke definere en grundmængde sprog, endsige en grundmængde sprog vedrørende aftaleret, og man kan derfor heller ikke tale om en repræsentativ delmængde heraf. De tekster, som er medtaget i de maskinlæsbare korpora kan snarere beskrives som en approksimeret delmængde af teksterne i de etablerede bibliografier; men samtidig er det uomgængeligt, at bibliografierne, selv om de dækker en afgrænset periode på ti år (1978-1987), og selv om projektgruppen har forsøgt at gøre dem så altomfattende som muligt, alligevel kun udgør hvad man kunne kalde en kunstig grundmængde, og de maskinlæsbare korpora i aftaleret kan følgelig maksimalt blive repræsentative i forhold til disse kunstigt etablerede grundmængder, jf. også Ole Lauridsens indlæg i nærværende publikation.

Det er således oplagt, at tekstudvælgelsen nok er foretaget på basis af nogle teoretiske overvejelser, men i sidste ende har diverse praktiske hensyn spillet en ikke uvæsentlig rolle. Det har således bl.a. også på grund af de juridiske systemers forskellighed været nødvendigt at justere fordelingen af tekstmaterialet inden for de enkelte teksttyper; og hvad angår fordelingen i henhold til teksttema, har det ikke været muligt at gennemføre denne inden for teksttyperne 1. love og andre regler og 4. kontrakter, idet disse normalt vedrører alle eller i hvert fald størstedelen af teksttemaerne.

Den enkelte tekst er en tekst i sin helhed; det vil sige en dom, et (afrundet) afsnit i en lærebog, en kontrakt, etc. Det betyder også, at tekstlængden undertiden overstiger

et i øvrigt tilstræbt maximum på 5.000 løbende ord.

2. De maskinlæsbare korpora

De maskinlæsbare korpora foreligger som ASCII filer. Hver enkelt tekst er forsynet med et filnummer, der indeholder en kode for, hvilken teksttype den pågældende tekst tilhører, samt et løbenummer inden for den enkelte teksttypekategori. Samme filnummer / tekstnummer findes på en liste over de i hvert enkelt korpus forekommende tekster / filer. De aftaleretlige tekstkorpora er tilgængelige for forskere ved de højere læreanstalter under forudsætning af, at de udelukkende anvendes til forskningsformål. Der henvises i øvrigt til omstående liste over tilgængelige maskinlæsbare korpora i Danmark.

3. Nogle teoretiske overvejelser vedr. fagsprog / fagsprogsforskning og korpuslingvistik

De ovenfor beskrevne tekstkorpora er som allerede anført de første i en række fagsproglige tekstkorpora, som er blevet etableret under Statens Humanistiske Forskningsråds initiativområde Fagsprog og fagsproglig kommunikation i erhvervslivet. At etablere fagsproglige tekstkorpora og bedrive fagsproglig forskning forudsætter en definition af, hvad fagsprog er for en størrelse, og netop dette spørgsmål har affødt en til tider ganske intens debat i de seneste år. Det sidste ord i denne sag er næppe sagt - eller skrevet - endnu, og nedenstående skal således opfattes som et af flere mulige svar på samme spørgsmål.

Set ud fra et lingvistisk synspunkt forekommer den traditionelle distinktion mellem almindeligt sprog og fagsprog ikke heldig, fordi der i begge tilfælde er tale om sprog, om realisationer af et og samme sprogsystem, der basalt set forudsætter en og samme sproglige kompetence. Hvad er det da, der betinger den ofte anførte skelnen mellem almindeligt sprog og fagsprog? Der er naturligvis frekvensforskelle inden for leksikken og mellem bestemte syntaktiske strukturer, men den måske mest afgørende forskel er ikke-lingvistiske faktorer som først og fremmest tekst- eller sprogproduktionsparametre: afsender-modtager forhold,

domæne og medium. Disse parametre betinger tekstens konstituering leksikalsk, grammatisk og tekstuel gennem nogle restringerede valg, der er betinget af sociokulturelle normer eller konventioner. I sin yderste konsekvens kunne en sådan definition medføre, at disse lingvistiske og ikke-lingvistiske parametre definerer en hvilken som helst ytring eller tekst som fagsprog. Omvendt synes en definition af fagsprog, som udelukkende inddrager kommunikation mellem fageksperter at være for snæver (jf. i øvrigt Ole Lauridsens bidrag til nærværende publikation). Det synes derfor at være mest operationelt, omend - vil nogen måske hævde - ikke det teoretisk set mest stringente, at definere fagsprog på basis af de allerede nævnte parametre: afsender-modtager relationer, domæne og medium således, at egentlige fagsproglige tekster er tekster, som omhandler nærmere definerede domæner i bestemte medier og med bestemte, nærmere definerede afsender-modtager forhold.

En sproglig ytring som f.eks. Programmet implementeres i computeren ... er således ikke nødvendigvis fagsprog, selv om den indeholder termene program og computer samt implementere, der er en del af en bestemt fagterminologi, samt en verbalfrase i passiv, som statiske set ofte forekommer i såkaldt fagsproglige tekster. Det afgørende vil her være tekstens klassifikation i de ikke-lingvistiske parametre: f.eks. afsender = firma / ekspert; modtager = ekspert, semi-ekspert eller lægmand; domæne: edb / software; medium = manual.

De maskinlæsbare korpora i aftaleret kan relativt enkelt indplaceres i de ovenfor nævnte ikke-lingvistiske sprogproduktionsparametre. Domænet er fastlagt som jura med subdomænet aftaleret; medium og afsender-modtagerrelationer kan for hver enkelt tekst afgøres på basis af teksttypologien, og de enkelte korpora kan herefter anvendes i leksikalske, lingvistiske eller tekstlingvistiske undersøgelser. Hvorledes man så skal anvende korpora i sådanne analyser er et spørgsmål for sig, som kort vil blive berørt nedenfor.

I modsætning til andre som f.eks. Jensen (1990) og

Vikner (1990), der definerer fagsprog som fagfolks sproglige kompetence, og som dermed repræsenterer en generativ sprogteori, betragtes fagsprog altså her som en bestemt sprogbrug, med andre ord, en bestemt måde at anvende et bestemt sprog eller sprogsystem som f.eks. dansk på. Denne sprogbrug har ganske givet en række leksikalske, grammatiske og pragmatiske karakteristika i forhold til sproget generelt; men indtil disse er nærmere beskrevet, må man tage til takke med at afgrænse denne sprogbrug på basis af primært ikke-lingvistiske faktorer.

Der er ved anvendelsen af korpora i sprogforskningen en række åbne spørgsmål, som man nødvendigvis må tage stilling til, inden man går i gang. Et af dem er spørgsmålet om lingvistisk kodering, der i alle tilfælde har én fordel, men samtidig også mindst en ulempe. Et koderet korpus bliver uomgængeligt (som bl.a. Jensen (1990:9) har beskrevet det) en teoriladet størrelse, idet en kodering med hensyn til f.eks. ordklasser forudsætter en bestemt lingvistisk teori. Da formålet med etableringen af de aftaleretlige korpora var og er, at de skal kunne bruges af en bredere kreds af lingvister, som ikke nødvendigvis alle arbejder på basis af en og samme sproglige teori, må det betragtes som en fordel, at disse korpora ikke er blevet koderet. (Enhver overvejelse med hensyn til automatisk over for manuel kodering eller økonomiske forhold lades i nærværende sammenhæng helt ude af betragtning.) Den manglende kodering betyder til gengæld, at der heller ikke er foretaget homografseparering eller lemmatisering af de enkelte ordstrengene i de maskinlæsbare korpora, og at man i en eventuel automatisk analyse f.eks. ikke umiddelbart vil kunne skelne mellem homografer som f.eks. substantivet en måtte og verbet at måtte; den enkelte bruger må altså selv foretage en kodering af disse korpora, hvor det skønnes nødvendigt.

Et andet spørgsmål er så den metode, som anvendes i den såkaldte korpuslingvistik. Lad det med det samme være slået fast, at der med korpuslingvistik ikke menes en bestemt lingvistisk skole, men blot dette, at man fra et systematisk sammensat korpus excerperer data til en sproglig analyse

(ligesom datalingvistik blot betyder, at man benytter en datamat i en sådan analyse). Der er altså tale om, at man på basis af en allerede eksisterende sprogteori foretager en kvalitativ analyse af udvalgte data, og det hævdes på ingen måde, at man på basis af korpus umiddelbart og rent induktivt kan extrapolere lingvistiske eller tekstlingvistiske strukturer.

Det er altså hensigten, at de fagsproglige tekstkorpora, sådan som de nu foreligger, skal kunne bruges som data i en nærmere bestemmelse af en given sprogbrug eller, hvis udgangspunktet er en generativ synsvinkel, som data til at teste hypoteser vedrørende fagfolks sproglige kompetence, jf. igen f.eks. Jensen (1990), Vikner (1990), eller Jensen, Sørensen og Vikner (1989). De forskellige typer af fagsproglig sprogbrug er på nuværende tidspunkt ikke særligt velbeskrevne, og resultaterne af de analyser, som i øjeblikket foreligger eller er under udarbejdelse vil forhåbentlig ad langs tilsammen kunne konstituere en slags fagsprogsgrammatikker.

Endelig kan man håbe, at der i de kommende år vil blive etableret en række korpora, som kan supplere de allerede foreliggende, f.eks. ved at tilføje nye sprog, nye domæner eller nye teksttyper til allerede eksisterende korpora, eller som kan udgøre en helt ny serie af maskinlæsbare korpora. At etablere maskinlæsbare korpora af en vis størrelse som eksempelvis den næsten kanoniske million ord, vil på grund af den teknologiske udvikling ikke være den uoverkommelige opgave, som det har været indtil for ganske nylig.

Litteratur:

- Dyrberg, Gunhild, Dorrit Faber, Steffen Leo Hansen & Joan Tournay (1988) "Etablering af et juridisk tekstkorpus" i Hermes 1/1988. Det erhvervssproglige Fakultet, Handelshøjskolen i Århus
- Jensen, Per Anker (1990) "Korpusarbejde, sprogbrug og sprogteori" i LAMBDA nr. 15. Institut for Datalingvistik, Handelshøjskolen i København
- Jensen, Per Anker, Finn Sørensen & Carl Vikner (1988) "Data, Korpora, and Linguistic Research" i Hermes 2/1989. Det erhvervssproglige Fakultet, Handelshøjskolen i Århus
- Vikner, Carl (1990) "Problematisering af anvendelsen af tekstkorpora i fagsproglig forskning" i LAMBDA nr. 15. Institut for Datalingvistik, Handelshøjskolen i København