Gunhild Dyrberg, Dorrit Faber,
Steffen Leo Hansen og Joan Tournay

Etablering af et juridisk tekstkorpus

Abstract

The project involves the collection of a corpus of English-French-Danish legal texts exemplifying different text types and different themes within the subject area of the law of contract. Each language will be represented by 1 million words. The text corpus will be computerized and is intended as the empirical data for various types of linguistic research such as text typology, text linguistics, syntax, semantics, etc.

The principles which govern the assembly of the corpus - strict definition and delimitation of the text universe, combination of text type and theme classification - are assumed to ensure representativity of the corpus.

The corpus forms part of a major project under the auspicies of the Danish Research Council for the Humanities, "Communication in Language for Special Purposes".

1. Baggrund og formål

For at styrke kommunikationen i erhvervslivet er der med støtte fra Statens Humanistiske Forskningsråd iværksat en række projekter, der under ét betegnes som initiativområdet "Fagsproglig Kommunikation".

Dette initiativområde er udsprunget af et forskningsprogram, som er udarbejdet af Handelshøjskolen i København, Handelshøjskolen i Århus og Handelshøjskole Syd, og som sigter mod at styrke den fagsproglige forskning.

Et led i dette forskningsprogram er oprettelsen af et engelsk-fransk-dansk juridisk tekstkorpus, som vil blive indlæst i en datamat, og som skal gøre det muligt at gennemføre en række forskellige empiriske undersøgelser, herunder teksttypologiske, tekstlingvistiske og syntaktiske undersøgelser af fagsproglige tekster. Projektet gennemføres i et samarbejde mellem Handelshøjskolen i København og Handelshøjskolen i Århus.

Gunhild Dyrberg og Joan Tournay, Institut for Fransk HHK, og Dorrit Faber, Institut for engelsk HHK, står for oprettelsen af det engelsk-fransk-danske korpus. Steffen Leo Hansen, Institut for Datalingvistik HHK, er konsulent, og Henning Bergenholtz, forskningsprofessor ved HHÅ, er både konsulent og koordinator for dette og evt. kommende korpora. Tilknyttet projektet er desuden Karen M. Lauridsen, Engelsk institut HHÅ, som står for supervisionen af den optiske indlæsning, og Helle Pals Frandsen, Institut for engelsk HHK, som deltager i udvælgelsen og den tekstuelle kodering af tekster til de engelske og danske korpora.

Det engelsk-fransk-danske korpus består af tre delkorpora, som etableres efter de samme principper, således at det indlæste tekstmateriale kan bearbejdes kontrastivt, samtidig med at det selvfølgelig kan danne udgangspunkt for enkeltsprogede projekter.

Disse principper er fastlagt under inddragelse af erfaringerne fra et tidligere projekt ved Handelshøjskolen i København med titlen "Database til terminologisk information og generering af ordbøger", kaldet Jurpilotprojektet, og er i øvrigt baseret på korpuslitteratur af bl.a. Henning Bergenholtz, Burkhard Schaeder og Steffen Leo Hansen.

2. Tekstkorpus

2.1 Hvorfor et tekstkorpus?

Formålet med at basere et projekt som det foreliggende på anvendelsen af et maskinlæsbart tekstkorpus er på den ene side - og først og fremmest - at udnytte muligheden for at indsamle og bearbejde store tekstmængder hurtigt og præcist ved anvendelsen af en datamat og tilgængeligt software som fx. ved optisk indlæsning af teksterne, ved at anvende et databasesystem til opbevaring af selve tekstkorpus og ved anvendelsen af programmel til lingvistiske undersøgelser af de indsamlede data. På den anden side er der det rent praktiske, men ikke mindre væsentlige sigte: en samling tekster som er lagret og tilgængelige for on-line søgning og genfinding i et databasesystem er som tekstmateriale tilgængeligt for mange andre end de implicerede forskere og kan anvendes til flere forskellige formål, og det kan til enhver tid opdateres ved at tilføre nye og relevante tekster.

Antallet af korpusbaserede projekter er da også vokset støt siden de første forsøg i denne retning i slutningen af 60´erne, og formålet har først og fremmest været at beskrive de sproglige data i det

maskinlæsbare tekstkorpus, enten i form af frekvensanalyser som typisk for de tidlige projekters vedkommende, eller med henblik på at bruge sprogbeskrivelsen som udgangspunkt for leksikografiske og terminologiske projekter.

2.2 Hvad er et tekstkorpus?

De maskinlæsbare tekstkorpora, som er ved at blive etableret nu under initiativområdet, bygger på følgende definition af et tekstkorpus:

Ved et tekstkorpus forstås en mængde af tekster der som data udgør det empiriske grundlag for en metodisk bearbejdning af specifikke egenskaber ved disse data, idet de som delmængde er udtaget fra en grundmængde på en sådan måde at delmængden antages at være repræsentativ for grundmængden med hensyn til omfang, struktur og arten af egenskaber.

Det skal understreges, at et tekstkorpus er en samling tekster, og at det som sådan adskiller sig fra det der kaldes belægsamlinger, dvs. en samling data som udvalgte eksempler på specifikke fx. morfologiske eller syntaktiske fænomener. Det skal endvidere bemærkes, at vi herefter bruger udtrykkene tekstkorpus og korpus som betegnelse for et tekstkorpus i ovenfor nævnte forstand.

Den definition som er givet ovenfor giver anledning til at komme ind på nogle væsentlige problemer knyttet til oprettelsen af et tekstkorpus, idet der ovenfor er nævnt egenskaber ved et korpus som omfang, struktur og repræsentativitet.

2.3 Principper for etablering af et tekstkorpus

Et tekstkorpus etableres med en ganske bestemt funktion for øje. Der er tale om dels en brugsfunktion som fx. et tekstkorpus til undersøgelse af fagsproglig kommunikation, og dels om en metodisk funktion, idet selve korpus som data skal kunne behandles og beskrives under anvendelsen af metoder fra områder som fx. teoretisk lingvistik, tekstlingvistik m.v. Såvel brugsfunktionen som den metodiske funktion har afgørende betydning for sammensætningen af et korpus. Ved sammensætning forstår vi kombinationen af et bestemt indhold, et givet omfang og en given strukturering af datamængden. Brugsfunktionen afspejler sig i valg af tekstmateriale og dermed i indholdet i korpus, og den metodiske funktion i det omfang og den struktur som et korpus har.

De her nævnte kategorier, indhold, omfang og struktur, og den

måde de realiseres på konkret i et korpus, er afgørende for det forhold, der består mellem på den ene side det maskinlæsbare tekstkorpus som en delmængde, og på den anden side det tekstunivers, grundmængden, som teksterne repræsenterer, dvs. for hele den problematik der er knyttet til begrebet repræsentativitet. Det skal bemærkes, at vi bruger ordet repræsentativ i den almindelige betydning og ikke i betydningen statistisk repræsentativ.

Da målet for en korpusanalyse ikke blot er at beskrive de data, der er repræsenteret i korpus, men tillige at kunne generalisere denne beskrivelse til også at gælde for data af samme art i grundmængden, må man nødvendigvis forsøge at sammensætte sit tekstkorpus på en sådan måde, at det forekommer rimeligt og sandsynligt at slutte fra observerede egenskaber i tekstkorpus til tilsvarende egenskaber i tekstuniverset.

Traditionelt arbejdes der derfor også med en antagelse om, at et tekstkorpus hhv. er og/eller skal være repræsentativt, og at et sådant korpus pr. definition tillader sådanne generaliseringer. Imidlertid synes det at være svært at finde frem til nogle præcise kriterier for hvad repræsentativitet er, ligesom det kan diskuteres, om man overhovedet kan lave et repræsentativt tekstkorpus.

I et forsøg på at løse denne problematik har vi på HHK, med udgangspunkt i de erfaringer vi har høstet i vores første, tidligere omtalte korpusbaserede projekt, opstillet nogle kriterier for det vi kalder et repræsentativt korpus. Disse kriterier består i nogle retningslinier for sammensætning af korpus, dvs. retningslinier for de beslutninger der vedrører indhold, struktur og omfang.

En afgørende betingelse for at kunne etablere et repræsentativt, maskinlæsbart tekstkorpus er, at det tekstunivers, dvs. den grundmængde man udtager teksterne fra, er en finit mængde, at man altså kan opremse samtlige elementer i mængden. En sådan betingelse synes imidlertid at gøre det helt umuligt at etablere et repræsentativt korpus, og i argumentationen mod et sådant korpus efterlyses gerne kriterier for sammensætning af et almensprogligt repræsentativt korpus for fx. moderne britisk-engelsk. Men hvis man forkaster ideen om et repræsentativt korpus, bliver også selve det korpusbaserede projekt en tvivlsom affære for så vidt som man ønsker at bruge korpus til at beskrive, forklare eller postulere egenskaber uden for korpus.

Vi har på HHK valgt en pragmatisk løsning på dette problem, idet det må være op til den eller de forskere, som etablerer korpus, at definere og sandsynliggøre den finitte grundmængde, der er grundlaget for deres maskinlæsbare tekstkorpus. Og det vil i praksis

sige, at det første man skal gøre er at opstille en bibliografi for det område, man ønsker at arbejde indenfor eller bearbejde og sandsynliggøre, at den til det givne formål er både nødvendig og tilstrækkelig. Med udgangspunkt i denne bibliografi som den finitte grundmængde kan man da etablere et korpus, som er repræsentativt mht. indhold, omfang og struktur.

Hvilket omfang et korpus skal have afhænger i høj grad af den tidligere omtalte metodiske funktion. Fra de meget omfattende leksikostatistiske undersøgelser som blev gennemført specielt i 60´erne og begyndelsen af 70´erne ved man, at et korpus, som fx. udelukkende skal anvendes til undersøgelse af ordhyppigheder, kan være væsentligt mindre end et korpus, der skal anvendes til undersøgelse af mere komplekse, sproglige strukturer. Der findes en lang række projekter, som primært stiler mod en sprogbeskrivelse, der omfatter de egenskaber ved et sprog, som optræder med en sådan regelmæssighed, at de kan beskrives med statistiske metoder. Mulighederne for at gennemføre hel- eller halvautomatisk korpusanalyse er imidlertid efterhånden blevet så gode, at ønsket og forventningen om at kunne arbejde med mere komplekse strukturer og også ud over sætningsgrænserne er vokset. Og i så fald vil det være andre kriterier, der bestemmer omfanget af korpus. Men helt præcist hvor, man skal sætte grænserne for omfanget af et korpus, er naturligvis ikke til at sige. Også på dette punkt spiller der mere pragmatiske faktorer ind i form af fx. lagerkapacitet og statistisk bekvemmelighed - det er således mere "bekvemt" at lave statistiske beregninger på et korpus med 1 mio. løbende ord.

Ud over de rent kvantitative overvejelser knyttet til beslutningen om omfanget af korpus, vil det før oprettelsen af et tekstkorpus være nødvendigt at tage stilling til hvilken struktur man vil lægge i sit korpus. Med struktur menes i denne sammenhæng ikke kun fordelingen af stikprøver i grundmængden, men der tænkes i nok så høj grad på den kvalitative struktur i det tekstunivers, som skal repræsenteres og på metoder og midler til at få denne struktur repræsenteret i det maskinlæsbare tekstkorpus.

I forhold til den grundlæggende bibliografi kan man således typisk strukturere korpus enten med udgangspunkt i en teksttypologi eller ud fra en tematisk klassifikation af det valgte område, eller ved en kombination af begge muligheder. Under udtagningen af stikprøver kan man endvidere sørge for at sikre sig en ligelig eller nærmere defineret relativ repræsentation af de enkelte teksttyper, ligesom man sørger for at fordele stikprøverne over samtlige tekster. Stiler man fx. mod et korpus på 1 mio. løbende ord, kan man udtage

otte stikprøver à 125.000 ord. Det vil imidlertid nemt føre til en atypisk repræsentation af enten en enkelt teksttype eller en enkelt forfatter. Udtager man derimod 500 stikprøver af 2000 løbende ord fordelt over hele tekstuniverset, vil man få et langt mere pålideligt korpus set i forhold til dette tekstunivers.

Endelig kan der tillige blive tale om en yderligere struktur i korpus, nemlig den der er bestemt af det databasesystem, man arbejder med, hvor de enkelte records indeholdende tekster typisk vil være forsynet med informationskategorier som "forfatter, titel, udgivelse, teksttype, emne" mv. Denne struktur er dog først og fremmest anvendelsesorienteret og beregnet på brugerne af korpus, der på denne måde får mulighed for at selektere i deres søgninger og arbejde på faste eller varierende delmængder af korpus.

Sammenfattende kan problematikken knyttet til beslutninger vedrørende korpusstruktur formuleres som problemet: hvordan sikrer man sig et kvalitativt repræsentativt korpus, og en mulig løsning på dette er at gå frem som beskrevet ovenfor.

Det er de her skitserede principper og overvejelser, der ligger til grund for de tekstkorpora, som er ved at blive oprettet, idet vi dog samtidig på baggrund af tidligere erfaringer tillader pragmatiske løsninger i de situationer, hvor det skønnes eller er nødvendigt, forudsat at sådanne løsninger i hvert enkelt tilfælde beskrives og dokumenteres.

Vi er således nået frem til at opstille en bibliografi for hhv. fransk og engelsk og har opstillet såvel en teksttypologi samt en tematisk klassifikation for det valgte område. Dermed har vi kortlagt det valgte områdes kvalitative systematik og kan bruge den som grundlag for valg af tekster til de maskinlæsbare tekstkorpora, der rent kvantitativt for hvert enkelt sprog bliver på 1 mio. løbende ord, sammensat af tekster på indtil 5000 løbende ord.

3. Det valgte emneområde

Da det overordnede formål er at fremme fagsproglig kommunikation, har vi valgt et emne, der har en central placering i erhvervslivet og derfor også i undervisningen på Handelshøjskolerne, nemlig aftaleret.

Aftaleret er imidlertid et meget omfattende retsområde, der kan ses som bestående af en generel del og en mere specifik del. Førstnævnte del behandler de generelle retsregler, der finder anvendelse på de konkrete kontrakttyper (køb, leje, lån etc.), der indgår i den specifikke del. Da den juridiske litteratur, der specifikt

behandler de enkelte kontraktforhold, i de fleste tilfælde er meget omfattende, har vi i første omgang afgrænset korpusemnet til de generelle aspekter af aftaleretten. På denne måde skabes der et fundament, som efter behov kan udvides med korpora inden for de specifikke kontraktforhold. (Som teksttype indgår konkrete aftaler naturligvis i korpus).

4. Teksttypologi

4.1 Principper for typologisering

En teksttypeklassifikation burde ideelt set tage sit udgangspunkt i en beskrivelsesmodel, der typologiserer tekster ud fra tekstlingvistiske analyseprincipper. Imidlertid findes der ikke så vidt det os bekendt en sådan teksttypologisk model, som umiddelbart kan anvendes til andet end en bestemmelse af tekster i brede kategorier som narration, instruktion, argumentation etc. Inden for det enkelte fagsprogsområde inddeles tekster følgelig ofte efter afsender/modtagerfaktoren eller efter funktion.

Etableringen af fagsproglige korpora som de her omtalte, ser vi som en forudsætning for, at man kan arbejde videre ad lingvistisk vej med en nærmere kortlægning af for det første hvad fagsprog er, altså hvilke sproglige træk der kan siges at karakterisere fagsproglige tekster i modsætning til (eller eventuelt i lighed med) almensproglige tekster - inden for almensprog har der jo i flere år eksisteret flere korpora, i hvert fald for engelsk og tysk således at der er et sammenligningsgrundlag at arbejde med - og for det andet de forskellige teksttyper, der kan bestemmes inden for særlige fagsprogsområder.

Det betyder for vores opstilling af teksttyper, at vi ikke baserer den på tekstanalytiske kriterier, men på vores viden om de tekstfunktioner, der er aktuelle inden for vores emneområde, og på afsender-modtager-faktoren. Samtidig er vi naturligvis styret af det overordnede formål med etableringen af disse korpora: at skabe et empirisk grundlag for analyser af bl.a. tekstlingvistisk og syntaktisk art. Endelig, men ikke mindst, er der det mere praktisk betonede hensyn, at jo flere teksttyper vi opdeler materialet i, jo mindre "plads" vil den enkelte teksttype kunne optage og jo spinklere bliver analysegrundlaget.

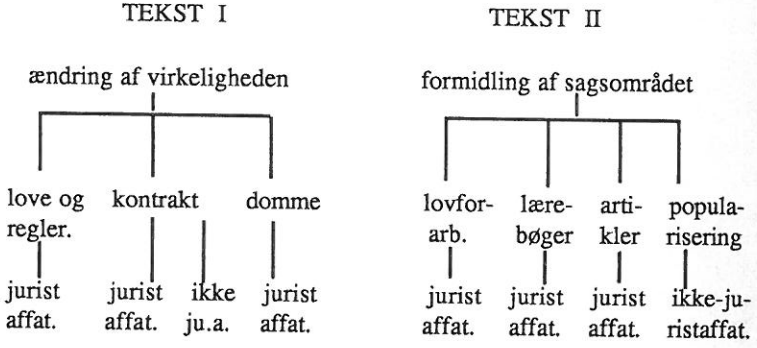
4.2 Den valgte teksttypologi

Disse betragtninger har medført, at vi har inddelt det valgte område i 6 tekstfunktioner, som vi mener er relevante for vores emne og korpusformål:

love og andre regler
lovforarbejder
domme
kontrakter
juridiske lærebøger
artikler i juridiske tidsskrifter

Efter overordnet funktion mener vi, at tekster med et aftaleretligt indhold kan inddeles i 2 hovedgrupper - en gruppe tekster, der ændrer "virkeligheden", og en anden der formidler et aftaleretligt sagsindhold.

Denne forskel i funktion, forventer vi, vil have betydelige sproglige og tekstuelle konsekvenser, som det vil være interessant at arbejde med ikke mindst i en kontrastiv analyse.



Hvad angår den tekstkategori, der ændrer virkeligheden, kan afsender foruden juristen også være den almindelige borger, typisk forretningsmanden, fordi der i de berørte sprogområders retssystemer er aftalefrihed. Dette indebærer, at borgerne frit kan indgå de kontrakter de ønsker, og at de med visse forbehold vil være bundet af de vilkår, de har aftalt, og i teorien i det mindste er det kun ved tvister, at eksperter/jurister behøver at blive involveret.

For den kategori tekster, der formidler et aftaleretligt sagsindhold, vil afsender ligeledes kunne være både fagmand og lægmand, idet journalister og andre naturligvis kan tænkes at skrive om disse forhold.

Fælles for de teksttyper vi har medtaget er, at de er affattet af fagmanden, juristen. Modtager og bruger er i første række også fagmanden, og kun sekundært lægmanden som bruger. Teksttyperne lovforarbejder, juridiske lærebøger og artikler er helt klart rettet mod andre eksperter eller kommende eksperter, mens den egentlige modtagergruppe for teksttyperne love og andre regler, domme og kontrakter, nemlig den almindelige borger eller forretningsmand, for den langt overvejende dels vedkommende har brug for juristen, fagmanden, som fortolker af indholdet eller som en form for varetager af dokumenternes indhold.

Tekster affattet af fagmand til lægmand med et populariseringsformål eller af lægmand til lægmand er derimod ikke repræsenteret i vores korpora. Det er helt klart at disse teksttyper er af tekstanalystisk interesse, idet man her kan iagttage f.eks. hvorledes en høj "ekspertfaglighedsgrad" omsættes til en lavere. Imidlertid gælder det for vores emneområde, den generelle del af aftaleretten, at disse tekster enten er meget vanskelige at fremskaffe (f.eks. privat korrespondance firmaer imellem om aftaleretlige forhold) eller de er ikke almindeligt forekommende, fordi de områder, hvor sådanne formidlende tekster vil være almindelige, især er de områder, der hører ind under den specifikke del af aftaleretten, f.eks. forbrugerret, selskabsret, lejeret, køberet osv. og som derfor falder uden for vort emneområde.

Det er således vores emneområde/afgrænsning af emneområdet der har betinget antallet og arten af teksttyper, der antages at repræsentere den typiske og anerkendte sprogbrug inden for området. Substandard tekster er derfor ikke inddraget. Dette har naturligvis visse tekstanalytiske konsekvenser, men vi tror, at mange forskere og undervisere i fagsprog og i oversættelse af juridiske tekster vil være enige med os i dette valg. Det er jo stadig således, at vi i forbindelse med oversættelse og kommunikation i det hele taget i høj grad har brug for at kende og kunne anvende den anerkendte sprogbrug, hvilket forudsætter en indgående og systematisk beskrivelse af fagsproglige sprogbrugsmønstre.

5. Tematisk klassifikation

5.1 Principper for klassifikation

For at sikre, at den delmængde, der udvælges fra vort tekstunivers (den etablerede bibliografi) til at indgå i det maskinlæsbare korpus, er repræsentativ for området, skal der som

tidligere anført foretages en klassificering af vort tekstunivers.

Vi har således som anført foretaget en indholdsmæssig og teksttypologisk strukturering af området, samt en vægtning af omfanget af tekst, der skal tildeles de enkelte emnekategorier i den tematiske inddeling.

Da et hovedmål for korpusprojektet er, at det skal kunne danne udgangspunkt for kontrastive og komparative undersøgelser, har vi lagt vægt på, at den tematiske klassifikation skal være anvendelig på alle tre delkorpora. Endvidere har erfaringerne fra Jurpilotprojektet (jfr. *Ark* 36, 15) vist, at det er vigtigt nøje at overveje, hvor høj en detaljeringsgrad, der er hensigtsmæssig for en sådan tematisk strukturering.

En for høj detaljeringsgrad kan, i de tilfælde hvor de enkelte landes retssystemer divergerer betydeligt, vanskeliggøre indplaceringen af de enkelte tekster i den tematiske struktur.

Konsekvensen af en for høj detaljeringsgrad er, at de enkelte tekster skal tildeles et uhensigtsmæssigt stort antal tematiske kategorier, fordi de enkelte emnekategorier i den tematiske struktur ikke er brede nok til at kunne rumme de aspekter, en tekst typisk vil berøre.

5.2 Den tematiske klassifikation

Ved udarbejdelsen af den tematiske klassifikation har vi taget udgangspunkt i dansk aftaleret og har bestræbt os på at nå frem til en tematisk klassifikation, der skulle være så generel, at den med tilføjelser og udeladelser for de to andre sprogs vedkommende kan bruges som fælles strukturering af alle de tre sprogområders aftaleretlige litteratur, på trods af at de involverede retssystemer er yderst forskellige. Endvidere har vi forsøgt at operere med en klassifikation, hvis kategorier er tilstrækkelig abstrakte og overordnede til at tillade en fornuftig fleksibilitet ved indplaceringen af de enkelte tekster i den tematiske struktur.

I arbejdet med den tematiske inddeling har vi inddraget den universelle decimalklassifikation (UDK), de kategorier som Jens Søndergaard anvender i sin danske juridiske bibliografi samt lærebogsinddelinger.

Disse har vi sammenholdt med franske og engelske inddelinger for at tage højde for, at relevante forskelligheder systemerne imellem kan indpasses i den fælles klassifikation.

For at sikre en korrekt indplacering af de udvalgte tekster i den tematiske struktur, har vi udarbejdet to udgaver af den tematiske

inddeling, nemlig en specificeret og en komprimeret udgave. Den komprimerede udgave indeholder 26 emnekategorier, og det er disse kategorier, der, for at tilgodese omtalte behov for abstrakte og generelle emnekategorier, finder anvendelse ved den tematiske klassifikation af teksterne.

Den specificerede udgave, hvor de 26 emnekategoriers underinddelinger er medtaget, skal tjene som vejledning, idet det kan være vanskeligt for ikke jurister, på grundlag af en meget overordnet klassifikation at fastlå, hvilken emnekategori en tekst, der fx. omhandler et snævert teoretisk delområde, skal tildeles.

De endelige udgaver af inddelingerne har efter udarbejdelsen været til høring hos fire jurister: lektor, lic.jur. Peter Blume, Københavns Universitet, der også var tilknyttet Jurpilotprojektet, lektor, cand.jur. Peter Møgelvang-Hansen, HHK, lektor, cand.jur. Niels Krogh-Hansen, Institut for fransk, HHK, og advokat, cand.jur. et mag. Jette Ronøe, ekstern lektor ved HHK i fransk juridisk sprog. Den komprimerede udgave ser ud som følger:

3%	1.	Aftalekategorier
35%	2.	Aftalens indgåelse og gyldighed
	2.1	betingelser for aftalens tilblivelse
	2.2	aftalers tilblivelse på anden vis
	2.3	aftaleindgåelse ved mellemmænd
	2.4	ugyldighed
	2.5	umulighed
15%	3.	Aftalens retsvirkninger
	3.1	aftalens retsvirkninger mellem parterne
	3.2	aftalens retsvirkninger over for tredjemand
	3.3	trediemandsløfte
5%	4.	Aftaleudfyldning og -fortolkning
22%	5.	Misligholdelse
	5.1	former for misligholdelse
	5.2	retsmidler ved misligholdelse
	5.3	fritagelse for ansvar
	5.4	voldgift
8%	6.	Overgang af ret eller forpligtelse
	6.1	Overdragelse af fordringer
	6.2	novation

6.3 delegation
6.4 subrogation

12% 7. Aftalens ophrør
7.1 opfyldelse
7.2 opsigelse
7.3 opgivelse

Foruden ovenstående tematiske strukturering af vort emneområde skal der for at sikre, at de tekster, der udvælges til at indgå i korpus, giver et repræsentativt billede af vort tekstunivers' sammensætning, foretages en vægtning af de enkelte emnekategoriers tekstmæssige andel af det maskinlæsbare korpus.

Denne vægtning er baseret på flere kriterier, der dog som vist nedenfor hænger nøje sammen. Ved tildelingen af tekstmængder til de enkelte kategorier i klassifikationen tages der således højde for:

- hvilke områder, der er vigtigst inden for aftaleret
- hvilke områder, der giver anledning til størst tekstproduktion
- hvilke områder, der læses mest.

Det skal bemærkes, at inden for juridisk sprog vil disse tre kriterier ofte alle pege på de samme områder, da de områder, der anses for at være de essentielle inden for et retsområde, her aftaleret, alt andet lige vil give anledning til en stor tekstproduktion, som vil blive læst intenst.

På basis af ovennævnte kriterier er vi i samråd med vore jurister nået frem til, at hovedinteressen knytter sig til emnekategorierne **2. Aftalens indgåelse og gyldighed** og **5. Misligholdelse.** Disse to emneområder behandles ofte i såvel domme, som artikler og lærebøger og indgår desuden som centrale komponenter i enhver aftale. De tildeles således henholdsvis 35% og 22% af den samlede teksmængde, der skal indgå i korpus.

Emnekategorien **4. Aftaleudfyldning og fortolkning** vægtes ikke specielt højt (5%). Området giver anledning til en del retssager og dertil hørende domme og anden tekstproduktion, men ikke i et omfang der berettiger til større tekstmæssig dækning.

Kategorierne **3. Aftalens retsvirkninger, 6. Overgang af ret eller forpligtelse** og **7. Aftalens ophør** vægtes ej heller specielt og tildeles henholdsvis 15%, 8% og 12%.

Endelig er kategorien **1. Aftalekategorier** vægtet så lavt som muligt (3%), dog således at det er muligt inden for den tildelte tekstmængde at få alle vore teksttyper repræsenteret. Den lave prioritering af dette område skyldes, at området mest er af teoretisk

interesse, og at det ikke i særlig høj grad er relevant for retstvister, idet det skal bemærkes, at problemer i forbindelse med formbundne og ikke formbundne aftaler falder ind under emnekategorien **2. Aftalens indgåelse og gyldighed.**

6. Bibliografi over aftaleret, fransk og engelsk

6.1 Sammensætning af bibliografien

Da formålet med oprettelsen af disse korpora er at styrke den fagsproglige forskning med henblik på at forbedre kommunikationen i erhvervslivet, har vi, også af hensyn til projektets omfang, valgt at foretage et synkront snit ved at sætte den tidsmæssige ramme for teksterne til en 10-årig periode, nemlig 1978-1987 incl. Dog vil nugældende love med tilhørende lovforarbejder og bekendtgørelser blive medtaget, selv om de er publiceret før denne periode.

Vi har bestræbt os på at udarbejde en emnedækkende bibliografi for den valgte periode, hvorved vi forstår en bibliografi, der - uden at man kan hævde, at den er komplet - må anses for at udgøre en tilstrækkelig og nødvendig tekstdækning af emneområdet.

Den danske bibliografi, der forventes færdiggjort i efteråret 1988, vil primært blive opstillet på grundlag af Jens Søndergaards danske juridiske bibliografi, som er emneinddelt.

I Frankrig og i Storbritannien findes der ingen emneinddelte juridiske bibliografier, hvilket har vanskeliggjort arbejdet med opstillingen af den engelske og den franske emnedækkende bibliografi. Det har først og fremmest været yderst tidskrævende at få sammensat en tilstrækkelig omfattende bibliografi, og dernæst har de mange forskellige kilders forskelligartede opstilling af de bibliografiske oplysninger gjort det nødvendigt at foretage en omfattende omredigering af de enkelte bibliografiske artikler, således at de fremtræder ens i den endelige bibliografi.

6.2 Bibliografiens teksttyper

I det følgende vil især arbejdet med den franske bibliografi blive diskuteret som illustration af de principper, vi har fulgt. Den engelske bibliografi vil blive omtalt, hvor det skønnes relevant at påpege træk, der er særegne for engelsk.

Samlet for den engelske bibliografis vedkommende skal det nævnes, at vi har afgrænset det meget store tekstmateriale til kun at omfatte britisk-engelske tekster. De angelsaksiske retsprincipper

finder udbredt anvendelse i store dele af den engelsk-sprogede verden, men selv om f.eks. en tekst, der er af australsk eller amerikansk oprindelse omhandler et generelt aftaleretligt forhold, er sådanne tekster ikke medtaget af de nævnte afgrænsningsmæssige årsager.

Love og andre regler

Den engelske aftalerets regler er stort set "common law" regler, og som sådan ikke nedfældet i en skreven lov. Der er enkelte tilfælde, hvor det er blevet fundet nødvendigt eller hensigtsmæssigt med nye lovgivningsmæssige bestemmelser, f.eks. om mindreåriges kontrakter, men dette er undtagelsen snarere end reglen. (Love som Sale of Goods Act, den engelske købelov, er naturligvis ikke medtaget i bibliografien, da den hører til den specifikke del af aftaleretten.)

En af konsekvenserne af den manglende nedskrevne lov er også, at der ikke findes (ministerielle) bekendtgørelser for området.

For den franske bibliografis vedkommende frembyder teksttypen love og andre regler ingen vanskeligehder, da civillovgivningen i Frankrig er samlet i en lovbog, nemlig i Code Civil. Således er samtlige de lovparagraffer, der omhandler den generelle del af aftaleretten i Frankrig optaget i Code Civil, og nye love og ændringer på området indføjes i de årlige udgaver af lovbogen. De øvrige teksttyper har derimod voldt os større problemer.

Lovforarbejder

Teksttypen lovforarbejder vil blive udmøntet i en række kommissionsbetænkninger, idet de to andre mulige eksempler på lovforarbejder, lovforslag og den nedskrevne debat i de 3 involverede landes parlamenter, ikke skønnes egnede til vores korpusformål. Lovforslag ligger tekstuelt helt tæt op ad loven, og parlamentsdebatterne er ikke en tekst, der er ment som en skreven sammenhængende tekst. Forskellige typer rapporter og betænkninger fra udvalg eller kommissioner er derimod vigtige tekster i forbindelse med en lovs historie, og samtidig tror vi, at de tekstuelt har visse særpræg.

Juridiske lærebøger

Den franske bibliografi over teksttypen juridiske lærebøger er

sammensat på grundlag af forlagskataloger fra praktisk talt samtlige juridiske forlag, suppleret med værker fra enkelte eksisterende bibliografiske værker (Bibliographie juridique générale 1986, Dalloz-Sirey) samt fra kartotekerne på det juriske hovedbibliotek i Paris, Bibliothèque spécialisée Droit et Science Economique, 2 rue Cujas.

Ud over de anerkendte lærebøger findes der en række værker beregnet enten til en hurtig og overfladisk gennemgang af emnet, f.eks. med henblik på studerendes eksamenslæsning eller som en introduktion til emnet. Det er vores skøn, at disse bøger ikke har den fornødne juridiske kvalitet, som vi har valgt at prioritere højt. Engelske lærebogsværker, der især behandler emnet ud fra "cases" eller udelukkende består af "cases", er heller ikke medtaget i bibliografien, da vi ikke ønskede disse i det maskinlæsbare korpus i og med at sådanne "case"-tekster er forkortede og bearbejdede udgaver af domme afsagt af domstolene.

Domme og artikler i juridiske tidsskrifter

Vi har opstillet den franske bibliografi over teksttyperne domme og artikler i juridiske tidsskrifter ved at gennemgå årsoversigterne fra perioden 1978 til 1987 incl. i de tre vigtigste franske juridiske tidsskrifter, nemlig *Semaine Juridique*, Editions Techniques, *Recueil Dalloz*, Jurisprudence générale Dalloz og *Revue trimestrielle de droit civil*, Sirey.

Vi har medtaget alle de titler, som de tre tidsskrifter har rubriceret under emnekategorien: contrats et obligations, eftersom de domme og juridiske artikler, der af tidsskrifternes redaktører er opført under denne kategori, må anses for primært at omhandle aspekter af generel aftaleretlig karakter. Domme og artikler, der er rubriceret under de specifikke aftaleforhold, f.eks. contrat de travail er ikke medtaget i de respektive bibliografier. Det skyldes, at selv om der nødvendigvis må indgå visse elementer af generel aftaleretlig karakter i sådanne domme og artikler, må de siges at falde uden for vort tekstunivers.

Hvad angår teksttypen juridiske artikler, har vi, for at sikre at bibliografien kommer til at indeholder juridiske artikler fra så mange tidsskrifter som overhovedet muligt, suppleret med artikler angivet i det bibliografiske værk: Index to Foreign Legal Periodicals.

Opstillingen af bibliografien over teksttypen domme har givet anledning til særlige problemer. Det skyldes, at de enkelte juridiske tidsskrifter ikke anfører parternes navne på dommene, men blot

forsyner dem med stikord og rubricerer dem emnemæssigt i årsoversigterne efter det hovedtema, som tidsskriftets redaktør skønner, at dommen angår.

Eftersom de enkelte stikord er baseret på hvilke delemner de enkelte domme skønnes at angå, og den emnemæssige rubricering også er foretaget på grundlag af et skøn fra de forskellige tidsskriftsredaktørers side, er det vanskeligt at undgå, at en eller flere domme bliver medtaget flere gange i vor bibliografi. Angivelsen af den domstol, der har afsagt de enkelte domme, og dommens dato kan ej heller anvendes som rettesnor for frasortering af eventuelle "gengangere", da der afsiges adskillige domme hver dag og da parternes navne som nævnt ikke er angivet i de bibliografiske årsoversigter.

Kun ved at finde hver enkelt dom frem og sammenligne dem, vil vi kunne frasortere alle gengangere. Da der imidlertid er tale om et meget stort antal domme, vil dette være praktisk umuligt, og vi mener ikke, at ulempen med "gengangere" i bibliografien er væsentlig.

I det maskinlæsbare korpus, hvor alle domme er gennemlæst inden indlæsning, vil der naturligvis ikke forekomme "gengangere".

Engelske jurister har adgang til en række forskellige domssamlinger, hvoraf vi dog har valgt kun at medtage et værk the All England Law Reports i bibliografien, idet der ikke hverken i udvalget af domme eller den sproglige rapporteringsform er væsentlige forskelle.

Den engelske bibliografi over artikler i juridiske tidsskrifter er etableret på basis af ti årgange af Index to Legal Periodicals, som imidlertid indeholder titler fra hele den engelsk-sprogede verden. Af disse titler er så excerperet dem, der har været publiceret i britiske tidsskrifter.

Kontrakter

Endelig gør der sig helt specielle forhold gældende for teksttypen kontrakter. Det er en komplet umulighed at etablere en emnedækkende bibliografi over kontrakter. Dette skyldes, at der dels findes et umådeligt stort antal kontrakter, dels at de fleste kontrakter er private og ikke kommer til offentlighedens kendskab. I langt de fleste tilfælde kender kun kontrahenterne indholdet.

Bibliografien over teksttypen kontrakter vil således primært omfatte publicerede formularsamlinger og fortrykte standardkontrakter. Endvidere vil den omfatte et mindre antal af aftaler, som vi

har fået stillet til rådighed af velvillige forretningsfolk og advokater.

6.3 Søgemuligheder i bibliografien

Det er besluttet, at de - indtil videre - tre grundlæggende bibliografier, der udgør projektets tekstunivers, skal indlæses i en database og således være tilgængelige for on-line søgning. Der er til dette formål udarbejdet et format for den bibliografiske record, der indeholder oplysninger om:

> Forfatter, titel, tidsskrift (navn, nr. datering), redaktion/domstol, udgave, udgivelsessted, forlag, udgivelsesår, teksttype samt angivelse af hvor vidt den pågældende titel er repræsenteret i det maskinlæsbare tekstkorpus.

Den endelige bibliografiske database vil gøre det muligt at gennemføre selektive søgninger fx. på tekster repræsenteret i det maskinlæsbare tekstkorpus eller på en eller flere af bibliografiens teksttyper i et eller flere af de repræsenterede sprog, ligesom den vil gøre det muligt for eksterne brugere at vurdere og anvende korpus.

7. Udvælgelse af tekster til det maskinlæsbare tekstkorpus

7.1 Principper for valg af tekster

Hovedformålet med oprettelsen af disse korpora er som nævnt at tilvejebringe et grundlag for forskellige empiriske undersøgelser, f.eks. tekstlingvistiske, teksttypologiske og syntaktiske undersøgelser. Dette formål er bestemmende for principperne for udvælgelsen af de tekster fra den emnedækkende bibliografi, der skal indgå i det maskinlæsbare korpus.

Ved en tekst forstår vi i denne forbindelse en relativt selvstændig, indholdsmæssig kohærent følge af sproglige ytringer. Således anses en dom, en kontrakt, en tidsskriftartikel eller et kapitel i en lærebog for at være én tekst, som medtages i sin helhed.

Vi vil dog bestræber os på, at tekstlængden ikke overskrider 5.000 løbende ord.

Vi opererer som tidligere nævnt med 6 teksttyper:

- love og andre regler
- lovforarbejder

- domme
- kontrakter
- juridiske lærebøger
- artikler i juridiske tidsskrifter

Da disse teksttyper i lige høj grad skal kunne være genstand for undersøgelser af forskere, sikres hver teksttype ved tekstudvælgelsen i princippet ligelig repræsentation, dvs. 16,66% af den samlede tekstmængde.

Men på grund af retssystemernes forskellighed vil det i nogle tilfælde ikke være muligt at finde tilstrækkeligt tekstmateriale til at gennemføre dette princip. Dette er eksempelvis tilfældet med teksttypen love og andre regler i det engelske korpus. Den således tiloversblevne "plads" vil blive anvendt til specielt interessante teksttyper, f.eks. kontrakter eller artikler.

Teksterne skal endvidere udvælges således, at hvert af klassifikationens temaer repræsenteres i overensstemmelse med den fordelingsnøgle, der er fastsat i samråd med jurister, nemlig:

- aftalekategorier 3%
- aftalens indgåelse og gyldighed 35%
- aftalens retsvirkninger 15%
- aftaleudfyldning og -fortolkning 5%
- misligholdelse 22%
- overgang af ret eller forpligtelse 8%
- aftalens ophør 12%

Efter at have vægtet teksttyperne og de enkelte temaer har vi regnet ud, hvor stor en procentdel af det samlede korpus, der skal dækkes af de forskellige teksttyper sammenholdt med temaerne.

Hvert korpus vil omfatte 1 mio. løbende ord, dvs. ca 3333 ns. På basis af den procentvise fordeling har vi udregnet fordelingen af teksterne i korpus angivet i normalsider. Disse beregninger anvender vi ved den konkrete udvælgelse af de tekster, der skal indgå i det maskinlæsbare korpus.

7.2 Kodering

De valgte tekster bliver derpå tekstuelt koderet, dvs. forsynet med bibliografiske oplysninger, angivelse af teksttype og tema fra klassifikationen, således at det bliver muligt at søge på bestemte teksttyper og temaer.

En lingvistisk kodering (dvs. en markering af f.eks. grammatiske kategorier) i lighed med den kodering, der blev foretaget for visse dele af Jurpilotprojektet, vil ikke blive gennemført. Vi har vurderet, at en generel lingvistisk kodering vil være af begrænset nytte for den enkelte forsker, samtidig med at den er meget ressourcekrævende.

Hvis en forsker ønsker en lingvistisk kodering, vil det være mere hensigtsmæssigt at foretage koderingen efterfølgende i overensstemmelse med projektets særlige behov.

8. Fremtidsperspektiver

Med oprettelsen af disse korpora imødekommes behovet for et fagsprogligt empirisk arbejdsgrundlag, og vi ser mange muligheder for spændende projekter inden for en række af lingvistikkens områder. Det vil især være oplagt at arbejde med kontrastive projekter både for specialeskrivende studerende og for sprogforskere generelt. De principper, der fastlægges for de ovenfor beskrevne korpora inden for initiativområdet, vil forhåbentlig kunne anvendes fremover, såfremt der findes midler til udvidelser, enten ved at de eksisterende korpora udbygges med den specifikke del af aftaleretten, eller ved at andre fag- og sprogområder inddrages.

Litteratur

Bergenholtz, Henning (1988): Korpusproblematik in der Computerlinguistik. Konstruktionsprinzipien und Repräsentativität, i: *Computational Linguistics. Ein internationales Handbuch computerunterstützter Sprachforschung und ihrer Anwendung*. Hrsg. von Istvan Batori u.a. Berlin/New York: de Gruyter (i trykken).

Bergenholtz, Henning/Schaeder, Burkhard (Hrsg.) (1979): *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*. Königstein/Ts 1979.

Hansen, Steffen Leo (1987): *Jurkorpus 3: Jurkorpus, indhold - omfang - struktur*. Terminologiafdelingen. Handelshøjdkolen i København 1987.

Hansen, Steffen Leo (1987): *Jurkorpus 5: Korpuslingvistik, teori - metode - praksis*. Terminologiafdelingen. Handelshøjskolen i København 1987.