

MASKINLÆSBARE TEKSTKORPORA I DANMARK

Af Karen M. Lauridsen (Handelshøjskolen i Århus)

Nedenstående liste beskriver i kort form de større maskinlæsbare tekstkorpora, som foreligger i Danmark. Originalteksterne er alle skrevne.

DANWORD

Etableret af: Bente Maegaard og Hanne Ruus (Københavns Universitet)

Korpus er indsamlet i perioden: 1975-80

Tekstmaterialet stammer fra perioden: 1970-74

Sprog (variation): dansk (skrevet)

Omfang: 1,25 millioner ord

Detaljer vedr. materialet: 5 x 1.000 tekstudsnit på hver 250 ord

Organisation af materialet: 5 kategorier på hver 250.000 ord: romaner, børnebøger, aviser, ugeblade og fagblade

Analyse: frekvenslister, konkordanser

Anvendelsesområde: frekvenslister, leksikalske studier

Adgang til korpus: kan anvendes til forskningsformål

Lagring af data: ASCII-filer på magnetbånd (KU)

Øvrige kommentarer: frekvenslister og en mere detaljeret beskrivelse af DANWORD findes i Maegaard, Bente & Hanne Ruus (1981a) Hyppige ord i danske børnebøger. København: Gyldendal. (1981b) Hyppige ord i danske romaner. København: Gyldendal. (1986) Hyppige ord i danske aviser, ugeblade og fagblade. København: Gyldendal. (1987) "The Compilation and Use of a Text Corpus" i Studies in Honour of Roberto Busa S.J. (pp. 103-122). Pisa.

DK87, DK88, DK89 og DK90

Etableret af: Henning Bergenholtz (Handelshøjskolen i Århus)
Korpus er indsamlet i perioden: 1987-88; 1988-89; 1989-90; 1990-91

Tekstmaterialet stammer fra perioden: 1987; 1988; 1989; 1990

Sprog (variation): dansk (skrevet)

Omfang: 4 x 1 million ord

Detaljer vedr. materialet: for hvert år, dvs. for hver 1 million ord, 200 tekster af ca. 5.000 ord

Organisation af materialet: 50 % fiktion, 25 % aviser, 25 % ugeblade. Alle teksterne er originale (ikke-oversatte) danske tekster, som blev udgivet for første gang i henholdsvis 1987, 1988, 1989 og 1990

Analyse: frekvenslister, konkordanser

Anvendelsesområde: forskellige former for lingvistisk, herunder også tekstlingvistisk, analyse

Adgang til korpus: teksterne foreligger som ASCII filer og er tilgængelige til forskningsformål for danske og udenlandske lingvister. Kopi på disketter fås ved henvendelse til professor Henning Bergenholtz, Handelshøjskolen i Århus, Fuglesangs alle 4, 8210 Århus V

Lagring af data: 3 1/2" eller 5 1/4" disketter (IBM eller dermed kompatible) samt Apple Macintosh

Øvrige kommentarer: nærmere beskrivelse findes i Bergenholtz, Henning (1988) "DK87: Et korpus med dansk almensprog" i Hermes 1/1988. Det erhvervssproglige Fakultet, Handelshøjskolen i Århus.

DANSK-ENGELSK-FRANSKE KORPORA I AFTALERET

Etableret af: Gunhild Dyrberg, Dorrit Faber, Steffen Leo Hansen, og Joan Tourney (Handelshøjskolen i København) og Karen M. Lauridsen (Handelshøjskolen i Århus)

Korpus er indsamlet i perioden: 1987-1989

Tekstmaterialet stammer fra perioden: 1978-1987

Sprog: dansk, engelsk og fransk (skrevet)

Omfang: 3 x 1 million ord

Detaljer vedr. materialet: originale britisk engelske, franske og danske tekster inden for emneområdet aftaleret; hvert enkelt delkorpus består af ca. 200 tekster på hver 5.000 ord

Organisation af materialet: for hvert sprog (for hvert korpus på 1 million ord) er materialet inddelt i følgende teksttyper: 1. Love og andre regler, 2. Lovforarbejder, 3. Domme, 4. Kontrakter, 5. Juridiske lærebøger, 6. Artikler i juridiske fagtidsskrifter. En nærmere beskrivelse af prin-

cipperne for distributionen af tekstmaterialet i de seks kategorier findes i Dyrberg, Gunhild, Dorrit Faber, Steffen Leo Hansen og Joan Tournay (1988) "Etablering af et juridisk tekstkorpus" i Hermes 1/1988. Det erhvervssproglige Fakultet, Handelshøjskolen i Århus

Analyse: frekvenslister; øvrige lingvistiske analyser under udarbejdelse

Anvendelsesområde: forskellige former for lingvistisk, herunder også tekstlingvistisk, analyse

Adgang til korpus: teksterne foreligger som ASCII filer og er tilgængelige til forskningsformål for danske og udenlandske lingvister ved henvendelse til Handelshøjskolen i Århus (Ole Lauridsen, Tysk institut), Fuglesangs alle 4, 8210 Århus V

Lagring af data: 3 1/2" eller 5 1/4" disketter (IBM eller dermed kompatible) samt Apple Macintosh

Øvrige kommentarer: Alle tekster er forsynet med information om kilde, teksttype, tema, osv.

JUR-korpora

Etableret af: Inge Gorm Hansen, Steffen Leo Hansen, Ole Helmersen, Bodil Nistrup Madsen, Hanne Puggaard, Joan Tournay, Charlotte Werther, Anne Zoëga (Handelshøjskolen i København)

Korpus er indsamlet i perioden: 1985-87

Tekstmaterialet stammer fra perioden: 1970ff

Sprog (variation): dansk, engelsk og spansk (skrevet)

Detaljer vedrørende materialet: 385.000 ord på dansk; 150.000 ord på engelsk; 402.000 ord på spansk; alle tekster tilhører emneområdet civilprocessen

Organisation af materialet: 1095 tekst på dansk; 518 tekster på engelsk; 1219 tekster på spansk

Analyse: det danske og det engelske materiale er delvis koderet m.h.t. ordklassekategorier

Anvendelsesområde: pilotprojekt vedrørende database til terminologisk information og generering af ordbøger

Adgang til korpus: on-line på Handelshøjskolen i København (Terminologiaafdelingen)

Lagring af data: database, status II

Øvrige kommentarer: projektet er beskrevet i ARK 36 (1986) Pilotprojekt vedrørende database til terminologisk information og generering af ordbøger. Terminologiaafdelingen, Handelshøjskolen i København.

DANSK-TYSK-SPANSK KORPORA I BIOTEKNOLOGI

Etableret af: Det tyske og det danske korpus: Bent Hauschildt, Ole Lauridsen og Henning Søndergaard (Handelshøjskolen i Århus) samt Theis Riiber (Handelshøjskolen i København). Det spanske korpus: Søren Kaas Andersen og Heribert Picht (Handelshøjskolen i København) samt Leon Jørgensen (Handelshøjskole Syd)

Korpus er indsamlet i: 1989

Tekstmaterialet stammer fra perioden: 1979-1989

Sprog: dansk, tysk og spansk (skrevet)

Omfang: 3 x 1 million ord

Detaljer vedrørende materialet: alle tekster tilhører emneområdet bioteknik og omfatter i alt væsentligt max. 5.000 ord (+ 10%)

Organisation af materialet: materialet er organiseret i fire grupper i henhold til afsender-modtager forholdene:

- I. ekspert til ekspert
- II. a. ekspert til lægmand
b. lægmand til ekspert
- III. lægmand til lægmand

Analyse: de første analyser er under udarbejdelse; der vil bl.a. foreligge frekvenslister

Anvendelsesområde: forskellige former for lingvistisk, herunder også tekstlingvistisk analyse

Adgang til korpus: teksterne foreligger som ASCII filer og er tilgængelige til forskningsformål for danske og udenlandske lingvister. Kopi af disketter fås ved henvendelse til Handelshøjskolen i Århus (Ole Lauridsen, Tysk institut), Fuglesangs alle 4, 8210 Århus V

Lagring af data: 3 1/2" eller 5 1/4" disketter (IBM eller dermed kompatible)

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