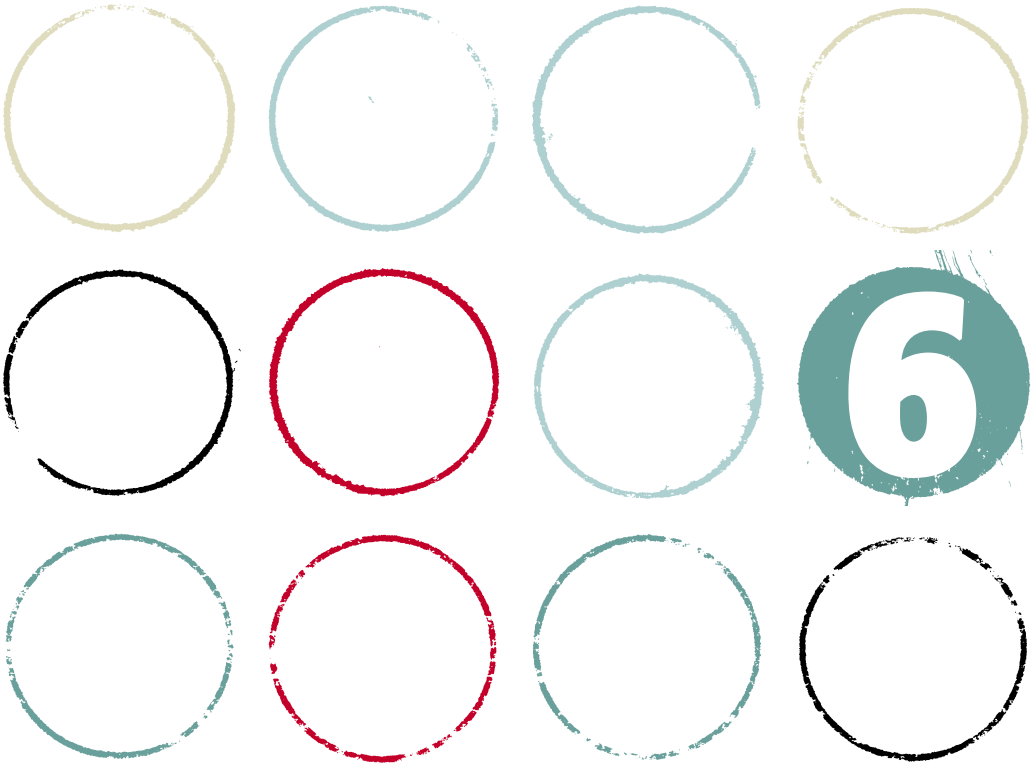
Forskningsrådet for Kultur og Kommunikation

STRATEGISK SATSNING PÅ
DANSK SPROGTEKNOLOGI



STRATEGISK SATSNING PÅ **DANSK SPROGTEKNOLOGI**

Udarbejdet af: **Bente Maegaard**, Københavns Universitet,
Center for Sprogteknologi, (formand)
Eckhard Bick, Syddansk Universitet
Paul Dalsgaard, Aalborg Universitet
Sabine Kirchmeier-Andersen, Copenhagen Business School
Ole Togeby, Aarhus Universitet
Berit Heer Henriksen, Københavns Universitet,
Center for Sprogteknologi, (sekretær)

TEMARAPPORTER FRA FORSKNINGSRÅDET FOR KULTUR OG KOMMUNIKATION

1. KULTURENS FREMTID – ÆSTETIK UDEN GRÆNSER
2. HUMANISTISK VIDEN I ET VIDENSAMFUND
3 LÆRING – KULTUR OG SUBJEKTIVITET
4. OMVERDEN, INDIVID OG SAMFUND – HUMANISTISK NATURFORSKNING
5. RELIGION – KULTUR – DEMOKRATI
6. STRATEGISK SATSNING PÅ DANSK SPROGTEKNOLOGI
7. BAG TALLENE – HUMANISTISK VELFÆRDSFORSKNING
8. ORDETS MAGT – HUMANISTISK FORSKNING I SPROG OG INDFLYDELSE
9. KULTUR OG SUNDHED – HUMANISTISK FORSKNING I KROP, SUNDHED OG SYGDOM
10. KOST OG KULTUR – HUMANISTISK FØDEVAREFORSKNING
11. IKT: ET HUMANISTISK ANLIGGENDE – TEKNOLOGI, MEDIUM, KOMMUNIKATION
12. UDFORDRINGER OG MULIGHEDER FOR HUMANISTISK FORSKNING

Strategisk satsning på
dansk sprogteknologi
2. oplag
Udgivet af: Forskningsrådet for Kultur og Kommunikation
Grafisk design: Marianne Dunker
Tryk: Clichéfa Tryk AS

ISBN: 87-90201-69-8

Indhold

Forord

Hermed fremlægger forskningsrådet en række tilbundsgående analyser af forskningstilstand og –behov på en række udvalgte humanistiske forskningsfelter i form af i alt 11 temarapporter.

Formålet med rapporterne er at styrke humanioras grænseflader til andre videnskaber og kontakten til det øvrige samfund. Det er frugten af et intenst arbejde med at nyformulere humaniora, som nu kommer til udtryk i temarapporterne.

I stedet for en fagdisciplinær tilgang har forskningsrådet udvalgt en række temaer. De enkelte temagrupper blev sammensat af håndplukkede forskere fra en række relevante fag. Mens temarapporterne så vidt muligt skulle skrives som helhedsblik over temaet, var det ikke et krav til arbejdsgrupperne, at de skulle dække alle relevante forskningsgrupper og -tilgange. Arbejdsgrupperne havde også fuld frihed til selv at vurdere og prioritere. Forskningsrådet modtog rapporterne til gennemsyn inden offentliggørelse og havde mulighed for at kommentere på dem, men de fremlægges uden at de på alle punkter skal tages som udtryk for forskningsrådets vurdering.

Rapporten Strategisk satsning på dansk sprogteknologi blev udarbejdet på bestilling af Statens Humanistiske Forskningsråd og første gang fremlagt i 2004. De første rapporter blev hurtigt udsolgt, og Forskningsrådet for Kultur og Kommunikation genudgiver derfor de gamle rapporter sammen med de nye.

Vi håber at rapporterne vil stimulere til ny forskning og samarbejde på tværs af fag og institutioner og stimulere dialogen med samarbejdspartnere uden for forskningsinstitutionerne.

Med venlig hilsen

Poul Holm
Formand
Forskningsrådet for Kultur og Kommunikation

Resumé

BEHOV FOR DANSK SPROGTEKNOLOGI Sprogteknologi (som omfatter taleteknologi og natursprogsbehandling) er et forskningsfelt, der meget hurtigt kan omsættes i produkter, som har stor betydning for alle i samfundet. Blandt de vigtigste kan nævnes: talesyntesemaskiner, der automatisk læser en skrevet tekst op; talegenkendelsesmaskiner, der kan omsætte en mundtlig tekst til skrift og således f.eks. tekste fjernsynsudsendelser løbende; intelligente dialogiske søgemaskiner, der efter mundtlig forespørgsel ikke blot kan finde steder på internettet, der indeholder oplysninger, men også selve oplysningerne; oversættelsesmaskiner, der oversætter skrevne eller mundtlige tekster fra et sprog til et andet; grammatik- og stilprogrammer, der tjekker korrekthed mv.

Effektiv sprogteknologi vil også kunne fremme det erklærede formål om parallelsproglighed, som er foreslået af Dansk Sprognævn, således at domænetab og manglende udbredelse af dansk forskning kan imødegås ved mere oversættelse til og fra dansk og ved, at det gøres nemmere at skrive på dansk.

Grundlaget for disse teknologier er til stede for de store sprog, og de er udviklet i en sådan grad, at de allerede anvendes i et vist omfang, men fordi Danmark er et lille sprogområde, er der ikke investeret tilsvarende i dansk sprogteknologi, og de systemer, der findes for dansk, er ikke nær så gode som for de store sprog.

Aktørerne i det sprogteknologiske landskab Der er ellers i Danmark meget gode muligheder både for at udvikle dansk sprogteknologi og for at tage det i anvendelse. Der er fire steder i landet (Aalborg Universitet, Syddansk Universitet, Københavns Universitet og Handelshøjskolen i København) sprogteknologiske forskningsinstitutioner med forskning på et højt niveau (men ikke i så stort omfang), og med et vist indbyrdes samarbejde, og Danmark er med mange pc'er, mange internetbrugere og højt uddannelsesniveau modent til at tage informationsteknologi i anvendelse. Men der er slet ikke investeret de samme beløb i sprogteknologi for dansk som for de store sprog, selvom det selvfølgelig koster det samme for dansk som for engelsk. For store udenlandske sprogteknologiske virksomheder er det ikke tillokkende at investere i et så lille sprog som dansk, og af danske sprogteknologiske virksomheder er der kun få og små. Derfor må samarbejdet mellem universiteterne og de private virksomheder styrkes og udbygges betydeligt. Danmark er, som en nylig international sammenligning har påpeget, lovende, men ikke førende inden for sprogteknologi. Større investeringer, både offentlige og private, er nødvendige for at sætte udviklingen i gang.

Sprogressourcer, basisteknologi og integreret teknologi Som forskningsfelt kan sprogteknologi på den ene led deles op i talesprogteknologi og (den øvrige) natursprogbehandling, og på den anden led i dataressourcer (tale- og tekstkorpusser), teknologiske ressourcer (databaser over danske lyd og ord og grammatikker over sætningstyper og teksttyper), basale teknologier (segmentering af løbende sprogligt input, konvertering mellem lyd og bogstav, parsere, der analyserer inputtet grammatisk og indholdsmæssigt) og integreret teknologi (maskiner og programmer, der kan læse tekster op, genkende tale, lave tekstredigering (f.eks. stavetjek) og automatisk klassificere og oversætte mellem flere sprog. Sprogressourcerne og den basale teknologi må nødvendigvis være til stede, for at man kan udvikle de integrerede anvendelser, men selvom der forskes flere steder i Danmark i alle dele af feltet, er ikke engang sprogressourcerne i dag til stede i tilstrækkeligt omfang.

Udfordringer og fokusområder Der er især fire områder, hvor der er brug for en styrkelse af forskningen: flere og større tale- og tekstkorpusser, robust talegenkendelse, bedre (mere generelle og robuste) parsere, oversættelse til og fra dansk og satsning på en avanceret og visionær anvendelse på dansk. Der er i Danmark forskningsmæssigt grundlag for, at vi kan blive førende på disse områder. Derfor foreslår vi en satsning på sprogteknologi. Kun dermed kan også danskere få glæde af sprogteknologiens muligheder.

En stor strategisk satsning

De investeringer, der er brug for i sprogteknologi, er i følgende størrelsesorden:

Tekst- og talekorpusser:	10 mio. kr.
Talegenkendelse:	10 mio. kr.
Grammatikker og parsere:	7,5 mio. kr.
Oversættelsesmaskiner:	20 mio. kr.
Visionær og avanceret integreret teknologi:	20 mio. kr.
Organisering af et tættere netværk mellem forskerne og et konsortium til samarbejde med erhvervslivet:	2,5 mio. kr.
I alt	70 mio. kr.

1. Behov for god dansk sprogteknologi

Sprogteknologi stifter vi bekendtskab med i mange forskellige sammenhænge, f.eks. når vi benytter stavekontrol i et tekstbehandlingsprogram, eller når vi bestiller biografbilletter gennem en automatisk interaktiv telefontjeneste. I fremtiden vil sprogteknologi komme til at spille en større og større rolle i moderne menneskers hverdag, men det er ikke sikkert, at danske brugere kan anvende deres danske sprog, når de betjener teknologien. Det vil være en kulturel falliterklæring, hvis vi er henvist til at bruge f.eks. engelsk eller helt undvære den nye teknologi, men det kan blive vilkårene, hvis der ikke satses intensivt på dansk sprogteknologisk forskning og udvikling.

Samtidig er forskning i sprogteknologi en kilde til vigtig, ny erkendelse inden for forskellige humanistiske fag og i tværfaglige forskningssamarbejder mellem humaniora og teknik.

1.1 SPROGTEKNOLOGI I DAG OG I FREMTIDEN Sprogteknologi inkorporerer information om menneskers sprog, og computere kan derigennem genkende, forstå, tolke og efterligne det menneskelige sprog i dets forskellige former. Sprogteknologi omfatter taleteknologi (maskinel behandling af talt sprog: at genkende og producere talt sprog) og natursprogsbehandling (maskinel behandling af skrevet sprog).

Sprogteknologi forenkler og forbedrer kommunikationen mellem menneske og maskine og hjælper mennesker til at kommunikere med hinanden. Næsten al information i dag behandles, lagres og fremfindes af computere, og uhyre meget (skriftlig) kommunikation foregår via computere. Derfor har sprogteknologi meget stor betydning for moderne menneskers indbyrdes kommunikation.

Behovet for god dansk sprogteknologi finder vi både hos borgerne, i erhvervslivet, i den offentlige forvaltning og i samfundet som sådan. Sprogteknologi er et hjælpemiddel i hverdagen, både for brede brugergrupper og for grupper med særlige behov, f.eks. handicappede. Teknologien kan effektivisere arbejdsprocesser markant, bl.a. ved at lette håndteringen af store informationsmængder.

Den sprogteknologiske forskning og udvikling giver grundlag for innovation og nye indtjeningsmuligheder for danske virksomheder. Landvindinger inden for sprogteknologi kan stimulere til innovation på mange forskellige områder, der har at gøre med sprog og kognition, f.eks. e-handel, e-læring (læring via computer, fjernundervisning m.m.) og computerspil. Dybest set er sprogteknologi et udtryk for drømmen om at kunne kommunikere med computeren på samme ubesværede måde som med andre mennesker. På den måde har sprogteknologi at gøre med kunstig eller indlært intelligens: Med denne teknologi kan vi forbedre computerens evne til at håndtere menneskets skrift, tale og kognition.

I dag har vi computere med så stor regnekraft, at det er realistisk at forestille sig, at maskinerne bliver i stand til at udføre de komplekse beregninger, der skal til, for at de kan lære at forstå, hvad vi beder om.

Fremtidsscenariet „pervasive computing" indebærer computeres indlejring i dagligdags genstande, f.eks. vægge, køleskabe og kaffemaskiner, og deres interaktion med mennesker og hinanden. Computerne vil interagere med mennesker på en mere integreret og menneskelig måde, end vi kender i dag. Integreret sprogteknologi gør det allerede nu muligt at få adgang til en computer i situationer, hvor man normalt ikke har hænderne frie, og i fremtiden kan dette bl.a. få betydning for læger og sygeplejersker, der i en operationsstue er travlt optaget med håndtering af patient og instrumenter: Via talesyntese (computerproduceret tale) kan en computer løbende give nuancerede oplysninger om patientens tilstand, f.eks. hans blodtryk og hjerte-

funktion, og ved hjælp af en talegrænseflade (interaktivt dialogsystem) kan læger og sygeplejersker skaffe sig adgang til patientens journal, få læst oplysningerne op og indtale nye oplysninger, som gemmes i journalen.

I det semantiske web kan sprogteknologien revolutionere fremtidens søgemaskiner ved at lære dem sprog, så de kan forstå, hvad vi søger, og besvare vores spørgsmål med fuldgyldige svar formet i naturlige sætninger. Spørger vi f.eks. en computer, hvilket år H.C. Andersen blev født, vil en søgning på nettet ikke kun besvares med henvisninger til et antal dokumenter. Computeren fortæller os simpelthen, at H.C. Andersen blev født i 1805, efter at have fundet oplysningen på nettet.

Tolkemaskiner vil i fremtiden kombinere talegenkendelse, talesyntese og oversættelse, og det vil give helt nye muligheder for kommunikation på tværs af landegrænser.

Alt dette udvikles imidlertid ikke nødvendigvis for dansk. Allerede i dag er udbuddet af dansk sprogteknologi relativt begrænset. Brugere med verdenssprogene engelsk og tysk er meget bedre stillet ifølge en nylig benchmarkundersøgelse, der sammenligner europæiske lande med hensyn til sprogteknologisk vareudbud, forskning, infrastruktur m.m. (*Benchmarking HLT progress in Europe (2003)*). Til hjælp for tekstbehandling findes der f.eks. stave- og grammatiktjekprogrammer for engelsk, der er meget bedre end for dansk. Der findes i højere grad generelle engelske programmer til informationssøgning, tekstresumering og interaktive dialogsystemer (talestyret menneske/maskinegrænseflade). Der findes programmer til maskinoversættelse (automatisk oversættelse mellem sprog), der laver bedre råoversættelser end de programmer, man kan finde for dansk. For de store sprog findes ikke blot generel talegenkendelse (computerens genkendelse af menneskelig tale), men også integrerede systemer til diktering, således at man kan diktere en tekst direkte ind i et dokument og derved slippe for at taste.

Danskerne har ligesom andre landes borgere en demokratisk ret til

at skaffe sig, forstå, bearbejde og producere information. Men små sprogområder som det danske risikerer let at blive tabere i et fremtidigt Europa, der ifølge rapporten om sprogteknologi i Europa er i fare for at udvikle sig i to hastigheder, hvad sprogteknologi angår. Danskerne skal sikres det høje sprogteknologiske niveau, som de har behov for, og som der er forskningsmæssigt potentiale til at udvikle. Sprogteknologi er et sprogpolitisk redskab, der bl.a. kan sikre, at danske borgere kan begå sig i det moderne globale it- og netværkssamfund.

1.1.1 Et hjælpemiddel i dagligdagen Sprogteknologi er et stadigt vigtigere og mere integreret hjælpemiddel i vores dagligdag. For handicappede og andre grupper med særlige behov kan sprogteknologi betyde muligheder, som man ellers var afskåret fra. Et interaktivt dialogsystem betyder f.eks., at det for visse grupper bliver muligt eller lettere at betjene en computer og derigennem tilegne sig information m.m. Omvendt kan it skabe voksende ulighed i samfundet, hvis teknologien ikke gøres tilgængelig for alle, inklusive personer der er ordblinde eller ikke læser så godt, ikke behersker fremmedsprog, har nedsat syn, er blinde, bevægelseshandicappede, afasiramte osv. Ca. 100.000-150.000 af de handicappede personer i Danmark har vanskeligt ved at benytte it-hjælpemidler, anslås det i rapporten *National Profile and LE Opportunity Map. Denmark* (1998).

Ofte betyder udviklingen af sprogteknologi, at det for både brede grupper og særlige grupper bliver lettere at kommunikere gennem computer eller telefon.

Sprogteknologi opfylder i dag behov som hjælpemiddel for mange forskellige brugergrupper i Danmark:

- Man kan udnytte talegenkendelsesprogrammer og ordprædiktion (det at f.eks. mobiltelefonen udfylder hele ordet efter ganske få indtastninger) til henholdsvis talestyret opringning og SMS i sin

mobiltelefon.

- Man kan få oplyst en persons navn, stilling og adresse med talesyntese efter at have tastet personens telefonnummer. Det sker gennem oplysningstjenesten på nummer 1811.
- Man kan køre bil og navigere og ved hjælp af bl.a. talesyntese løbende få oplysninger om navigationen i retning af et givet mål.
- Man kan modtage internetbaseret fjernundervisning i f.eks. grammatik eller sprog.
- Man kan afmærke tekst på en vilkårlig dansk hjemmeside, trykke på en afspilleknap og herefter få teksten læst højt med talesyntese. Denne service findes på adressen www.adgangforalle.dk.
- Man kan få oplæst sine e-mails ved hjælp af talesyntese.
- Man kan benytte hjælpeværktøjer, der på forskellig vis letter læse- og skriveprocessen for funktionshæmmede og personer med læse- og skrivevanskeligheder.

1.1.2 Effektivisering af arbejdsprocesser Den sprogteknologiske udvikling åbner mange muligheder for effektiviserings- og rationaliseringsgevinster rundt omkring i det danske samfund. Sprogteknologi kan effektivisere omgangen med de store mængder information, der hver dag produceres og udveksles eller lagres og hentes i offentlig og privat administration. Med sprogteknologi kan det gøres lettere at fremfinde, behandle og formidle information, som det f.eks. sker ved elektronisk borgerbetjening og elektronisk dokumenthåndtering.

Virksomhederne og den offentlige forvaltning kan effektivisere sine arbejdsprocesser med sprogteknologi, f.eks.:

- De kan få en mere entydig faglig kommunikation ved at udarbejde termbaser, der systematisk beskriver de faglige begrebers indhold, afgrænsning og indbyrdes relationer. Det har f.eks. virksomheden Nordea gjort.
- De kan få en mere effektiv vidensorganisering og videnshåndtering gennem indekseringsprogrammer m.m. til elektronisk dokumenthåndtering. Det har f.eks. Kommunernes Landsforening gjort.
- De kan effektivisere brevskrivning og journalisering ved hjælp af dikteringsværktøj med talegenkendelse trænet til bestemte anvendelsesdomæner. På Vejle Sygehus' røntgenafdeling har dikteringsværktøj f.eks. betydet, at læger selv kan indtale journalerne, hvilket sparer sekretærerne tid og kan fremskynde patientbehandlingen.
- De kan have en automatisk telefonpasningsmaskine, f.eks. som produktet Emma, der kører hos Danmarks Radio.
- De kan gennem talegenkendelse få udfyldt spørgeskemaerne i en telefonbaseret interviewundersøgelse på en enklere og billigere måde end før.
- De kan lette oversættelsesarbejde ved hjælp af en computers råoversættelse. Det gør f.eks. firmaet Lingtech.

1.1.3 Innovation og indtjeningsmuligheder I fremtiden vil mange af de sprogteknologiske delelementer, vi kender eller forestiller os i dag, kunne bygges sammen til nye og mere avancerede systemer. Mange af disse anvendelser udgør innovationsmuligheder for dansk erhvervsliv. Forskning i dansk sprogteknologi kan gøre virksomhederne i stand til at producere dansksprogede løsninger til det danske marked for siden hen at overføre teknologien til produkter for andre sprog.

Sprogteknologi er en forudsætning for et avanceret it-samfund, og om it har regeringen sagt:

Det samfund, der er bedst til at bruge ny teknologi, vil klare sig bedst i den internationale konkurrence. En effektiv udnyttelse af it giver grundlag for større effektivitet, nye produkter og bedre tilbud til borgerne. It vil i de kommende år være en afgørende faktor for vækst i økonomien og dermed grundlaget for velfærden i Danmark. (Den danske regerings regeringsgrundlag fra 26. november 2001).

Der er f.eks. innovationsmuligheder i:

- Maskinoversættelsesprogrammer, der kan oversætte til dansk fra forskellige sprog og hjælpe virksomheder med at oversætte brugerdokumentation mv. Oversættelsesprogrammer vil også gøre det flersprogede internet mere brugbart for alle.
- Intelligente søgeprogrammer til nettet, der kan håndtere homografer, synonymer osv.
- Programmer til automatisk klassifikation og indeksering.
- Programmer til automatisk resumering.
- Interaktive talegrænseflader, f.eks. til navigation på hjemmesider eller til navigation i lydbøger for blinde.
- Skrivestøtteprogrammer, der kombinerer ordprædiktion, stavekontrol, grammatikkontrol, synonymordbog og talesyntese.
- Læsestøtteprogrammer, der kombinerer talesyntese, indbygget skærmlæser (talesyntese for f.eks. navigationsmuligheder på skærmen), markering af ord, der oplæses, og automatisk afkortning af tekster.

1.2 ET SPROGPOLITISK REDSKAB For at få det fulde udbytte af sprogteknologi skal man kunne betjene teknologien via sit modersmål, det sprog, man udtrykker sig bedst på.

Kulturminister Brian Mikkelsen fremfører i sin sprogpolitiske redegørelse, at:

> *Regeringen ser det som et mål at sænke barriererne for udvikling af dansk sprog- og taleteknologi, således at danske brugere og industri kan få lettere adgang til de samme grundlæggende sprog- og taleteknologiske værktøjer, som man har inden for andre sprogområder.* (Mikkelsen, Brian, kulturminister: Sprogpolitisk redegørelse, 18.12.2003).

I Kulturministeriets udspil til en dansk sprogpolitik, Sprog på spil (2003), blev det anbefalet at sætte ind mod det danske sprogs domænetab, dvs. tendensen til, at dansk bliver fortrængt af andre sprog på bestemte domæner. I Danmark oplever vi i øjeblikket, at det i det videnskabelige domæne bliver stadigt mere udbredt at skrive eller undervise på engelsk, og i dele af erhvervslivet bruger en række virksomheder med internationalt fokus et andet sprog end dansk som koncernsprog.

Kulturministeriet anbefalede at sigte mod parallelsproglighed, dvs. at der „bruges dansk, ikke i stedet for, men ved siden af vor tids internationale hjælpesprog". I udspillet foreslås det f.eks. at sætte ind med elektroniske termbanker og en mere systematisk oversættelse af dokumenter, således at man sikrer det danske sprogs termudvikling og muligheden for at formulere sig på sit danske modersmål. På den måde forebygger man f.eks., at sprogbarrierer sænker det faglige niveau på universiteterne eller giver befolkningen en vanskeligere adgang til forskningsresultater m.m. I rapporten *Sprogpolitik på de danske universiteter* (2003) vurderer Rektorkollegiet imidlertid, at det vil være for dyrt at satse på en systematisk oversættelse af forskningsresultater.

Små sprogområder som det danske har særligt stort behov for sprogteknologi til at sikre nationalsproget mod domænetab. Det er nødvendigt, at alle sprogteknologiske hjælpeværktøjer findes på dansk i høj kvalitet, hvis det danske sprog skal bruges og udvikles på alle domæner i fremtiden.

Sprogteknologi kan bruges sprogpolitisk:

- En oversættelsesmaskine, der letter oversættelse mellem dansk og f.eks. engelsk, kan bringe Danmark nærmere parallelsproglighed på udsatte domæner. Råoversættelserne skal ganske vist efterbearbejdes grundigt før offentliggørelse, men det vil trods alt blive mindre uoverkommeligt end nu at sørge for, at tekster, hjemmesider osv. både kommer i en dansk og en engelsk version.
- Tværsproglige søgemaskiner kan sikre, at danske brugere kan bruge deres modersmål, når de søger på nettet, og at udlændinge kan finde danske informationer.

- Oversættelse af danske hjemmesider fra dansk til andre sprog bidrager til, at dansk sprog og tanker udbredes.
- Grammatik- og stiltjekkere kan bruges til udviklingen og realiseringen af virksomheders sprogpolitik. I fremtiden vil endnu mere intelligente programmer kunne indgå i tekstproduktionen og derved støtte tekstproduktion også på dansk.
- Det, at der findes dansk talesyntese, har betydet, at de navigeringsprogrammer, man har i biler, nu taler dansk – i begyndelsen talte de engelsk.

1.3 HUMANISTISK OG TVÆRFAGLIGT FORSKNINGSUDBYTTE Sprogteknologi er et humanistisk forskningsfelt med et stort potentiale for samarbejde mellem forskellige discipliner. Forskning i sprogteknologi kan bidrage til at øge forskningssamarbejdet på tværs af eksisterende faggrænser, eftersom den ikke blot involverer datalogi og sprogvidenskabens forskellige discipliner, men også medievidenskab, tekstvidenskab, informationsvidenskab, ingeniørvidenskab og fysik (akustik).

1.4 SPROGTEKNOLOGI FOR DANSK – EN NØDVENDIGHED Der er behov for forskning i og udvikling af sprogteknologi, der kan resultere i gode sprogteknologiske produkter og tjenester for dansk. Sprogteknologi er et hjælpemiddel i dagligdagen, et middel til effektivisering af arbejdsprocesser, en kilde til innovation og et sprogpolitisk redskab.

At sikre god dansk sprogteknologi nu og i fremtiden kræver både økonomiske ressourcer, en stor forskningsindsats og et veludviklet samarbejde mellem aktørerne. Til gengæld får man foruden god dansk sprogteknologi også et vigtigt forskningsmæssigt udbytte inden for forskellige fag inden for og uden for humaniora.

2. Aktørerne i det sprogteknologiske landskab

Danmark står på spring for at udnytte moderne informations- og kommunikationsteknologi og deltage i udviklingen af den. På det punkt rangerer Danmark som nr. 5 blandt 102 lande, kun overgået af USA, Singapore, Finland og Sverige, fremgår det af rapporten *Global Information Technology Report* 2003-2004.

Paratheden i forhold til sprogteknologi består for det første i et højt dansk forskningsniveau på området. Rapporten *Benchmarking HLT progress in Europe* (2003) konstaterer, at Danmark har en stærk tradition for natursprogsbehandling og en velfunderet taleforskning. For det andet og tredje består paratheden i et generelt højt uddannelsesniveau og en veludviklet digital infrastruktur. Tal fra Danmarks Statistik viser, at 81 % af befolkningen havde adgang til internettet i hjemmet eller på arbejdet i andet halvår af 2003, og 63 % af befolkningen benyttede internettet ugentligt, 41 % dagligt (*Befolkningens brug af internet 2. halvår* 2003, 2003). Behovet for sprogteknologi er med andre ord lige så stort i Danmark som i store sprogområder og bliver ikke mindre af, at dansk tales af færre mennesker.

Imidlertid er det ikke i sig selv nok at have parate brugere og gode forskningsresultater, hvis det skal lykkes at sikre den sprogteknologiske forskning og udvikling i Danmark. De nødvendige økonomiske ressourcer skal også være til stede, og der skal være en effektiv teknologioverførsel til markedet. Hidtil har disse sidste to betingelser ikke været opfyldt.

Problemet skyldes dels, at de danske virksomheder på området har været for få i antal og kun små eller mellemstore i størrelse. Dels tales dansk af så få mennesker, at der ikke har været et økonomisk bære-

dygtigt marked for dansk sprogteknologi. Og endelig har der ikke været nær så store offentlige investeringer i sprogteknologi som f.eks. i England, Tyskland og Finland. Resultatet er, at danske borgere ikke tilbydes sprogteknologi på nær samme niveau som borgerne i førende europæiske lande.

EU og Norden har gennem årene bidraget til den danske forskning i sprogteknologi, men adgangen til EU-midler er blevet vanskeligere, idet EU mener, at det enkelte land har hovedansvaret for sit eget sprog, og støtten har ikke haft et omfang, der sikrede Danmark en position som et af de førende lande, hvad angår sprogteknologi.

2.1 FORSKNINGSAKTIVITET Der er etableret stærke miljøer for sprogteknologisk forskning og udvikling i Danmark, og der er stærke traditioner for samarbejde forskningsmiljøerne imellem.

Den forskningsmæssige aktivitet inden for sprogteknologi kan f.eks. måles i, at der var ca. 60 forskerårsværk i 2002 og produktion af ca. 30 ph.d.er i perioden 1995-2002. I 2002 var der 44 forskerårsværk afsat til faget datalingvistik, og i perioden 1995-2002 blev der produceret 17 ph.d.er i faget. For taleteknologis vedkommende var tallene henholdsvis 15 og 12.

Disse tal stammer bl.a. fra en dataindsamling foretaget i sommeren 2002 i forbindelse med udarbejdelsen af *Oplæg til dansk it-forskningsstrategi* (2002). Alle danske universiteter, forskningsinstitutioner mv. blev spurgt, om de udøvede it-forskning, og hvordan de ville karakterisere den. Især afkrydsningerne i spørgeskemaernes kategorier datalingvistik og taleteknologi tegner et billede af aktivitetsniveauet på det sprogteknologiske område. Men den faktiske aktivitet kan være større i og med, at området har berøringsflader til mange andre fag.

Den danske forskning i sprogteknologi er spredt på en række forskningsinstitutioner, men de fleste af de ansatte er koncentreret nogle få steder. Fire større danske forskningsinstitutioner har sprogteknologisk forskning som et erklæret satsningsområde:

- Aalborg Universitet: Institut for Elektroniske Systemer (Afdeling for Kommunikationsteknologi) samt Institut for Kommunikation.
- Syddansk Universitet: Institut for Fagsprog, Kommunikation og Informationsvidenskab; Laboratoriet for Naturlige Interaktive Systemer; Institut for Sprog og Kommunikation.
- Københavns Universitet: Center for Sprogteknologi; Institut for Almen og Anvendt Sprogvidenskab.
- Handelshøjskolen i København: Institut for Datalingvistik.

2.1.1 Aalborg Universitet

Center for Personkommunikation, nu Center for Teleinfrastruktur

Center for Personkommunikation blev oprindelig etableret for en femårsperiode i 1993, finansieret af Statens Teknisk-Videnskabelige Forskningsråd, siden fulgt op af endnu en femårsperiode. Siden 1. januar 2003 har centret været fuldt integreret i Institut for Elektroniske Systemer, Afdeling for Kommunikationsteknologi, Aalborg Universitet. Centret er finansieret af universitetet og Statens Teknisk-Videnskabelige Forskningsråd i fællesskab. Centrets finansiering fra STVF ophørte med udgangen af 2002, men centrets forskning indgår som væsentligt bidrag i oprettelsen af et nyt center, Center for Teleinfrastruktur, fra januar 2004.

Ved Center for PersonKommunikation, forskningsgruppen Tale- og Multimedie Kommunikation, drives der forskning inden for forskellige områder. Områderne er naturlig tale, sprogbehandling og intelligente multimediesystemer, der er fokuseret på robusthed og brugervenlighed. Forskningen drives i sammenhæng med integration i anvendelser i kontorarbejde, via webbrowsing eller over faste og trådløse netværk. Forskningen dækker genkendelse og forståelse af naturlig tale, tekst-til-tale-syntese, intelligente multimodale systemer samt grundlæggende værktøjer og korpusser med akustiske taledata (et korpus er en afgrænset mængde af talte eller skrevne data, som er indsamlet efter klart definerede kriterier, og som er til-

gængelige i elektronisk form).

I forbindelse med talegenkendelse og forståelse af naturligt sprog arbejdes der i kontekst af interaktive talestyrede dialogsystemer på grundlag af bl.a. parsere (program, der sætter computeren i stand til at analysere og strukturere løbende datainput, f.eks. tekst eller tale). Forskningen i tekst-til-tale er rettet mod nye metoder til etablering af forståelig og naturligt lydende syntetisk dansk tale. Herunder modellering af intonation samt datamining med henblik på udvikling af metodikker til fleksibel dannelse af nye stemmer på grundlag af akustiske korpusser.

Inden for interaktive dialogsystemer er forskningen rettet mod teorier og metoder til brugeranalyser vedrørende anvendelser, som der er adgang til fra desktopcomputere og håndbårne terminaler over såvel det faste som det trådløse netværk.

Gruppens forskning er endvidere rettet mod anvendelser af taleteknologi og natursprogsbehandling for personer med nedsatte funktioner (afasi, ordblindhed, nedsat bevægelighed m.m).

Institut for Kommunikation Ved Institut for Kommunikation forskes der i kommunikation i bred forstand, dvs. kommunikation mellem mennesker, via tekster og i medier. Dette udmøntes i forskning og undervisning inden for en række videnskabelige discipliner: litteraturvidenskab, lingvistik (studier af morfologi, syntaks osv.), psykologi, filosofi/videnskabsteori, humanistisk datalogi, medievidenskab, interpersonel kommunikationsteori, organisationsteori og læringsteori.

Andre forskningsgrupper på Aalborg Universitet Der er en række andre forskningsgrupper, der er relevante i denne sammenhæng, herunder Forskningsgruppen Naturlige og Formelle Sprog, Center for Lingvistik og Center for Digitale, Interaktive Medier.

2.1.2 Syddansk Universitet Ved Syddansk Universitet er den sprogteknologiske forskning primært koncentreret på NIS-laboratoriet (Odense), Institut for Sprog og Kommunikation (Odense) samt Institut for Fagsprog, Kommunikation og Informationsvidenskab (Kolding).

Laboratoriet for Naturlige Interaktive Systemer Laboratoriet for Naturlige Interaktive Systemer satser især på det teknologiske aspekt og er involveret i en række store europæiske initiativer som f.eks. European Network of Excellence in HLT (ELSNET), Natural Interactive Communication for Edutainment (NICE), Natural Interactivity Tools Engineering (NITE), International Special Interest Group on Discourse and Dialogue (SIGdial) og Speechdriven Interfaces for Consumer Devices (SPEECON).

Institut for Sprog og Kommunikation Institut for Sprog og Kommunikations sprogteknologiske satsning er baseret på de tværsproglige VISLprojekter (VISL: Visual Interactive Syntax Learning), der beskæftiger sig med især udvikling af natursprogsparsere, computer aided language learning (CALL), maskinoversættelse og produktion af sprogteknologiske ressourcer (korpus-opmærkning, f.eks. Korpus 90/2000). VISL har bilaterale samarbejder med en række europæiske partnere, men er også involveret i flere nordiske forskningsnetværk (f.eks. vedrørende navnegenkendelse og syntaktiske træbanker).

Institut for Fagsprog, Kommunikation og Informationsvidenskab
Institut for Fagsprog, Kommunikation og Informationsvidenskab er aktiv på sprogteknologiske områder: terminologi og leksikologi (f.eks. DANTERMbank SYD og NorNa-søgemaskinen), maskinoversættelse (f.eks. engelsk-dansk Comprendium), strukturering og design af information i faglige tekster (f.eks. Informationssøgning med XML) og modellering af sproglige aspekter af fagsproglig kommunikation (bl.a. HPSG-grammatik). Instituttet deltager i OntoQuery-projektet (Ontology-based querying).

2.1.3 Københavns Universitet

Center for Sprogteknologi Center for Sprogteknologi blev i 1991 oprettet som en sektorforskningsinstitution under Ministeriet for Videnskab, Teknologi og Udvikling, men er fra 1. januar 2004 blevet fusioneret med Københavns Universitet under det humanistiske fakultet. Centrets økonomi er baseret på en årlig basisbevilling fra ministeriet, på eksterne midler fra forskningsprogrammer og på andre eksterne midler. Centret er et nationalt center for sprogteknologi og har til formål at udføre og fremme strategisk forskning og kommerciel udvikling inden for sprogteknologi og datalingvistik i Danmark. Via internationalt samarbejde skal centret skaffe ny viden til Danmark, nyttiggøre den i en dansk sammenhæng og bidrage til den internationale videnskabelige udvikling på området. Det er en af Center for Sprogteknologis nye organisatoriske opgaver at samle de vigtigste sprogteknologiske miljøer i Danmark i et forskningskonsortium for sprogteknologi.

Ved centret forskes der i datalingvistik, teoretisk lingvistik, leksikografi, terminologi, dansk sprog, en række fremmedsprog, datalogi og ingeniørvidenskab. Centrets forskning falder i to dele, dels ressourcer, dels teknologier (både basisteknologier og integrerede teknologier). Inden for sprogressourceområdet skal fremhæves arbejdet med den store SprogTeknologisk Ordbase (STO). Herudover foregår der forskning i ontologier, leksikalsk semantik, navnegenkendelse mv. Der arbejdes med formel dansk sprogbeskrivelse og dansk datamatisk grammatik i et samarbejde med Handelshøjskolen i København.

Inden for området integrerede teknologier er maskinoversættelse traditionelt det vigtigste område for centret; det er også det område, hvor det er lykkedes bedst at skabe samarbejde med erhvervslivet. Herudover arbejdes der med multimodale systemer, hvor sprogteknologi integreres, med sprogteknologisk beriget informationssøgning og semantisk web, videnshåndtering, kontrolleret sprog samt resumering.

Center for Sprogteknologi har gennem årene haft en række projekter, der sigtede mod dokumentation af sprogteknologien i Danmark og Europa, oplysningskampagner om anvendelse af sprogteknologi mv.

Institut for Almen og Anvendt Sprogvidenskab Institut for Almen og Anvendt Sprogvidenskab er hjemsted for uddannelsen i datalingvistik på Københavns Universitet og har været et arnested for udviklingen af datalingvistik i Danmark. Instituttet deltog i arbejdet med at forbedre og videreudvikle den danske talesyntese i Videnskabsministeriets talesynteseprojekt, især med henblik på prosodi (tryk og rytme i talt sprog).

Andre forskningsgrupper på Københavns Universitet Der foregår relevante aktiviteter på en række af de sproglige institutter, herunder især på Institut for Nordisk Filologi og Romansk Institut.

2.1.4 Handelshøjskolen i København

Institut for datalingvistik Datalingvistik har siden 1997 været satsningsområde ved Handelshøjskolen i København, hvilket har medført øgede interne bevillinger til Institut for Datalingvistik til forskningsaktiviteter i størrelsesordenen 1-1,5 mio. kr. om året. I de senere år er ca. 60 % af instituttets forskning eksternt finansieret. Instituttet varetager undervisning og forskning inden for sproglig it, dvs. de aspekter af informationsteknologien, der involverer sprog. Med sine 150 aktive studerende er instituttet det største uddannelsessted for datalingvister i Danmark. Instituttet har oprettet et grundforskningscenter i 2002 og et i 2003.

Medarbejderne ved Institut for Datalingvistik, Handelshøjskolen i København, forsker i formel beskrivelse af ordforråd og grammatik, i datamatisk terminologi, semantik og vidensmodellering samt på processeringssiden i korpusbaserede maskinlæringsteknikker og statistisk processering. Resultaterne af forskningen afprøves i f.eks. auto-

matisk oversættelse, indholdsbaseret informationssøgning, dialogsystemer og brugergrænseflader, som kan håndtere talt eller skrevet input. Der arbejdes med forskellige typer af almensproglige og fagsproglige korpusser, og instituttet har udviklet værktøjer til søgning, opmærkning og statistisk bearbejdning af skrift- og talesprog. Instituttet samarbejder med tilsvarende miljøer i Danmark om at tilvejebringe de nødvendige ressourcer, herunder en dansk sprogbank, og deltager bl.a. i opbygningen af den store orddatabase STO. Institut for Datalingvistik deltager i et stort antal nationale og internationale forskningsprojekter, både i EU-regi og under de nordiske forskningsråd.

Center for Computational Modelling of Language Grundforskningscentret Center for Computational Modelling of Language har til opgave at forske i nye metoder til analyse og datamatisk håndtering af naturlige sprog, herunder talesprog. Der arbejdes både med regelbaserede og statistiske modeller inden for fonologi (udtalelære), morfologi (læren om ordenes kategorier og bøjningsformer), syntaks og semantik, herunder især maskinindlæring af grammatik og ordforråd på basis af ensprogede og parallelle flersprogede korpusser. Centret har netop opnået bevilling som center under Statens Humanistiske Forskningsråd for 2004-2006 samt en bevilling (sammen med Center for Sprogteknologi) til forskning i maskinoversættelse baseret på parallelle korpusser.

Center for Terminologiske Ontologier Grundforskningscentret Center for Terminologiske Ontologier har til opgave at udvikle teorier og metoder til formel beskrivelse af terminologiske ontologier, specielt polyrelationelle begrebssystemer.

DANTERMcentret DANTERMcentret er en erhvervsdrivende fond oprettet i 1998 af Terminologigruppen, en paraplyorganisation for alle terminologiske aktiviteter i Danmark. Centret blev oprettet med centerkontraktbevillinger fra Statens Humanistiske Forskningsråd og Erhvervsfremme Styrelsen. Handelshøjskolen i København har afsat midler til at sikre en videreførelse af DANTERMcentret efter centerkontraktens ophør. Det forventes dog, at centret fremover i overvejende grad kan finansieres ved indtægtsdækket virksomhed. Centret har adresse på Handelshøjskolen i tilknytning til Institut for Datalingvistik. Instituttets medarbejdere deltager i forskningsprojekterne, som gennemføres i tilknytning til centrets kontraktprojekter, og erfaringer fra DANTERMcentret anvendes i undervisningen i form af cases og baggrundsviden.

DANTERMcentret beskæftiger sig bl.a. med virksomheders termbaser, sprogpolitik og strategi for anvendelse af sprogteknologiske værktøjer i bredere forstand, f.eks. oversættelseshukommelsessystemer og elektroniske eller netbaserede ordbøger.

2.1.5 Andre universiteter og forskningsinstitutioner

Ovenfor er gennemgået de vigtigste koncentrationer af sprogteknologisk forskning, men relevant forskning finder også sted andre steder, bl.a. på Handelshøjskolen i Århus, hvor der på Tysk institut forskes i sprogteknologi med henblik på sprogundervisning, og hvor både Center for Leksikografi og Center for Medicinsk Fagsprog anvender sprogteknologiske metoder.

Uden for universiteterne kan man nævne Det Danske Sprog- og Litteraturselskab, der anvender sprogteknologiske metoder i korpusarbejdet.

2.1.6 Samarbejde mellem fag og mellem forskningsinstitutioner

Sprogteknologi er et humanistisk og tværfagligt forskningsfelt, og det afspejler sig i en stærk dansk tradition for forskningssamarbejde.

Den sprogteknologiske forskning kan bidrage til ny erkendelse om sprog, kognition, videnshåndtering og meget mere. Når man arbejder

med at sætte det danske sprog i system for en computer, får man samtidig en bedre forståelse af sproget, og derfor udgør forskning i sprogteknologi en spændende udfordring til den lingvistiske forskning i fonetik (udtalelære), grammatik, semantik, pragmatik (ordenes/enhedernes betydning i kommunikationssituationen) osv. Korpusarbejdet, der er en del af det empiriske grundlag for forskning i sprogteknologi, leder f.eks. hele tiden til ny viden med hensyn til, hvordan det danske sprog i praksis tales og skrives. Fremtidens metoder og teknikker til korpushåndtering vil med stort udbytte også kunne anvendes til f.eks. historisk sprogforskning og forskning i sprogpsykologi, fremmedsprog og fremmedsprogspædagogik. Ligeledes bidrager arbejdet med at udvikle sprogteknologiens ontologier, termbaser, grammatikker osv. til en voksende viden om dansk begrebsapparat, ordforråd, sprogopbygning m.m.

Samtidig med at sprogteknologi er et humanistisk forskningsfelt, er det i høj grad et emne for forskning på tværs af fag og fakulteter. Forskningen foregår i et meget tæt samarbejde mellem sprogvidenskab, informations- og medievidenskab, datalogi og ingeniørvidenskab. I fremtidens it-samfund er sprogteknologi indlejret i komplicerede it-løsninger, og behovet for tværfaglig kreativitet vil vokse yderligere.

Danske forskeres tradition for at samarbejde med hinanden om forskning i sprogteknologi er en væsentlig årsag til, at den danske forskning står i en så lovende position, som den gør i dag.

Der er mange eksempler på godt dansk forskningssamarbejde mellem institutioner og på tværs af fag, når det gælder sprogteknologi:

- OntoQuery er et samarbejde mellem: Center for Sprogteknologi, Danmarks Tekniske Universitet, Handelshøjskolen i København, Roskilde Universitetscenter og Syddansk Universitet.
- Udforskning af Dansk Ordforråd og Grammatik (UDOG) havde deltagere fra: Center for Sprogteknologi, Odense Universitet, Institut for Erhvervsforskning, Handelshøjskole Syd, Aalborg Universitet og Handelshøjskolen i Århus.
- Dansk Syntetisk Tale (DST), udvikling af den danske talesyntese til ubegrænsede danske tekster, havde tre forskningspartnere: Center for PersonKommunikation, Københavns Universitet og Tele Danmark.
- Den sprogteknologiske ordbog STO er et samarbejdsprojekt mellem: Center for Sprogteknologi, Københavns Universitet, Handelshøjskolen i København og Syddansk Universitet.

2.2 SAMSPIL MED ERHVERVSLIVET Danske forskningsmiljøer og dansk erhvervsliv har samarbejdet om en række sprogteknologiske projekter, og samarbejdet har illustreret, at dette er en god model. Men der er for lidt af denne type samarbejde med det resultat, at teknologioverførslen ikke fungerer godt nok, og forskningsresultaterne har svært ved at nå ud til brugerne som markedsklare produkter og tjenester.

Hertil kommer, at erhvervslivet under ingen omstændigheder kan løfte den store opgave, det er at sikre de nødvendige økonomiske ressourcer til forskning i sprogteknologi for dansk.

Der er 15-20 kommercielle udviklere og leverandører af sprogteknologi i Danmark, bl.a. Ankiro Aps, Max Manus A/S og Mikro Værkstedet A/S. Heraf er nogle få meget store, mens mange er meget små. Derudover er der 10-15 andre leverandører af sprogteknologiske værktøjer og løsninger. Virksomhedernes begrænsede antal og størrelse sætter grænser for deres evne til at finansiere og medfinansiere sprogteknologisk forskning, især når det danske købermarked er lille, og udsigten til økonomisk udbytte er tilsvarende lille. I udlandet er der store firmaer inden for feltet, men der er ikke international efterspørgsel efter at udvikle sprogteknologi for dansk.

Hvis man overlader finansieringen af den sprogteknologiske forskning til markedet, leder det til generel underinvestering. Sikringen af grundforskningens kontinuitet, omfang og kvalitet går tabt, bl.a. fordi grundforskning typisk ikke falder ind under virksomhedernes

mere kortsigtede, kommercielle forskningsinteresse.

Eksempler på, at dansk forskning og dansk erhvervsliv har samarbejdet om forskning i og udvikling af dansk sprogteknologi:

- VID, et forsknings- og udviklingsprojekt, der vedrører bl.a. informationssøgning og dokumentproduktion: Foruden Center for Sprogteknologi omfatter projektet på den ene side virksomhederne Bang & Olufsen A/S, Zacco A/S og Nordea A/S, som i dette projekt udgør teknologiens brugere, og på den anden side Navigo Systems A/S og Ankiro, som er teknologiproducenter. Projektet har modtaget støtte fra Center for IT-forskning.
- PaTrans, et fuldautomatisk maskinoversættelsessystem, der oversætter engelsksprogede patentdokumenter fra engelsk til dansk: Systemet er udviklet for oversættelsesfirmaet Lingtech A/S af Center for Sprogteknologi, og systemet oversætter mere end tre mio. ord årligt.
- Identifikation og anonymisering af navne (IDANNA): Projektet går ud på at forske i navnegenkendelse og udvikle et automatisk værktøj, som via anonymisering gør ellers fortrolige dokumenter tilgængelige for udenforstående. De anonymiserede data bruges til at træne og udvikle et sætningsmønster (mønster for tale på sætningsniveau inden for et bestemt domæne) til talegenkendelse i dikteringsværktøj for advokater. Projektet er et samarbejde mellem Center for Sprogteknologi og Max Manus A/S. Philips indgår i projektet ved at være det firma, som varetager selve udviklingen af dikteringsværktøjet. Projektet finansieres med en bevilling fra Statens Humanistiske Forskningsråd og af de to samarbejdspartneres medfinansiering.
- It-baseret stave- og skrivehjælp, som skal resultere i en brugervenlig stave- og grammatikkontrol, udviklet af Dansk Videnscenter for Ordblindhed, Mikro Værkstedet A/S og Grammarsoft i fællesskab. Udviklingsprojektet har fået bevilget den fornødne økonomiske støtte fra bl.a. IT- og Telestyrelsen, Egmont-Fonden og Tips/lottobevillingen.
- Hjælpemiddel til læseindlæring for ordblinde og personer med læsevanskeligheder (Computerlæs), et forsknings-/forstudieprojekt til opbygning og testning af en prototype og metoder til brug for ordblindes selvindlæring af konstaterede læsevanskeligheder: I projektet deltog Center for PersonKommunikation, Københavns Universitet, Instrulog, ScanDis samt Mikro Værkstedet A/S, og projektet blev støttet af Forsknings- og Udviklingscentret for Hjælpemidler og Rehabilitering.
- Nummer-til-Navn, telefontjeneste, der i dag kan benyttes via nummer 1811: Et samarbejdsprojekt mellem Center for PersonKommunikation og Tele Danmark, som finansierede projektet sammen med Aalborg Universitet.

2.2.1 Virkemidler

Forskningssamarbejde med erhvervslivet kan antage mange former lige fra præcise kontrakter, hvor det ønskede resultat er forholdsvis klart formuleret, over mere løst samarbejde til egentlige sponsorater, hvor virksomhedens udbytte snarere er kontakter eller forbedret image. Samarbejdet kan f.eks. have form som ved Aalborg Universitets samfinansieringsmodel, ved centerkontrakter eller ved støtte gennem Center for IT-forskning.

Aalborg Universitets samfinansieringsmodel Ved Aalborg Universitet har det Teknisk-Naturvidenskabelige Fakultet gennem en længere årrække medfinansieret eksterne projekter, hvis indsats ligger inden for fakultetets forskningsmæssige indsatsområder, og som foregår i samarbejde med erhvervslivspartnere. Ordningen kræver, at fakultetet – inden projektet er endeligt aftalt med den eksterne (private) partner – er med i projektets budgetlægning.

For hvert forskningsårsværk medfinansierer fakultetet den halve indsats, mens den eksterne partner finansierer den anden halvdel. Til

gengæld skal den forskningsansatte formidle sine forskningsresultater ved at deltage i undervisning i form af projektvejledning, ph.d.-vejledning eller kursusafvikling.

Ved siden af denne ordning har fakultetet en ordning, der vedrører indkøb af udstyr til forskningsarbejdet. Også her gælder det, at fakultetet betaler halvdelen af omkostningerne, mens den eksterne partner betaler den anden halvdel.

Endelig har fakultetet en ordning, der på bestemte betingelser giver alle fakultetets ansatte VIP'er dækning af udgifter til forskningsformidlingsrejser.

Centerkontrakter En centerkontrakt var et trekantssamarbejde mellem f.eks. to forskningsinstitutioner, tre-fire virksomheder og et Godkendt Teknologisk Serviceinstitut (GTS-institut). Virksomhederne medfinansierede 50 % af aktiviteterne ved eget arbejde, og de modtog ikke direkte tilskud i ordningen. Det teknologiske serviceinstitut finansierede 25 % af sin del af aktiviteterne og modtog bevillinger fra Innovationssøjlen under Ministeriet for Videnskab, Teknologi og Udvikling (tidligere via det daværende Erhvervsfremme Styrelsen).

Centerkontrakten skulle udspringe af et konkret innovationsbehov i virksomhederne og opbygge kommercielt orienteret knowhow i den deltagende teknologiske service. Et resultat af centerkontrakten kunne være, at der udvikles og markedsføres helt nye produkter samtidig med, at der bliver publiceret en stribe forskningsartikler. DANTERM centret ved Handelshøjskolen i København er et eksempel på en sprogteknologisk institution, der har indgået en centerkontrakt.

Betingelserne for centerkontrakter har i øvrigt meget til fælles med betingelserne for Center for IT-forskning og Statens Humanistiske Forskningsråds pilotprojekter, f.eks. IDANNA. Men Center for IT-forskning krævede ikke GTS-deltagelse, hvilket har været en fordel i visse tilfælde.

I dag er centerkontrakterne afløst af innovationskonsortier under Ministeriet for Videnskab, Teknologi og Udvikling.

2.2.2 Barrierer for teknologioverførsel

Dansk erhvervsliv har ganske vist generelt et forholdsvis stort innovationspotentiale og en høj grad af konkurrencedygtighed, men engagementet i sprogteknologi er relativt begrænset, og der savnes brede kanaler for teknologioverførsel, vurderer rapporten *Benchmarking HLT progress in Europe* (2003).

I en handlingsplan fremfører den danske regering ligeledes, at samspillet mellem forskning og erhverv er for dårligt *(Nye veje mellem forskning og erhverv – fra tanke til faktura* (2003)). Samtidig konstaterer man, at mange af især de små eller mellemstore virksomheder, der ikke har erfaring med samspil med vidensinstitutioner, selv vurderer, at de ville stå stærkere i den fremtidige konkurrence med et samspil. I *Oplæg til dansk it-forskningsstrategi* (2003) argumenteres der for, at et tættere samarbejde mellem universiteterne og erhvervslivet vil accelerere innovation inden for it.

Både erhvervsliv og forskning kan have gavn af et samarbejde. Danske virksomheder har en objektiv interesse i indholdet af den danske sprogteknologiske forskning, fordi sprogteknologi foruden effektiviseringsgevinster kan betyde nye indtjenings- og ekspansionsmuligheder for virksomhederne. Forskerne får feedback på deres forskningsresultater og inspireres i mødet med den empiriske verden.

Der kan være mange grunde til, at det hidtil har været svært for forskere at samarbejde med erhvervslivet og omvendt. De to miljøer har forskellige vilkår, interesser og kulturer, som det f.eks. ridses op i *Oplæg til dansk it-forskningsstrategi* (2003):

- Universiteterne arbejder med lange tidshorisonter, hvor virksomhederne arbejder med korte.
- Universiteterne skal publicere originale resultater af international standard, mens virksomhederne skal tjene penge.
- Universiteternes indfaldsvinkel er teoretisk, mens virksomhedernes er praktisk.
- De områder, der forskes i, kan være nogle andre end dem, der produktudvikles i.

- Det åbne universitetsmiljø harmonerer dårligt med virksomhedernes interesse i at holde kortene tæt til kroppen, f.eks. vil forskere måske gerne publicere tidligere i processen, end virksomhederne er interesserede i.

Regeringen peger i sin handlingsplan *Nye veje mellem forskning og erhverv - fra tanke til faktura* (2003) på en række muligheder for at forbedre samspillet mellem dansk forskning og erhvervsliv:

- Forslag om at videreføre og udvide den forsøgsordning, der i øjeblikket giver virksomheder mulighed for et skattefradrag på 150 % af udgifter til forskning. Skattefradraget skal gøre det nemmere og mere attraktivt for virksomheder at indlede samarbejde med vidensinstitutionerne.
- Forslag om, at universiteterne indfører mere fleksible og attraktive orlovsmuligheder for forskere, der ansættes midlertidigt i private virksomheder i forbindelse med projekter. I øjeblikket kan forskere ikke altid frikøbes i tilstrækkeligt omfang pga. deres undervisningsforpligtelser.
- En målrettet markedsføringskampagne, der skal udbrede kendskabet til erhvervsforsker-initiativet.
- Forslag om, at institutioner går sammen om fælles selskaber til at arbejde med teknologioverførsel. Med en omlægning af det statslige bevillingssystem vil regeringen gøre det muligt at medfinansiere samarbejdsaktiviteter mellem vidensinstitutioner og virksomheder om etableringen af fælles vidensnetværk, der kan fremme varige samarbejdsrelationer mellem de forskellige aktører.

2.2.3 Fokus på samspillet forskning-erhverv

Der er etableret gode kontakter mellem dansk forskning og erhverv på det sprogteknologiske område, og der er gennemført vellykkede samarbejdsprojekter. Men generelt foregår teknologioverførslen ikke effektivt nok. Desuden er dansk erhvervslivs potentiale for investeringer langtfra en tilstrækkelig garanti for, at der sker forskning og udvikling inden for dansk sprogteknologi i nødvendigt omfang.

Samarbejde mellem forskning og erhverv og mellem offentlige og private investeringer kan ellers give store resultater. Det er det tyske Verbmobilprojekt et godt eksempel på. Det er et sprogteknologisk oversættelsesprojekt til oversættelse mellem tysk, japansk og engelsk. Det offentlige skød 115,1 mio. DM i projektet, mens erhvervslivet bidrog med 51,4 mio. DM. Projektet var udbytterigt for både de deltagende forskere og virksomheder, og det resulterede bl.a. i en vellykket teknologioverførsel og en række innovative produktløsninger: Projektet gav således anledning til 11 patenter, 20 nye produkter og 8 nye virksomheder. Projektet løb fra 1993 til 2000 og er en hovedårsag til, at Tyskland i dag er førende i Europa på det sprogteknologiske felt. Dansk erhvervsliv har ikke tysk erhvervslivs tyngde, men man kan lære af eksemplet, at store offentlige investeringer i sprogteknologi er givet godt ud, især når det sker i et samspil med industrien.

2.3 OFFENTLIGE INITIATIVER Der har været en række offentlige initiativer på området sprogteknologi, og der er adskillige eksempler på tværrådsligt samarbejde om sprogteknologi.

2.3.1 Statens Humanistiske Forskningsråd Statens Humanistiske Forskningsråds strategiplan for 1998-2002 satte fokus på temaet Sprog og Kunst, herunder „sprog og erkendelse". Herigennem blev bl.a. givet støtte til projektet Semantisk udbygning af en Constraint Grammarformalisme for dansk, engelsk og portugisisk. I projektet arbejdede man bl.a. med semantisk inspirerede syntaksforbedringer i grammatikkerne for de nævnte sprog. Projektet forløb fra 1. januar 1999 til 31.december 2001.

Tilbage i 1991 lancerede Statens Humanistiske Forskningsråd initiativet Eksperimentel Sprogvidenskab, og inden for dette indsatsom-

råde har forskningsrådet bl.a. finansieret projektet UDOG. Formålet med projektet var dels at udforske nogle af det danske sprogs basale elementer og deres samspil, dels at udvikle og sammenligne forskellige metoder til ordbeskrivelse på basis af korpusser og korpusværktøjer. Projektet løb over fem år fra 1993.

SHF har i 2003 afsat 4,5 mio. kr. til oprettelse af et SHF-center for 5 år til udvikling af computermodeller for sprog.

2.3.2 Statens Teknisk-Videnskabelige Forskningsråd I sin strategiplan for 1998-2002 prioriterede Statens Teknisk-Videnskabelige Forskningsråd bl.a. området Informatik og elektronik, herunder taleteknologi med vægt på teknikker, der kan bringe mængden af fejlgenkendte ord ned, sproglige ressourcer til træning af systemer samt tværsproglig viden om fonetik og lingvistik.

2.3.3 Tværrådsligt samarbejde om strategiske forskningsprogrammer Programkomiteen for it-forskning, der er nedsat af Forskningsforum, har udmøntet 2003-midlerne, heriblandt 40 mio. kr. afsat på finanslov 2003 til forskningsinitiativer inden for it-forskning, Netværkssamfundet, it-læring og pædagogik og integreret produktion.

Det forrige it-forskningsprogram har bl.a. ydet støtte til det tværfaglige projekt OntoQuery, der har et væsentligt sprogteknologisk aspekt og bl.a. har resulteret i en dansk prototype på et ontologibaseret informationssøgningssystem. Projektet blev påbegyndt i august 1999 og slutter i august 2004.

I strategiplanen for tværvidenskabelig forskning 1998-2002 blev programmet Tværrådsligt Informatik Program (TIP) introduceret. Et af programmets prioritetsområder var Kommunikation og medier, og herunder var Sprog og tale et fokusområde, der angik teknologisk og sproglig forskning i forståelsen af sproglige strukturer som basis for konstruktion af informationsteknologi.

Bevillingssystemet Center for IT-forskning blev oprettet i 1996 og lukket med udgangen af 2002. Bevillingssystemets formål var at styrke dansk it-forskning i samarbejde med dansk erhvervsliv.

Forskningsprogrammet Program for Informatik blev introduceret i et supplement til forskningsrådenes strategiplaner for 1993-97 og blev finansieret af tre råd i fællesskab: Statens Teknisk-Videnskabelige Forskningsråd, Statens Naturvidenskabelige Forskningsråd og Statens Jordbrugs- og Veterinærvidenskabelige Forskningsråd. Programmet var en samlet investering på ca. 65 mio. kr. årligt, og det ydede bl.a. støtte til projektet Spoken Language Dialogue Systems. Dette projekts formål var at udvikle prototyper af talestyrede menneske-maskinedialogsystemer, dvs. systemer, der inden for et begrænset emne er i stand til at føre en talt dialog med en person. Projektet var planlagt at være fireårigt med start i 1991.

2.3.4 Andre initiativer Handlingsplanen *Handicap ingen hindring* fra januar 2003 er et led i regeringens it- og telepolitiske redegørelse *IT for alle – Danmarks fremtid* fra 2002. Handlingsplanens investeringspulje på fem mio. kr. har sikret støtte til fire projekter, der skal medvirke til at gøre it mere tilgængelig for bevægelseshæmmede, blinde og svagsynede samt ordblinde og andre, som har det svært med korrekt stavning.

Den danske regering bevilgede i 1998 midler til talesyntese gennem forsknings- og udviklingsprojektet DST. Systemet er i stand til at interface til Microsoft SAPI, og ved hjælp af programmellets udviklingsplatform er virksomheder i stand til at indbygge det til diverse anvendelser. Projektkontrakten havde Speech-Ware, et ApS ejet af TDC og Nordjyllands Videnpark. Ud over de midler, der blev tilført projektet fra Forskningsministeriet, havde projektet delvis egenfinansiering. Projektet løb fra 1999 til 2001.

Forskningsrådene har administreret 47 mio. kr., som den danske regering afsatte i perioden 1997-2000 til forskning i multimedia.

Under prioritetsområdet Teknologier og værktøjer var et af emnerne flersprogethed og brugen af natursprog i it-systemer, og Staging-projektet var finansieret herunder. Projektet gik ud på at gøre computerbrugere i stand til ved hjælp af tale, håndbevægelser m.m. at føre en sammenhængende dialog med autonome agenter i en virtuel bondegård.

2.3.5 Hidtil ingen stor strategisk satsning Der har været en række offentligt støttede aktiviteter, projekter og initiativer på forskningsfeltet sprogteknologi, og det har givet forskningen et solidt udgangspunkt. Men man må konstatere, at der hidtil ikke har været en større, strategisk, offentlig økonomisk satsning på dansk sprogteknologi.

Udgifterne ved at udvikle sprogteknologi for et sprog er ikke proportionale med, hvor mange der taler sproget. Der skal investeres samme summer for at udvikle sprogteknologi for et lille sprogområde som dansk, som hvis det gjaldt f.eks. engelsk eller tysk. Derfor giver det mere mening at sammenligne forskellige landes udgifter i absolutte tal end pr. indbygger. I udlandet har man haft varierende fokus på forskning i sprogteknologi.

Tyskland er det land, der opnår den højeste placering i sammenligningen af de europæiske landes sprogteknologiske niveau i rapporten *Benchmarking HLT progress in Europe* (2003). Den høje placering afspejler Tysklands stærke tradition for at understøtte forskningen og finde anvendelse for den: koordinere offentlige og private investeringer, fremme stærke forskningsmiljøer og bidrage til internationalt samarbejde.

Holland har en stærk tradition for at yde støtte til sprogteknologi, og et godt eksempel på det er det hollandske initiativ Basic Language Resources Kit (BLARK), (*A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch*, (2002)). BLARK definerer et sæt basale sprogteknologiske ressourcer for nederlandsk, som er nødvendige for, at man kan udvikle sprogteknologiske produkter, og man er

nået langt i de praktiske overvejelser vedrørende prioritering, tilgængeliggørelse, omkostninger osv. Fra januar 2004 har et sprogteknologisk konsortium for forskellige institutter været ansvarlig for administrationen, vedligeholdelsen og distributionen af sprogressourcer og forskellige hollandsk-flamske sprogteknologiske projekter, som er tilvejebragt med offentlig støtte.

Storbritannien har været et af de førende europæiske lande med hensyn til sprogteknologi de sidste tyve år og har i dag en førende position. Siden det fireårige SALT-program, der begyndte i 1990'erne, har der ganske vist ikke været et egentligt, større støtteprogram for området sprogteknologi, men Storbritannien drager fordel af at være et stort sprogområde, og en stor del af den fortsatte aktivitet omkring engelsk sprogteknologi kommer fra tilsvarende aktiviteter i USA og andre lande.

Finland har ydet markant statslig støtte til landets sprogteknologiske forskning. Den teknologiske udviklingscentral Tekes (forskningsrådet) yder i stor stil økonomisk støtte og ekspertbistand til virksomheders og universiteters forsknings- og udviklingsprojekter. Et sprogteknologisk program til en anslået pris af 80 mio. finmark har indtil videre resulteret i fem sprogteknologiske projekter, og i forlængelse heraf er de finske myndigheder ved at finansiere et nationalt universitetsnetværk, der skal styrke forskningen i sprogteknologi.

I Norge har man med programmet Kunnskapsutvikling for norsk språkteknologi (KUNSTI) sat sig for at styrke den norske grundforskning i sprogteknologi, ikke mindst gennem kompetenceopbygning, dvs. oprustning af de norske sprogteknologiske forskningsmiljøer. Programmet løber over seks år fra 2001 til 2006 og har for tiden en samlet bevilling på 75 mio. norske kroner (*KUNSTI – kunnskapsutvikling for norsk språkteknologi. Programplan*, 2001). Herudover har man udarbejdet en organisationsmodel, driftsmodel og finansieringsplan for en stort set offentligt finansieret norsk sprogbank (*Samling og tilgjengeliggjering av norske språkressursar*, 2002).

Det svenske erhvervs- og teknikudviklingsråd og det humanistisk-

samfundsvidenskabelige forskningsråd havde et fælles sprogteknologisk program 1990-1996 og et nyt 1997-1999. Med hjælp fra førstnævnte har man bl.a. opbygget et Svensk OrdNet og arbejdet med tekstanalyse for informationssøgning, korpusbaseret leksikon og analyseværktøj.

På grund af sprogteknologiens tværvidenskabelige og tværfakultære natur vil det også i fremtiden være oplagt for Statens Humanistiske Forskningsråd og Statens Teknisk-Videnskabelige Forskningsråd at samfinansiere forskning i sprogteknologi.

Et udvalg under Ministeriet for Videnskab, Teknologi og Udvikling har afgrænset, evalueret og anbefalet sprog- og taleteknologi som et af otte fremtidige fokusområder i en kommende dansk it-plan (*Oplæg til dansk it-forskningsstrategi* (2003)). I en pressemeddelelse fra 16.01. 2004 oplyser ministeriet, at regeringen over de næste otte år vil placere 16 milliarder kr. i en ny fremtidsfond, der skal styrke de bedste danske vidensmiljøer inden for højteknologiske forskningsfelter som bio-, nano-, informations- og kommunikationsteknologi (*Ny milliardfond skal styrke dansk højteknologi*, 2004). Hvis dette bliver en realitet, skulle der være gode muligheder for at foretage en stor strategisk satsning på sprogteknologi i Danmark.

2.4 UDBYTTE AF INTERNATIONALT SAMARBEJDE Hvis den danske forskning i sprogteknologi tilføres flere økonomiske ressourcer, vil danske forskere kunne deltage mere i det internationale arbejde på området, end der er mulighed for i dag.

Danmark er et lille sprogområde, og derfor har danske forskere en særlig interesse i at samarbejde med forskere i udlandet, vurderer rapporten *Benchmarking HLT progress in Europe* (2003). Den danske befolkning har generelt gode fremmedsprogskundskaber, og i forskningsmiljøerne er der god basis for at dyrke tværsproglig forskning vedrørende sprogteknologi, et eksempel på dét er forskningen i dansk-portugisisk maskinoversættelse ved Syddansk Universitet.

Forskernes tværsproglige potentiale kan skabe en platform, hvorfra man på den ene side drager nytte af udenlandske erfaringer og på den anden side selv bidrager til verdensforskningen.

Danske forskere kan bidrage med den særlige danske forskningstradition, der består i en udpræget humanistisk tilgangsvinkel til sprogteknologi. Man fokuserer meget på forskningen i sprogteknologis sproglige aspekt, mens man i f.eks. Portugal, Brasilien og USA har en mere datalogisk tilgang. Den danske forskning i sprogteknologi udmærker sig ved at have et bredt lingvistisk udgangspunkt og interesserer sig for udforskning af forskellige paradigmer – hermed adskiller danske forskere sig fra forskere med en mere snævert formel grammatisk indfaldsvinkel til sprogteknologi.

EU's EUROTRA-projekt inden for maskinoversættelse gav for det første danske forskere en eminent indføring i international forskning i datalingvistik og natursprogsbehandling med særligt henblik på maskinoversættelse. Danske forskere samarbejdede med de førende europæiske eksperter på området, og projektet har givet varige samarbejdsrelationer med flere europæiske miljøer. For det andet var danske forskere i et samarbejde med et privat firma (Lingtech, aftager) i stand til at udvikle et system, PaTrans, til oversættelse af en bestemt teksttype fra engelsk til dansk.

2.4.1 Europæisk sprogteknologisk agentur Det kan være fordelagtigt for den danske forskning på feltet at arbejde for, at et eventuelt kommende europæisk sprogteknologisk agentur trækkes til Danmark. Et sådant agentur anbefales i rapporten *Benchmarking HLT progress in Europe* (2003). Agenturet skal bl.a. forvalte infrastrukturfonde til gavn for især små sprogsamfund, der har svært ved at finansiere udvikling af sprogressourcer og basisteknologier. I forvejen har Danmark gode argumenter for en dansk placering. Danmark er et lille sprogsamfund med stort multilingvalt potentiale, og danske forskere behersker nyttige metoder til at få meget ud af de ressourcer, der er til rådighed –

disse metoder kan andre lande være interesseret i.

2.4.2 EU-midler Hidtil har den offentlige danske støtte til forskning i sprogteknologi været suppleret af midler fra EU. Sprogteknologi har været med som selvstændigt område i både 2., 3., 4. og 5. rammeprogram. I EU's 6. rammeprogram er størstedelen af satsningsområdet imidlertid blevet indlejret i andre områder, og hvis man vil have penge til forskningen gennem EU i dag, må man som ansøger selv påvise, hvor sprogteknologi eventuelt passer ind i tværfaglige initiativer.

EU-programmernes ændrede satsning er et udtryk for, at behovet for tværfaglig forskning i sprogteknologi er vokset. Samtidig er det imidlertid også et udtryk for, at Kommissionen mener, at faget sprogteknologi nu er modent, og at der derfor ikke længere er behov for investeringer på samme niveau, og ikke behov for investeringer i grundforskning. De enkelte lande må tage ansvaret for deres eget sprog (subsidiaritetsprincippet). Dette argument gælder kun for de store sprog, men det gennemføres altså for alle.

Behovet for tværsproglighed er vokset, og behovet for dygtige danske forskere vil vokse i og med, at dansk forskning er nødt til at vise, at den har noget at bidrage med, hvis udlandet skal gøres interesseret i at beskæftige sig med dansk. EU's ændrede prioritering af støtte til forskning i sprogteknologi betyder, at vi kan forvente, at Danmark i fremtiden selv må finansiere en større del af sin sprogteknologiske grundforskning.

Gennem de sidste 20 år har EU bidraget til mange forskellige sprogteknologiske projekter, som danske forskere har deltaget i, f.eks.:

- Inden for maskinoversættelse har det mest fremtrædende EU-projekt været EUROTRA (1984-92), et maskinoversættelsesprojekt igangsat af EF-kommissionen. EUROTRA kom senere til at udgøre en del af grundlaget for det danske system PaTrans. Andre EU-projekter har fokuseret på udvikling af værktøjer, der kombinerer oversættelse med bl.a. rapportgenerering og resumering: Communicating Through the Language Barrier (LINGUANET) (1995-98) og Language Technologies for Police and Emergency Services (SENSUS) (1998-2000).
- Inden for talegenkendelse og interaktion har der f.eks. været EU-støtte til projektet ONOMASTICA (1993-95), en indsamling af tekstuelle data og hertil hørende standardtransskription for 11 europæiske sprog (transskribere: skrive talt sprog i fonetisk alfabet). Her blev der bl.a. trænet et selvlærende transskriptionssystem. To andre omfattende EU-projekter var talegenkendelsesprojekterne Integration and Design of Speech Understanding Interfaces (SUNSTAR) (1989-93) og Real World Applications of Robust Dialogues (REWARD) (1993-97).
- Arbejdet med sprogteknologiske ordbøger har EU støttet gennem projektet PAROLE (1996-98) og SIMPLE (1998-2000). Projekterne dannede basis for udviklingen af den store danske sprogteknologiske ordbog STO.
- Opbygningen af europæiske talekorpusser er bl.a. blevet støttet gennem EU-projekterne Speech Databases for Creation of Voice Driven Teleservices (SPEECHDAT) (1996-99) og Speech Databases for Voice Driven Teleservices and Control in Automotive Environments (SPEECHDATCAR) (1998-2000).
- EU har bidraget til korpusbaserede grammatikker gennem støtte til projektet Large-Scale Grammars for EU-Languages (LS-GRAM) (1994-96). Det ligeledes EU-støttede Linguistic Specifications for Danish (LINDA) (1994-1995) havde til formål at bidrage til en formel beskrivelse af det danske sprog.
- EU's bidrag til fælleseuropæiske retningslinjer og standarder for sprogressourcer er f.eks. kommet til udtryk gennem projekterne Expert Advisory Group on Language Engineering Standards (EAGLES) (1997-1999) og International Standards for Language Engineering (ISLE-HLT) (2000-2003).
- For at udbrede kendskabet til og fremme anvendelsen af sprogtek-

nologi i Europa igangsatte EU i 1996 Human Language Technologies Opportunity Promotion in Europe (EUROMAP Language Technologies) (1996-2003). Projektet mundede bl.a. ud i rapporten *Benchmarking HLT progress in Europe*, 2003.

2.4.3 Nordisk samarbejde Nordisk sprogteknologisk forskningsprogram 2000-2004 er et forskningsprogram, som Nordisk Ministerråd har igangsat. En programkomité udpeget af Nordisk Ministerråd har det overordnede ansvar for programmet, som administreres af Nordisk Forskeruddannelsesakademi (NorFA). Programmet har tre prioritetsområder: Computerstøttet indlæring af nordiske sprog, Tværsproglig informationsbehandling på de nordiske sprog samt Interaktion mellem menneske og computer ved hjælp af naturligt sprog.

Programmet støtter bl.a. dokumentationscentre for sprogteknologiske forskningsresultater i de nordiske lande samt et formaliseret samarbejde mellem centrene i form af Nordisk netværk for dokumentationscentre for sprogteknologiske forskningsresultater (NorDokNet).

NorFA har desuden arrangeret lingvistiske sommerskoler og seminarer for forskere på det sprogteknologiske område.

Danske forskere er aktive i nordiske projekter, og det nordiske netværk og samarbejde er en ressource, som danske forskere kan trække på, men Norden er ikke en kilde til økonomiske midler i stort omfang. Danmark kan ikke gøre sig håb om at finansiere en væsentlig del af sin forskning, sine sprogressourcer og sine basisteknologier ad denne vej, da midlerne er begrænsede, og da de især anvendes til støtte af netværkssamarbejde.

2.5 BEHOV FOR OFFENTLIG STRATEGISK SATSNING Med en stor strategisk satsning på sprogteknologi netop nu ville man få meget for pengene: En stor del af det grundlæggende forskningsarbejde er allerede gjort takket være mange års initiativer og projekter og de stærke danske forskningsmiljøer, som er gode til at samarbejde og få meget ud af de midler, der er til rådighed. En satsning er særligt velanbragt nu, hvis man skal have størst mulig gavn af det oparbejdede forskningspotentiale og gøre Danmark til et af de førende lande på området.

Danmark befinder sig på EU-gennemsnittet, hvad angår forskning, udvikling og teknologioverførsel for førstegenerationssprogteknologi, men Danmark har behov for, at der er adgang til mere avancerede sprogteknologiske løsninger i fremtiden. Dette kræver, at investeringerne i forskningen øges. Disse investeringer kan man ikke overlade til det private erhvervsliv, for det danske marked er for lille til, at virksomhederne kan forvente et stort afkast, og de danske virksomheder er for små til at kunne investere rigtig stort. Samtidig er mulighederne for at få del i midler fra EU's rammeprogrammer indskrænket. Derfor må Danmark gennem offentlige investeringer sikre den forskning, der skal til.

Herudover skal teknologioverførslen til markedet forbedres. Det kræver brede samarbejdskanaler mellem forskning og erhvervsliv at omsætte forskningsresultater til konkrete sprogteknologiske produkter og tjenester. På grund af miljøernes og virksomhedernes begrænsede størrelse er der i Danmark først og fremmest brug for samarbejde, ikke konkurrence. Ligesom forskningsmiljøerne kan opnå større resultater gennem samarbejde, kan danske virksomheder have gavn af at slå sig sammen om sprogteknologiske projekter.

En stor strategisk satsning skal både styrke forskningen og lette teknologioverførslen for sprogteknologi. Desuden kan satsningen medføre, at der kan udbydes nye tværfaglige uddannelser på de danske universiteter.

3. Sprogressourcer, basisteknologi, integreret teknologi

En række grundlæggende sprogressourcer og teknologier danner forudsætningerne for, at man kan udvikle god dansk sprogteknologi til danske brugere. Produkter og tjenester på markedet er enkle eller mere integrerede teknologier udviklet af basisteknologier. Det samme gælder for forskningen: Avanceret forskning i sprogteknologi kræver næsten i alle tilfælde adgang til basisressourcer og basisteknologier.

De bedste betingelser for sprogteknologisk innovation får man ved at sikre, at basisteknologierne er tilgængelige for alle offentlige organisationer og private virksomheder, der ønsker at udvikle specifikke sprogteknologiske hjælpemidler. I lande med hovedsprog – USA, England, Tyskland og Frankrig – har man tilstrækkeligt udviklet basisteknologi til, at man er kommet godt på vej med hensyn til at etablere en sprogteknologisk industri med et bredt udvalg af sprogteknologiske moduler, som bliver integreret i produkter, der kommunikerer intelligent med mennesker. Der er, som det vil fremgå, skabt et godt fundament af danske sprogressourcer og basisteknologier, men det er ikke godt nok til at sikre, at der udvikles integreret sprogteknologi for dansk i fremtiden.

Inden for både taleteknologi og natursprogsbehandling kan man sondre mellem sproglige dataressourcer, teknologiske ressourcer, de basale teknologier og de integrerede teknologier:

	Taleteknologi	Natursprogsbehandling
Dataressourcer	Talekorpusser	Tekstkorpusser
Teknologiske ressourcer	Fon-, difon-, trifondatabaser	Orddatabaser, grammatikker, tesaurusser, ontologier
Basale teknologier	Konvertering mellem lyd og bogstav og omvendt	Parsere på grundlag af grammatikker og ordtransfersystemer
Integrerede teknologier	Oplæsningsmaskine Tekstningsmaskine	Stave- og grammatiktjek, søgemaskiner, tekstredigering Tekstresumeringsmaskiner
	Oversættelsesmaskiner	

3.1 DATARESSOURCER Til ressourcerne regnes dataressourcerne, dvs. talekorpusser og tekstkorpusser, som igen kan være opmærkede eller uopmærkede, og de teknologiske ressourcer, dvs. databaser, der er udtrukket af korpusserne. Korpusser er samlinger i elektronisk form af en afgrænset mængde af talte eller skrevne sproglige data (f.eks. dialoger eller tekster), som er indsamlet efter klart definerede kriterier, beskrevet efter en given elektronisk standardform.

Tekst- og talekorpusser danner grundlaget for al sprogteknologisk forskning i dag. For at kunne udvikle og tilpasse et sprogteknologisk produkt til en bestemt opgave, f.eks. et maskinoversættelsessystem eller en talegrænseflade, har man brug for eksempler på den sprogbrug, som systemet skal kunne håndtere. Disse eksempler har man i stort tal i et korpus. Man skelner inden for samlingerne af tekster mellem uopmærkede og opmærkede korpusser og mellem monolingvale (etsprogede), multilingvale, multimodale og multimediale (som inddrager flere medier) korpusser. Korpusser optages efter principper, der sikrer, at de faktorer, der har betydning for den forekommende variation, også er repræsenteret i korpusset. Det gælder f.eks. sprogbrugernes fordeling på alder, køn, dialekter og udtale. Ved talekorpusser er der brug for en særdeles god lydkvalitet.

Ligesom det er vanskeligt at give klare mål for korpussets omfang, er det vanskeligt at give klare kriterier for, hvilke tekster der bør indgå i et korpus, for dette afhænger af, hvilken type undersøgelse man ønsker at foretage. Et korpus, som skal bruges til at afdække forekomsten af anglicismer i danske aviser, vil nødvendigvis være anderledes sammensat end et korpus, som skal danne udgangspunkt for undersøgelsen af den juridiske diskurs i danske domme. Og et korpus, der skal danne udgangspunkt for ordbogsproduktion, skal nødvendigvis have en helt anden sammensætning end et korpus, der skal bruges til at udvikle en talegrænseflade til en mobiltelefon eller et dikteringsprogram.

Anvendeligheden af resultatet af en korpusundersøgelse står og falder derfor med, om det er muligt at vurdere, hvilke tekster der er indgået i undersøgelsen, og efter hvilke kriterier oplysningerne er trukket ud. Dette kan kun lade sig gøre, hvis man forsyner den rå tekst med yderligere information, den såkaldte opmærkning. Der findes to typer af opmærkning: ekstralingvistisk opmærkning og lingvistisk opmærkning

Den ekstralingvistiske opmærkning i form af en såkaldt „header" eller et teksthoved omfatter information, der knytter sig til hele teksten og beskriver omfang (antal ord), oprindelse (forfatter, årstal, forlag osv.) og indhold (sprog, medium, genre og emne).

Den lingvistiske opmærkning i form af tags eller metatekst knytter information til de enkelte enheder i teksten og giver grammatiske oplysninger (morfologisk om ordenes kategorier og bøjningsformer eller syntaktisk om ordenes funktion i større enheder), semantiske oplysninger og diskursoplysning (om f.eks. taleture i en dialog), pragmatikoplysninger, prosodioplysninger (om tryk og rytme i talt sprog) samt oplysninger om dialogsegmenter.

For sprogteknologiske anvendelser er det vigtigt at råde over grammatisk opmærkede tekstkorpusser, og her venter der en stor arbejdsindsats forude, selvom enkelte projekter allerede er gennemført eller påbegyndt på området (f.eks. Arboretum- og DDT-træbankerne, PAROLE-ordklasseopmærkning og den CG-syntaktiske opmærkning af 56 millioner ord i Korpus90/2000).

Talesprog skal transskriberes, hvis det skal indgå i korpusser og bruges til træning af computerprogrammer.

3.2 TEKNOLOGISKE RESSOURCER Til de teknologiske ressourcer regnes lyddatabaser, orddatabaser og grammatikker.

3.2.1 Lyddatabaser På grundlag af transskriberede talekorpusser kan man lave en database over de typer eller mønstre af lyde, der forekommer i korpusset, det, der i sprogvidenskaben kaldes fonemer, dvs. den mindste adskillende enhed i talesproget, det, der parallelliseres med et bogstav i skriften.

Et fonem kan beskrives ved en såkaldt Markovmodel, der er en statistisk repræsentation af et mønster i de akustiske sammenhænge, hvori det forekommer i naturlig tale. Ved talegenkendelse modelleres en fonembaseret Markovmodel ofte, ikke ved sandsynligheden for et enkelt fonem, men ved en såkaldt trifon, der i sin simpleste matematiske model beskrives ved hjælp af tre sandsynlighedsfunktioner: for fonemet før, for det fonem, der beskrives, og for fonemet efter. For de to yderste lyddele afhænger sandsynlighedsfunktionen dels af en del af lyden i fonemet selv, dels af lydens påvirkning fra nabofonemet. Sandsynlighedsfunktionen for den midterste lyddel afhænger af lyden i den centrale del af fonemet. Sandsynlighedsfunktionerne trænes og fastlægges på grundlag af frekvensanalyserede talesignaler for alle sprogets fonemer. I praksis omfatter træningen flere tusind trifonmodeller.

Ved akustiske talesignaler forekommer der koartikulation – det forhold, at efterfølgende ord udtales sammenflydende. Koartikulationen fører ofte til flere, lige sandsynlige ordkæder. Hvert ord er sammenføjet af flere trifoner, f.eks. vil ordet *seks* være sammenføjet af de

fire trifoner [_,S,æ], [s,Æ,g], [æ,G,s], [g,S,_]. Ordet otte kan både være sammenføjet af tre trifoner [_,Å,d], [å,D,øh], [d,ØH,_] eller af to trifoner [_,Å,d], [å,D,_].

Ved talesyntese arbejdes der ikke med trifoner, men med difoner, det vil sige sandsynligheden for en lyd efter en anden lyd; ordet seks opbygges af difonerne [_,S], [s,Æ], [æ,G], [g,S] og [s, _].

En forudsætning for, at man kan 'træne' computeren til at genkende og producere fonemer, er således, at der findes store databaser med talesprog i forskellige varianter.

3.2.2 Orddatabaser På grundlag af tekstkorpusser laves orddatabaser eller sprogteknologiske ordbøger, der systematisk beskriver et sprogs ord, hvad angår morfologi, syntaks og semantik. Man kan sondre mellem monolingvale og multilingvale orddatabaser og mellem grammatiske orddatabaser, tesaurusser, ontologier, domænespecifikke termbaser og udtaledatabaser.

Maskinel anvendelse stiller specielle krav til ordbogsdata, fordi computeren kun kan arbejde med fuldt tilgængelige og fuldt eksplicitte oplysninger. Oplysningstyperne kan i stor udstrækning være de samme som i en almindelig ordbog, selvom det sprogteknologiske aspekt påvirker valget af dem (f.eks. er der ikke brug for etymologiske oplysninger i sprogteknologiske ordbøger beregnet for maskinoversættelse). Oplysningerne skal altid formuleres eksplicit, entydigt, udtømmende og formaliseret i overensstemmelse med det fastlagte formelle beskrivelsessprog.

Sprogteknologiske ordbøger kan have meget omfattende og komplicerede oplysningsstrukturer; informationerne kan være forbundet med hinanden f.eks. i en relationel struktur (ved hjælp af tabeller og sammenkædninger), dvs. der kan søges på mange forskellige måder i materialet, og man kan udnytte og restrukturere indholdet efter behov. På grund af den særlige strukturering kalder man dem ofte „orddatabaser" – også for ikke at forveksle dem med elektroniske ordbøger, der er en elektronisk version af en almindelig ordbog for mennesker.

Indholdet i sådan en orddatabase er ikke umiddelbart læseligt eller forståeligt for en menneskelig bruger, fordi oplysningerne er udtrykt i et formelt beskrivelsessprog. På den anden side er det vigtigt at fastholde, at til trods for de skitserede forskelle er der ikke tale om to helt adskilte „ordbogsverdener", men om forskellig realisering af samme. Og netop de særlige krav om systematisk, entydig og præcis beskrivelse af opslagsenheders lingvistiske (f.eks. morfologiske og syntaktiske) egenskaber gør indholdet i en sprogteknologisk ordbog til en værdifuld kilde også for ikke-sprogteknologiske anvendelser. Anvendelse kan ske efter en konvertering af de formaliserede data til naturlige og almenleksikografiske oplysninger, som så kan udnyttes på forskellig vis i traditionel leksikografi.

Et markant eksempel på en dansk sprogteknologisk ordbog er STO, SprogTeknologisk Orddatabase. STO står nu foran sin færdiggørelse pr. 1.marts 2004, og den vil indeholde 80.000 opslagsord fra almensprog og seks udvalgte, afgrænsede fagområder: edb, miljø, sundhed og helse, finans, forvaltning samt handel & erhverv. Ordene i ordbasen vil være mærket med morfologiske oplysninger, og 45.000 af dem vil være mærket med syntaktiske oplysninger, mens 10.000 vil være mærket med semantiske oplysninger.

De „delordbøger", der knytter sig til bestemte fagområder i STO, kan ikke sammenlignes med traditionelle fagordbøger. I modsætning til fagordbøger forudsætter de f.eks. ikke, at brugeren har et bestemt modersmål, og de er heller ikke „retningsbestemte", dvs. beregnet for enten læsere eller producenter af tekster. De kan snarere betegnes som basisordbog til ikke-fagkyndig brug, nemlig maskinel anvendelse, hvorfor informationsindholdet er anderledes end i en fagordbog.

Den afgørende forskel er, at fagordbøger primært fokuserer på fagtermer og er baseret på et fags systematik. De kan således indeholde megen encyklopædisk viden om det pågældende fag, men ofte væsentligt mindre information om bøjningsmorfologi, konstruktions-

muligheder osv. Fagrelaterede ord i STO er beskrevet primært efter de samme datalingvistiske principper som det almensproglige ordforråd.

EU-projektet PAROLE har haft en særlig betydning for udvælgelse og strukturering af STO's oplysningstyper, idet man ved at lægge sig op ad de samme standarder har skabt en orddatabase, der kan sammenlignes med tilsvarende orddatabaser for 10 andre sprog.

Hvis man vil udvikle sprogteknologi af tilstrækkelig høj kvalitet, har man ofte brug for domænespecifikke termbaser, ontologier (modeller over, hvilke elementer der kan tales om inden for et bestemt emne eller område), begrebsnet (netværk af relationer mellem begreber i en ontologi), domænemodeller (modeller over ontologien inden for et domæne) eller tesaurusser (begrebsordbøger).

Orddatabaser er også nødvendige for talegenkendelse, hvis der skal genkendes ikketrivielt sprog, dvs. udtaleundtagelser, som der er mange af på dansk. Sprogteknologiske udtaleordbøger er ordbøger, der systematisk beskriver mønstre af ordenes mulige udtale. Hvert ord foreligger i form af et eller flere mønstre for mulige udtalevarianter.

3.2.3 Grammatikker I et sprogteknologisk system er det grammatikken (eller grammatikkerne), der gør det muligt at beskrive strukturen og funktionen af de sproglige data, som systemet håndterer, hvad enten disse data er talt sprog eller skrevet sprog. „Forståelse" eller produktion af natursprog ved hjælp af automatiske systemer kræver således grammatikker, der kan fortolkes og kompileres som computerprogrammer og bruges af systemets parser.

Grammatikken fungerer som bindeled mellem orddatabaserne, der beskriver ordenes sproglige potentiale, og den samlede mening i en ytring. Grammatikker findes på flere niveauer: På det laveste niveau beskriver grammatik ords bøjning, afledning og sammensætning, på det midterste niveau beskriver grammatikker sætningsbygning og ordføjning, og på et højt niveau kan man forestille sig grammatikker, der beskriver strukturen i hele tekster eller i en dialog. Sprogvidenskabelige sætningsgrammatikker adskiller sig fra hinanden ved at fokusere i forskellig grad på syntaktisk form og kommunikativ funktion eller mening. *Feltskemaanalysen* for dansk, for eksempel, beskriver ords funktion gennem deres position i sætningen, mens *Chomskyansk frasestrukturgrammatik* beskriver sætningers struktur af umiddelbare bestanddele (sætningstræer). Begge systemer egner sig bedst for sprog med fast ordfølge. En grammatik baseret på *grammatiske led* (subjekt, objekt osv.) og en *dependensgrammatik*, der tager udgangspunkt i ords styring af hinanden, er bedre til at analysere sprog uden fast ordstilling (f.eks. latin og russisk), for de kompenserer den større frihed i ordstillingen gennem morfologisk og leksikalsk information, f.eks. kasus eller valens.

I den senere tid er der kommet mulighed for at kombinere træk fra de forskellige paradigmer i én model, og man må ikke forveksle den deskriptive værdi af et grammatisk forskningsparadigme med den metodologiske værdi i computerlingvistikken. I sprogteknologisk henseende er det således vigtigt, at en grammatik er robust og dækkende for hele sproget, effektivt implementeret og ikke mindst korpuskontrolleret. Kun en robust og korpusbaseret dansk computergrammatik vil kunne fungere som hjerte i fremtidens brugernære anvendelser som f.eks. spørgsmålsvarsystemer, andre dialogsystemer, maskinoversættelse, fjernundervisning.

Blandt udfordringerne i de kommende år vil derudover være udviklingen af specialmoduler, der bl.a. kan bestemme, hvad anaforer (f.eks. *han* eller *den*) henviser til; der kan registrere, hvornår der henvises til samme tid, rum og person; der kan genkende navne; der kan entydiggøre den massive flertydighed, der findes i alt naturligt sprog (f.eks. ***på*** *taget,* ***på*** *hospitalet,* ***på*** *fredag,* ***på*** *4 dage, lodderne på vægten*), eller som kan angive, hvorledes leksikalske størrelser eller grammatiske strukturer oversættes fra et sprog til et andet (transfer). Ikke blot skal disse moduler laves, men de skal også integreres med hinanden, før man kan opbygge integrerede teknologier som dialogsyste-

mer, ekspertsystemer eller maskinoversættelse.

En særlig og meget stor udfordring ligger i at lave grammatikker, der tager højde for forskellen i de hyppigste grammatiske strukturer i talt og skrevet sprog. Talesprogsgrammatikker og skriftsprogsgrammatikker må have de samme grundlæggende regler, men det er ikke indtil nu beskrevet, hvorledes reglerne adskiller sig fra hinanden og på hvilke felter.

3.3 BASALE TEKNOLOGIER Basisteknologier bygger på og arbejder videre med de sproglige ressourcer således, at de kan blive mere anvendelige. Der er altså tale om en slags hjælpeprogrammer, som er nødvendige for, at der kan udvikles et egentligt sprogteknologisk produkt. Basisteknologierne segmenterer, analyserer, opmærker eller konverterer det sproglige (talte eller skrevne) input eller sammenligner to parallelle input eller producerer (genererer) et nyt natursprogligt output. De enkelte basisteknologier kan kun fungere på grundlag af sprogressourcerne, f.eks. kan segmentering have brug for en ordbog, en parser have brug for en grammatik mv.

Segmenteringsværktøjer opdeler ord- og talestrømmen i meningsfulde segmenter: ord, led og sætninger. Analyseværktøjer, dvs. parsere, analyserer sammenhængen mellem segmenterne i en større struktur. Opmærkningsværktøjer forsyner segmenterne med den information, som er fremanalyseret af segmenterings- eller analyseværktøjerne. Konverteringsprogrammer konverterer lyd til bogstaver i talegenkendelsen, og bogstaver til lyd i talesyntesen. Parallelliseringsværktøj undersøger ligheder og forskelle mellem to forskellige repræsentationer af det samme. Genereringsværktøj producerer på grundlag af formelle strukturer tekst eller tale.

Forskellige basisteknologiers gennemløb af en sprogressource kan tilføje ny information til ressourcen i flere omgange. Hver ny information giver input til næste gennemløb, f.eks. bliver ordene i en tekst identificeret i første gennemløb, derefter bliver ordenes ordklasser slået op i ordbogen i andet gennemløb, og teksten opmærkes med oplysning om ordklasse og bøjningsform. I tredje gennemløb bliver ordene kombineret til grammatiske ledhelheder, f.eks. ledsætninger ved hjælp af grammatikreglerne. Basisteknologierne kan anvende statistiske metoder, hvor programmer trænes på store datamængder og derved automatisk opbygger grammatikker og sætningsmønstre.

3.3.1 Talegenkendelse En vigtig basisteknologi, som endnu ikke er udviklet tilfredsstillende for dansk, er talegenkendelse: konvertering fra lyd til skrift. Det ultimative formål med automatisk talegenkendelse er at genkende og forstå tale. Indtil nu har det i praksis kun været muligt at opretholde tilstrækkeligt høj genkendelsesrate for begrænsede anvendelsesdomæner. Der har været arbejdet med talegenkendelse siden 1950'erne, men de væsentligste fremskridt er sket efter 1980, hvor man begyndte at anvende statistiske metoder til modellering af det talte sprogs basale elementer, fonemerne. Senere, i 1980'erne, inddrog man desuden lingvistiske metoder, som udnytter kendskab til sprogets grammatiske, semantiske og pragmatiske forhold.

Talegenkendelse benytter tekniskmatematiske metoder til behandling af akustiske talesignaler, herunder statistisk baserede metoder til analyse og beregning af estimater på sandsynlige udfald for ord i talesignalet. Ordenes mulige føjninger genkendes ligeledes ved hjælp af statistiske databaser over sætningsmønstre. For at computeren kan genkende naturlig tale, må den have adgang til store tale- og tekstkorpusser.

Den computerbaserede bearbejdning af det digitaliserede talesignal, f.eks. en talt sætning, indledes med, at computeren på grundlag af en fast opdeling af signalet i tidsintervaller, f.eks. af 20 millisekunders varighed, analyserer signalets frekvensindhold. Derefter sammenlignes resultaterne fra hvert tidsinterval med et stort antal lydmønstre, ordmønstre (ordmodeller for tale) og sætningsmønstre. Med

dagens teknik består et talegenkendelsessystem af analyse- og sammenligningsmoduler samt hertil knyttede databaser over fonemer, ord og sætninger.

Efter denne indledende bearbejdning foretager computeren et skøn ved hjælp af et sandsynlighedsmål og foreslår mindst én mulig opdeling af talesignalet. Først findes en kæde af fonemer svarende til det talte signal, som herefter sammenlignes med de mulige ordudtaler, der findes i databasen med ordmønstre. Det fører til en kæde af ord, som igen sammenlignes med sætningsmønsteret, der indeholder mulige sætninger. På grundlag heraf beslutter computeren hvilken sætning, der med størst sandsynlighed repræsenterer talesignalet.

En talegenkenders ordgenkendelsesrate er et mål for f.eks., hvor mange ord ud af et givet antal ord i et testmateriale der i gennemsnit genkendes korrekt. Sætningsgenkendelsesraten vil umiddelbart være lavere, idet den teoretisk set er lig med ordgenkendelsesraten opløftet i en potens, der er lig med antal ord i sætningen. Udnyttelse af sætningsmønsteret i forbindelse med genkendelsen medfører imidlertid, at sætningsgenkendelsesraten øges i forhold til den teoretiske bundgrænse.

Der er mange faktorer, der afgør et systems genkendelsesrate. Ud over selve genkendelsesprincippet påvirkes raten af en række „ydre" forhold som fysisk støj fra brugsomgivelserne, netværksintroduceret støj i faste og trådløse netværk samt størrelse og sværhedsgrad af både ordforråd og delsprog.

Der er udviklet grundlæggende programmel til dansk talesyntese, men et tilsvarende projekt har man endnu ikke set for dansk talegenkendelse. Man er her henvist til videreudvikling af udenlandsk produceret talegenkendelsesprogrammel, der før ibrugtagning skal „lokaliseres" (trænes) til danske trifoner, det danske sprog og til de enkelte anvendelser.

3.3.2 Talesyntese For talesyntesens vedkommende er det ultimative formål at genskabe naturligt lydende og forståeligt talesignal svarende til en vilkårlig dansk tekst. Indtil nu er man for dansk i stand til at præstere syntetisk dansk tale med god naturlighed og forståelighed.

Man har arbejdet med mekaniske og akustiske modeller for talesyntese helt tilbage i ca. 1780 (Kratzenstein), men udviklingen tog først for alvor fart med computerens indførelse. I starten blev talesyntesens lydgenerering baseret på modellering af de akustiske forhold, når mennesker taler, i form af såkaldt formantbaseret talesyntese. Senere har sammenføjning af indspillede lydsegmenter været meget anvendt. Hvert difon består af et udsnit fra et talesignal med en udstrækning, der begynder omkring midten af et fonem og slutter omkring midten af det efterfølgende fonem. Akustisk set indeholder en difon information om overgangen mellem de enkelte fonemer, hvilket er vigtigt for den syntetiske tales naturlighed. Det lydmæssige basismateriale for difonteknikken består af en database med store mængder akustiske difonmønstre – i alt flere tusind.

Grundlaget for computerens syntetiske talesignal er en skreven tekst, som transformeres via flere forskrifter til en sammenføjet kæde af difoner. Først ekspanderer en algoritme alle forkortelser, specialtegn og lignende til tilsvarende fuldtekst, eventuelt ved samtidig at udnytte information fra den omgivende tekst. Derefter omsættes de enkelte ord til en kæde af fonemer f.eks. ved opslag i en database over udtaler for ord og egennavne samt ved at udføre en syntaktisk analyse af teksten med henblik på bestemmelse af ordkategorier. Videre forsynes fonemkæden med en række markeringer, der bl.a. fastlægger, på hvilken måde det syntetiske talesignal skal udtales med hensyn til sætningsrytme og toneleje, karakteriseret ved henholdsvis fonemernes varighed og talesignalets grundtone (pitch).

Til slut ombrydes fonemkæden til en tilsvarende difonkæde, og de enkelte lydsegmenters styrke, tidsmæssige varighed og overgange modificeres, for at den computerdannede tale skal lyde så naturlig og forståelig som muligt. Det syntetiske talesignal genereres herefter på com-

puteren, hvorfra det på digital form sendes til f.eks. et lydkort i en pc.

Udviklingen af dansk talesyntese er sket ved hjælp af regelbaserede teknikker og ud fra en række statistiske modelbeskrivelser over det danske sprogs difoner. At benytte difoner medfører den begrænsning, at den syntetiske tale ikke fleksibelt kan bringes i stand til nemt at variere talens karakteristika; benyttes der f.eks. difoner, der oprindelig er indtalt af en mandlig taler, vil den syntetiske tale fortsat indeholde denne persons karakteristiske præg. Tilsvarende for syntetisk tale, der er dannet på grundlag af difoner fra en kvindelig indtaler.

Med f.eks. en > 600 MHz pentiumcomputer kan man i dag anvende syntetisk tale i mange praktiske sammenhænge.

3.3.3 Parsere En parser er et program for den procedure, computeren bruger til ved hjælp af en ordbog og en grammatik at genkende, analysere og „forstå" inputtekster.

Det sprogteknologiske fokus har traditionelt ligget på morfologien: lemmatisering (genkendelse af de leksikalske opslagsord) og entydiggørelse af ordklassen, og på syntaksen, dvs. sætnings- eller ytringsgrammatikken. Lemmatisering har allerede fundet anvendelse i f.eks. søgesystemer og i programmer, der trækker tekniske termer ud af tekster, og der opnås i dag ca. 97 % korrekthed med statistiske systemer og 99 % med regelbaserede. Udfordringen ligger nu på det syntaktiske plan og ikke mindst i integrationen af syntaktisk grammatik og semantisk-pragmatiske strukturer. Her kunne succeskriteriet være at kunne tildele hvert ord i en ytring en korrekt syntaktisk eller semantisk kategori med 95 % sikkerhed.

I Danmark foregår der anvendelsesorienteret og korpusbaseret forskning inden for de datalingvistiske skoler: *Head Driven Phrase Structure Grammar* (HPSG), *Lexical Functional Grammar* (LFG) og traditionen *Constraint Grammar* (CG). Som et eksempel på en computerbrugbar feltskemagrammatik kan nævnes sætningsanalysemodulet i den danske del af EUROTRA-projektet.

HPSG er en såkaldt kontekstfri grammatik, der deler sætningen i stadigt mindre bestanddele, ned til ord- eller morfemniveauet, men benytter sig af leksikalsk viden (valens, selektionsrestriktioner) til også at hæfte grammatiske funktionskategorier på de enkelte grene i trædiagrammet.

Lexical Functional Grammar er ikke kontekstfri og skelner i analysen mellem C-struktur (selve sætningstræet, dannet af en kontekstfri grammatik) og F-struktur (over funktioner og andre attributter). Constraint Grammar er en modulær-progressiv metode, der i flere skridt opløser ordenes grammatiske flertydighed og derved som resultat har en grammatisk analyse af sætningen, mens træstrukturer må afledes sekundært. Udfordringen er på dette område at lade skolerne supplere hinanden i stedet for, at de blot konkurrerer.

Både kontekstfrie og kontekstfølsomme grammatikker benytter sig traditionelt af lingvistskrevne regler og detaljerede leksika, men kan til en vis grad forbedres med statistiske metoder. Constraint Grammarteknologien er allerede blevet brugt til opmærkning af danske korpusser, der så igen kan danne grundlag for lingvistreviderede såkaldte træbanker og eksperimenter med probabilistiske metoder i den grammatiske analyse og med maskinlæring, dvs. at maskinen, mens den analyserer tekster, opbygger probabilistiske procedurer for den følgende analyse.

I dette arbejde er der brug for træbanker, dvs. databaser over sætningsmønstre. Det er i praksis indtil nu ikke muligt at få genkendt sætningsmønsteret med en grammatik, der beskriver hele sproget. I anvendelser med talegenkendelse arbejdes der derfor ofte med en stærkt begrænset mængde af sætningsmønstre (med et delsprog). Et fragment af en database over sætningsmønstre kan f.eks. bestå af de mulige ordføjninger, der skal til for at genkende et vilkårligt af tallene fra nul til nioghalvfems. Træning af et parserprogram sker ved hjælp af omfattende mængder elektronisk datamateriale fra det aktuelle anvendelsesdomæne (f.eks. røntgenrapporter, lægejournaler eller juridiske dokumenter).

Et særligt problem udgør den semantiske parsing, dvs. automatisk uddragning af de oplysninger, som ligger i den tekst, som parses. Der findes endnu ikke generelle modeller for, hvad indholdet eller oplysningerne i en tekst egentlig er, men inden for meget begrænsede anvendelsesområder (f.eks. flybestilling, flynavigering) har man fungerende systemer.

Man kan også tænke sig parsere, der analyserer hele tekster i bestemte veldefinerede genrer, således at det, som man kan kalde tekstens indhold, er repræsenteret formelt i computeren. Sådanne analyser er nødvendige for både resumeringsprogrammer og oversættelsesmaskiner, men på disse områder er der ikke nogen programmer for de store sprog, som vi blot kan overtage. Her har Danmark på grund af sin mangfoldighed af sprogvidenskabelige skoler en fordel i forhold til lande, hvor en enkelt skole dominerer.

3.3.4 Transfermoduler Grundlaget for et transfermodul er en parsing af hele teksten, som beskrevet ovenfor, og en tosproget ordbog (transferordbog). Hertil kommer et genereringsmodul, der gendanner teksten på målsproget.

Sådanne transferordbøger skal være meget detaljerede, hvis de skal give en brugbar oversættelse. Der er set nok morsomme eksempler på, hvordan de eksisterende oversættelsesmaskiner laver fejl, der gør teksten helt uforståelig. Forudsætningen for en vellykket oversættelse er en entydiggørelse af de ord, der skal oversættes, og en transferordbog, der specificerer de leksikalske størrelser på begge sprog både grammatisk og semantisk og pragmatisk. Leksikalske enheder er enkeltord og termer, men også sammensatte ord og flerordsenheder. Korrekt behandling af kollokationer har også stor betydning for korrektheden af oversættelsen.

Som sagt ligger hovedproblemet i oversættelse i at bestemme de rette ækvivalenter for ord og flerordsenheder, men der skal mere til for at producere en korrekt oversættelse. Ofte kan sætningsmønsteret fra kildesproget ikke direkte overføres til målsproget (enkle eksempler er f.eks. engelske -ingformer, der svarer til danske relativsætninger, eller det faktum, at artikelbrugen er forskellig fra sprog til sprog). Der er derfor behov for et meget stort kontrastivt arbejde med at kortlægge forskelle og ligheder mellem sprogene og beskrive dem formelt i den bedst egnede grammatik, evt. i en kombination af ovennævnte grammatiktyper. Herefter kan transferregler, der behandler sætningskonstruktioner, beskrives.

Dette arbejde med at konstruere parsere, der kan analysere alle sætningstyper og kan anvende meget store ordbøger „in real time", er kun i sin begyndelse, og der er mange forslag til, hvorledes det skal foregå. Sådanne hurtige og robuste parsere med store og generelle grammatikker og ordbøger findes ikke tilgængelige for dansk. Dette er også et af de steder, hvor statistisk baseret sprogteknologi kan bidrage; en af mulighederne i fremtidens transferforskning er at konstruere et hybridt system, hvor der dels anvendes lingvistisk kontrastiv viden, dels anvendes statistisk viden om, hvordan ord og sætninger normalt oversættes.

3.4 INTEGRERET TEKNOLOGI Kun hvis vi har den nødvendige basisteknologi i veludviklet og tilgængelig form, kan vi udvikle integreret sprogteknologi til praktisk anvendelse for danske brugere.

Forskellig integreret teknologi forudsætter forskellige basisteknologier, f.eks.

- Håndteringen af store informationsmængder vil i fremtiden kunne gøres lettere med udvikling af programmer til tekstresumering, indeksering, tekstklassifikation, intelligent søgning osv. Udviklingen af disse programmer kræver basisteknologierne segmentering, analyse og generering, orddatabaser og ontologier.
- Udvikling af programmer til sikring af entydighed i den faglige kommunikation samt til virksomhedernes vidensorganisation og

videnshåndtering (knowledge management) kræver adgang til bl.a. termbaser, orddatabaser, ontologier og enkel parsing.
- Et semantisk web forudsætter, at nettets tekster opmærkes med en række nøgleord, og dette vil kunne foregå delvis automatisk i takt med, at der udvikles semantiske opmærkningsprogrammer til det. Ontologier anvendes også her.
- Talegenkendelsen for dikteringsværktøjer skal have støtte både af en lyddatabase, et lydbogstavkonverteringsystem og en grammatisk parser.
- For udviklingen af f.eks. dansk stavekontrol og forbedrede søgefaciliteter på nettet er det f.eks. nødvendigt at have adgang til sprogteknologiske ordbøger og grammatisk viden.

Integreret teknologi integrerer mange af basisteknologierne i en applikation. I et interaktivt dialogsystem til naturlig tale, der bl.a. består af en talegenkender, benyttes denne til at konvertere talt input og information til en symbolsk form, der kan forstås af det informationssystem, som skal besvare stillede spørgsmål.

Talegenkenderen benyttes sammen med et semantisk analysemodul, hvis opgave er at udtrække semantisk meningsfulde dele af stillede spørgsmål. Det semantiske modul kan også være indbygget som en del af talegenkenderen.

Talesyntese udnyttes som den del af dialogsystemet, der prompter brugeren for information og giver systemets svar tilbage til denne. Dannelse af systemets svar kræver adgang til et „symbol-til-tekst" modul, der kan konvertere f.eks. tabeloplysninger til meningsfuld dansk tekst.

Et dialogsystem består foruden af talegenkender og talesyntese (hver eventuelt udbygget med sine specielle lingvistiske moduler) også af et modul, der håndterer kontrol af dialogsystemets overordnede opgave. Denne dialogorganisator holder styr på flow i dialogen med henblik på fortsat fokus på den aktuelle del af den samlede dialog. Dialogorganisatoren virker som brugergrænseflade mellem brugeren og systemet.

I praksis forekommer der fejl under genkendelsen, og bruger kan „tale forkert", hvorfor computeren også skal være i stand til at rette fejl i inputtet.

Maskinoversættelse er også et eksempel på integreret teknologi. Her anvendes et analysemodul (parser, grammatik og ensproget ordbog) til at analysere kildesprogsteksten, der udarbejdes et transfermodul som beskrevet ovenfor, og endelig skal der produceres en tekst på målsproget; dette gøres af et syntese- eller genereringsmodul. Det er en særlig udfordring at producere sætninger på målsproget, der er grammatisk korrekte og gengiver kildesprogsætningens mening, samtidig med at sætningerne er formet på en måde, så teksten virker naturlig. Det samme indhold kan på de fleste sprog udtrykkes ved flere forskellige ordstillinger, der alle er grammatisk korrekte; men de vil ikke alle være lige naturlige, når de indgår i en tekst.

3.5 SPROGRESSOURCERS OG BASISTEKNOLOGIERS TILGÆNGELIGHED Basisteknologier og sprogressourcer for dansk er udviklet og tilgængelige i forskelligt omfang. I følgende oversigt er det forsøgt opgjort for hver enkelt type ressource, i hvilket omfang den findes og er tilgængelig. Der er anvendt følgende notation:

I betyder, at basisteknologi eller data ikke findes.
U betyder, at basisteknologi eller data findes, men ikke er tilgængelige.
T betyder, at basisteknologi eller data findes og er tilgængelige.
Tallene fra 1 til 10 anvendes til at beskrive kvaliteten og omfanget af ressourcerne.

Udvalget har ikke gennemført en tilbundsgående undersøgelse af, hvad der findes. Nedenstående bygger således på udvalgets kendskab til eksistens og tilgængelighed af danske sprogteknologiske ressourcer og basisteknologier. Udvalget mener, at nedenstående giver et

meget godt billede af situationen.

I. Dataressourcer			
Uopmærkede talekorpusser	U	4	Sparsomt tilgængelig
Opmærkede talekorpusser	U	4	Sparsomt tilgængelig
Uopmærkede tekstkorpusser	T	3	
Opmærkede tekstkorpusser	U	2	Meget lidt findes
Multilingvale korpusser	T	2	Meget lidt findes
Multimodale korpusser	U	1	Sparsomt tilgængelig
Multimediale korpusser	-		
II. Teknologiske ressourcer			
A. Tale:			
fondatabaser	U	1	
difondatabaser	U	8	
trifondatabaser	U	8	
udtaleordbog	U	8	Sparsomt tilgængelig
B. Tale og skrift			
orddatabaser			
domænespecifikke	T	5	
generelle	T	8	
termdatabaser	T	3	
tesaurusser og ontologier	U	2	Sparsomt tilgængelig
sætningsmønstre	U	1	
sætningsdatabaser (træbanker)			
domænespecifikke	I		
generelle	T	2	
grammatikker	U	6	
III. Basale teknologier			
A. Tale			
Lydsegmentering	U	6	
konvertering fra lyd til fonem	U	5	
konvertering fra lyd til bogstav	U	4	
konvertering fra bogstav til lyd	U	5	Sparsomt tilgængelig
genkendelse af ikke-modersmål	I		
adaption til taler	I		Teknikken findes
adaption til domæne	I		Teknikken findes
prosodigenkendelse	I		En vis forskning findes
difonsyntese	U	8	
syntese af fonemvarianter	U		
prosodiplanlægger	U		
B. Tale og skrift			
sætningsgrænsegenkendelse	T	8	Sparsomt tilg. for tale
navnegenkendelse	T	5	Delvis tilgængelig
lemmatisering	T	9	
morfologisk analyse	U	7	
morfologisk syntese	U	7	
ordklassebestemmelse	T	9	
parsere			
Syntaktisk	T	8	
Semantisk			
Entydiggørelse af ordbetydninger	U	3	Sparsomt tilgængelig
Pragmatisk			
Referentgenkendelse	I		
Tekstgenerering	U	3	
Ordtransfersystemer	U	3	
IV. Integrerede teknologier			
Oplæsningsprogram	T	8	
Tekstningsmaskine (til tv)			
Diktereprogram	I		
Stave- og grammatiktjek	T	8	
Intelligent søgning	T	6	
Tekstredigering	T	6	
Maskinoversættelse, avancerede	U		
Maskinoversættelse, enkle	T	3	

Tekstresumeringsmaskiner	T	4
Automatisk klassifikation		I
Automatisk indeksering	U	2
Vidanshåndtering med sproglig viden		I
Tolkesystem	I	

Det skal bemærkes, at en stor del af de integrerede teknologier, der findes, er kommercielle, og det betyder, at de ikke er tilgængelige for danske forskere til yderligere forskning.

Danske sprogressourcer er i øjeblikket spredt rundt omkring i forskellige institutioner i Danmark. I næsten alle aktive sprogteknologiske forskningsmiljøer i Danmark samles og udvikles der sprogressourcer og søgeværktøjer til forskningsbrug.

Korpuslingvistikken i Danmark har været i langsom fremmarch fra slutningen af 1970'erne med Maegaard og Ruus' (1981) ordhyppighedsundersøgelser til i dag, hvor langt de fleste sproglige forskningsmiljøer anvender eller ønsker at anvende tale- eller tekstkorpusser og korpuslingvistiske metoder i større eller mindre omfang.

Selvom der sandsynligvis har været en ret kontinuerlig aktivitet på området de sidste 30 år, opleves det, som om udviklingen er forløbet i spring med nogle ganske få højdepunkter – styret af de store korpusudgivelser, der har fundet sted undervejs. Således har Bergenholtz' aviskorpus DK87-90 domineret anvendelsen af almensproglige korpusser i de forløbne 10 år, for først nu for alvor at blive afløst af korpusset bag *Den Danske Ordbog* (Det Danske Sprog- og Litteraturselskab 2003-). Dette sidste korpus er i foråret 2003 gjort delvis tilgængeligt via internettet. Inden for det fagsproglige område har vigtige milepæle i korpusudviklingen især været det aftaleretlige (Dyrberg et al. 1991) og det genteknologiske korpus (Lauridsen & Andersen 1993). Inden for taleteknologi har man bl.a. gjort brug af specialkorpusser som Bysoc-talesprogskorpusset. Disse store og gode indsamlingsinitiativer har medført en del korpusbaserede undersøgelser, men antallet og omfanget af tilgængelige korpusser er langtfra tilstrækkeligt.

Et stort problem for korpusbrugerne har været og er stadig tilgængeligheden af teksterne og taleoptagelserne. De fleste tekster og taleoptagelser er beskyttet i henhold til lov om ophavsret, og det betyder, at de fleste korpusser i deres helhed kun må anvendes af en snæver kreds af forskere.

Et andet problem er, at korpusser som rene tekstsamlinger og som rene akustiske data er utilstrækkelige, og at der derfor i stigende grad arbejdes på at berige dem med forskellige former for opmærkning.

Et tredje problem er, at teksterne i forskellige institutioners korpusser er lagret og opmærket i forskellige formater, og at de derfor ofte kun kan bearbejdes med særligt udviklede værktøjer.

Hvad angår tilgængelige monolingvale sprogteknologiske ordbøger, frigives den sprogteknologiske ordbase STO 1. marts 2004. Med sine 80.000 indgange og omfattende syntaktiske (for 45.000 indgange) og semantiske (for 10.000 indgange) mærkning vil den kunne blive et vigtigt grundlag for sprogteknologisk forskning fremover.

3.6 FRA SPROGRESSOURCER OG BASISTEKNOLOGIER TIL VISIONÆR SPROGTEKNOLOGI Der er et aktivt forskningsmiljø i Danmark på det sprogteknologiske felt, og der er skabt et godt grundlag med hensyn til at etablere basisteknologi og sprogressourcer for dansk. Men for at kunne udnyttes effektivt skal ressourcer og basisteknologier i højere grad videreudvikles og tilgængeliggøres på landsplan. Hvis dette ikke sker, kan man frygte, at basisteknologierne slet ikke udvikles for dansk i den målestok, som er nødvendig, og så må danskerne i fremtiden klare sig med engelsk sprogteknologi med de følger, dette har for brugen af det danske sprog.

4. Udfordringer og fokusområder

Den sprogteknologiske forskning står i dag over for en lang række udfordringer, og mange af dem er af en karakter, der kræver tværfagligt samarbejde.

I Danmark er der god grund til at have særligt fokus på forskningsområderne tekst- og talekorpusser, talegenkendelse, grammatik, parsing og formel sprogbeskrivelse samt en dansk oversættelsesmaskine. En opprioritering af disse områder er nødvendig for udviklingen af fremtidens danske sprogteknologi, både af den slags, som udlandet allerede har, og af den mere visionære slags, som kun fremtiden kender.

4.1 FORSKNINGSMÆSSIGE UDFORDRINGER FOR SPROGTEKNOLOGIEN Sprogteknologi omfatter både taleteknologi og natursprogsbehandling. De to felter har udviklet sig fra forskellig basis, men den udvikling, der sker i fagområdet, gør, at det nu er meget vigtigt, at de to fagområder samarbejder tæt. Man kan sige, at så længe hovedproblemet er overhovedet at kunne genkende ord i talegenkendelse, så er de højere niveauer, som f.eks. grammatik, i natursprogsbehandling ikke så relevante. Men så snart man vil integrere talegenkendelse i et avanceret dialogsystem eller i et dikteringsværktøj, bliver de andre niveauer i den lingvistiske beskrivelse yderst relevante. En af de væsentligste udfordringer her er, at de to sider af sprogteknologi ikke blot skal kombinere moduler, men at viden fra begge områder skal integreres på mange niveauer. F.eks. bidrager viden fra sætningsgrammatikken til at afgøre, hvilket ord genkenderen har fat i, ud af flere mulige hypoteser.

Dette samarbejde og den synergi, det afføder, skabes bedst gennem fælles projekter. Det tyske Verbmobil-projekt er netop et eksempel, hvor forskningsadministratorerne gennem formgivningen af et projekt har sikret sig dels et tværfagligt samarbejde mellem taleteknologer og natursprogsforskere, dels et samarbejde mellem forskning og erhvervsliv.

Sprogteknologi har et meget stort potentiale for samarbejde med mange andre discipliner, bl.a. multimodalitet, kommunikation, informationssøgning, datalogi, sprogpædagogik. En del samarbejde er allerede i gang, men der kan skabes meget mere. Nedenfor beskrives nogle af disse tværfaglige muligheder.

4.1.1 Regelbaserede og statistikbaserede systemer Danske forskere i sprogteknologi har en meget bred lingvistisk baggrund, hvilket giver gode muligheder for at vælge de mest hensigtsmæssige teorier til forskellige problemstillinger. Traditionelt har natursprogsbehandlingen anvendt regelbaserede metoder, dvs. man beskriver en grammatik for det danske sprog ved at beskrive alle de regler, der gælder. Forskere i taleteknologi har tilsvarende traditionelt benyttet statistisk baserede metoder til f.eks. genkendelse, det gælder både på ordniveau og på sætningsniveau.

Fordelen ved regelbaserede systemer er, at beskrivelsen af sproget bliver veldefineret: Forskeren analyserer og udarbejder regler og reviderer dem, indtil de passer. Ulempen kan, alt afhængigt af den måde, teorien organiserer regler på, være, at samlingen af regler og undtagelser bliver uoverskuelig. I forbindelse med at computerne er blevet meget kraftigere, er statistisk baseret natursprogsbehandling blevet en alternativ mulighed; der forskes i statistisk analyse af tekster, statistisk maskinoversættelse osv. Fordelen ved disse metoder er, at computeren selv opbygger „reglerne“, baseret på en meget stor tekstmængde, og at det er forholdsvis nemt at ændre systemet, hvis man skal behandle en anden type tekst. Ulempen er, at disse systemer

normalt ikke giver lige så gode resultater som et godt regelbaseret system. Samtidig kræver de meget store tekstmængder for at blive gode.

Internationalt forskes der for tiden i metoder til at skabe hybride systemer, hvor den regelbaserede tilgang bruges, hvor den er bedst, mens den statistikbaserede bruges, hvor den uden vanskelighed giver det rette resultat. Der er ingen tvivl om, at fremtiden for naturlig sprogsbehandling ligger i en helt ny måde at kombinere regelbaserede og statistisk baserede metoder, og at Danmark bør være med til at udvikle dette paradigme.

4.1.2 At forstå en tekst Den ultimative udfordring for sprogteknologien er at få en computer til at „forstå" en tekst, hvad enten der er tale om en talt ytring eller en skrevet tekst. Der er forskel på, hvor dyb en forståelse de enkelte opgavetyper kræver, derfor findes der allerede nu forskningsresultater, der viser, at computeren kan reagere korrekt inden for en meget begrænset verden.

4.1.3 Generel kontra domænespecifik sprogbeskrivelse Lingvistisk forskning søger normalt at beskrive det generelle, og det gælder også sprogteknologisk forskning. Der er imidlertid flere grunde til, at det er meget vigtigt at beskæftige sig med domænespecifik sprogteknologi.

Det er vigtigt, at sprogteknologi faktisk fungerer for den teksttype, den er beregnet til. Det betyder, at forskning i de måder, domænespecifikt sprog adskiller sig på leksikalsk og syntaktisk, er vigtig. For områder, hvor sprogteknologien står over for store faglige udfordringer (f.eks. talegenkendelse, maskinoversættelse), er det en fordel at starte med et udsnit af virkeligheden og teste teoriernes bæredygtighed her. I visse tilfælde kan sådanne udsnit endda give grundlag for en teknologioverførsel og markedsføres som nicheprodukter.

Men domæner kan ofte ikke klart adskilles, hverken fra andre domæner eller fra almensproget, og en fuldstændig atomisering i domænespecifikke systemer ville være helt uhåndterlig, samtidig med at det ville være en videnskabelig falliterklæring. Der ligger altså en udfordring i at beskrive de generelle og de domænespecifikke egenskaber formelt på en måde, så de kan sameksistere.

4.1.4 Flersproget sprogteknologi Som beskrevet er der et udtalt behov for udveksling af informationer og kommunikation på tværs af sprog, et behov, som sprogteknologien skal hjælpe med til at imødekomme. Danmark har en god tradition for kontrastiv lingvistik. Imidlertid er de udfordringer, som sprogteknologien står over for, ret specielle. For at en computer kan benytte en tosprogsordbase korrekt, f.eks. i forbindelse med maskinoversættelse, må det være beskrevet helt nøjagtigt, hvordan man vælger den rigtige målsprogsækvivalent ud af en række mulige. Dette gælder naturligvis ikke blot for enkeltord, men også for flerordsudtryk og på det grammatiske niveau for sætningskonstruktioner.

4.1.5 Orddatabaser, ontologier mv. Orddatabaser (hvad der svarer til ordbøger for mennesker) er helt nødvendige byggestene i et integreret system. Der er investeret i en sprogteknologisk orddatabase STO, som netop er ved at være færdig. Der er selvfølgelig utroligt mange udvidelsesmuligheder ud over selve størrelsen og domænedækningen: STO kan forsynes med udtaleoplysninger, STO kan kobles til tilsvarende orddatabaser på andre sprog, hvorved man får en tosproget eller flersproget orddatabase, mv. Men de vigtigste områder inden for forskningen på ordniveau synes at være dels forskning i kollokationer (formel definition, automatisk genkendelse), dels forskning i ontologier eller begrebssystemer.

Kollokationer er flerordsenheder, hvis syntaktiske og semantiske egenskaber ikke kan afledes af sprogets generelle regler (i „finde sted"

har „finde“ f.eks. ikke bevaret sin generelle betydning). Behandlingen af kollokationer indgår både, når man forsøger at finde betydningen af en sætning, f.eks. i forbindelse med informationssøgning, og når man skal oversætte til andre sprog (jf. „take place“).

Ontologier i betydningen begrebssystemer er et aktivt forskningsemne i Danmark. Det er vigtigt at kende ords og begrebers forhold til hinanden i forbindelse med alle former for forståelse af sætninger (hvis en tekst refererer til en hund og lidt senere omtaler „dyret“, skal computeren kunne indse, at der er tale om det samme). Det er imidlertid meget bekosteligt at udvikle ontologier for alle sprogets ord, så den helt store udfordring for sprogteknologien i disse år ligger i at finde metoder og principper, der kan hjælpe med en automatisk udvikling af ontologier.

4.2 FOKUS PÅ SPROGRESSOURCER OG BASISTEKNOLOGI De fire områder, hvor der er størst brug for sprog- og taleteknologisk grundforskning er:

- samlede, berigede og tilgængeliggjorte tekst- og talekorpusser med tilhørende sproglige analyseredskaber (gerne i form af en Dansk Sprogbank)
- robuste talegenkendelsesprogrammer
- robust parser med tilhørende grammatikker
- oversættelse (med systemdesign, flersprogede ordbøger og transferregler).

Disse fire områder er nødvendige og kan tilsammen dække alle de tidligere nævnte forskningsmæssige udfordringer i et vist omfang. Endvidere udgør de en ramme for et eller flere mere avancerede integrerede projekter, som der gives ideer til i slutningen af dette kapitel.

4.2.1 Tekst- og talekorpusser Inden for sprogteknologi opleves et stigende behov for korpusser i mange forskellige domæner og inden for forskellige genrer, således at nye sprog- og taleteknologiske programmer hurtigt vil kunne tilpasses til nye anvendelsesområder. Dette behov accelererer, fordi moderne sprogteknologi kræver adgang til meget store tekstmængder.

Med hensyn til tekstkorpusser til udvikling af natursprogsbehandling vil det være hensigtsmæssigt, at indsamlingen af tekster i første omgang fokuserer på:

- først og fremmest etablering af fagsproglige korpusser
- parallelle (oversatte) tekster inden for de samme domæner
- dialoger.

Fokus på bestemte fagsproglige domæner skal imødekomme, at danske virksomheder i stigende grad anvender elektronisk dokumenthåndtering og knowledge management.

Fokus på parallelle (oversatte) tekster skal imødekomme et stigende behov for udvikling af værktøjer til maskinstøttet oversættelse. Parallelle tekster er desuden velegnede til systematisk bearbejdning af flersproglig terminologi samt studier af oversættelsesstrategi.

Fokus på dialoger skal sikre et tilstrækkeligt empirisk materiale til at muliggøre forskning i dialogsystemer, dvs. natursproglige grænseflader.

Med hensyn til talekorpusser vil det være vigtigt at orientere sig mod talesprog i typiske genrer. Talesproget former sig i ekstrem grad efter den givne aktivitetstype. Det spontane, uformelle talesprog forholder sig til den velforberedte foredragsstil næsten som to forskellige sprog til hinanden, såvel hvad angår grammatik og leksikalsk selektion som lydlig realisering.

Visse talegenrer er velundersøgte af fonologer og taleteknologer, især de formelle typer, som lægger sig tæt op ad oplæst tale eller informationstæt diskurs (rejsebureau, oplysningstjenester etc.). Andre talegenrer er endnu yderst sporadisk udforsket af grammatikere, fonologer og taleteknologer – paradoksalt nok især dem, som er allermest benyttede, nemlig de spontane og uformelle: den „almindelige samtale“.

Forskning inden for talesprogets strukturerende principper er afhængig af korpusser, som dækker et bredt udsnit af de hyppigste aktivitetstyper. I Norden er det p.t. kun Sverige, der har et velstruktureret talesprogskorpus dækkende et bredt udsnit af aktivitetstyper. Norge er på vej og vil have et sådant korpus om ca. 1 år. Danmark har et lille antal gode transskriptionskorpusser, men ingen, der opfylder ovenstående. En ressource af denne art ville kunne fungere som et stærkt tiltrængt referencekorpus. Derved ville dansk talesprogsforskning (såvel offentlig som privat) blive boostet i betydelig grad – og komme på niveau med den svenske og norske.

I nært samarbejde med internationale korpusinitiativer (f.eks. Text Encoding Initiative/PAROLE) bør man sikre, at fremtidens danske korpusser overholder internationale standarder for opmærkning.

For at de rå eller opmærkede tekst- og talekorpusser kan bruges i forsknings- og udviklingssammenhænge, kræves der programmer, som kan trække data ud af teksterne, præsentere dem på en overskuelig måde og bearbejde dem statistisk. Der findes allerede i dag et antal værktøjer til disse opgaver, men de må udvides, videreudvikles og kobles til programkomponenter til statistisk analyse.

4.2.2 Talegenkendelse De teknologiske muligheder for at anvende talegenkendelse, taleforståelse og talesyntese er i dag så gode, at teknikken inden for få år kan forventes at få en væsentlig udbredelse, f.eks. i forbindelse med talegrænseflader, i forbindelse med mobiltelefoner og håndbårne terminaler, offentlige informationstjenester, i faste og trådløse telenet samt internet og som hjælpemiddel i undervisning og i handicappedes hverdag. Der arbejdes i øjeblikket på at få dansk talesyntese til at lyde mere naturlig, og der er store forventninger til, at taleteknologi vil finde nye anvendelser, f.eks. til simultan undertekstning for døve efter genkendelse af udvalgte emner (ord) fra offentlige nyhedstjenester.

Hvis Danmark skal være med fremme med disse brugervenlige typer af systemer, må der investeres i forskningen i talegenkendelse. For at bringe forskningen videre er det væsentligt, at der etableres forskning på multidisciplinær basis, dvs. teknisk-videnskab, fonetik, lingvistik og probabilistisk mønsterdannelse.

Et talegenkendelsessystems anvendelighed er ikke alene bestemt af genkendelsesraten, idet f.eks. også brugervenlighed med mulighed for online fejlretning indgår. Dette peger på, at forskning i talestyret menneske/maskine-interaktion tilsvarende er en del af det fremtidige kompleks af multiple discipliner, som er nødvendige for at kunne skabe velfungerende og robuste talegenkendere til markedet.

Der er f.eks. innovationspotentiale i talegenkendelsesprogrammel i en opdelt client/server-struktur, hvor client sørger for forbehandling og kodning af akustisk input, og genkendelse foregår i en server, der har langt større hukommelse og processorkraft.

I udlandet arbejdes der intenst på at udvikle teknikker til at spore udvalgte ord og/eller navne automatisk ved hjælp af genkendelsesteknikker på løbende talesignaler fra f.eks. en nyhedsudsendelse. Samme teknikker er bl.a. relevante for automatisk indeksering af denne type data og kan benyttes til automatisk indeksering, hvis den tilsvarende tekst findes i elektronisk form.

4.2.3 Grammatik, parsing, formel sprogbeskrivelse Grammatikker er en del af næsten enhver integreret sprogteknologisk anvendelse. Man kan anvende enkle grammatikker til enkle dialogsystemer, mens hele sproget må dækkes, hvis man skal oversætte f.eks. EU-tekster. Man er altså nødt til at have gode, omfattende danske grammatikker, hvis visionerne om dansk sprogteknologi på højt niveau skal opfyldes.

Inden for feltet parsing og grammatik har Danmark som tidligere beskrevet en fin tradition. Der findes delvise grammatikker for dansk inden for tre af de førende grammatikteorier, CG, LFG, HPSG. Der findes også foreløbigt arbejde inden for statistisk parsing. Imidlertid er der ingen af de eksisterende grammatikker, der bare tilnærmelsesvis

dækker det danske sprog, således som det skrives i dag. Der vil være behov for en større satsning på dansk sprogteknologisk grammatik, og den bør gennemføres i samarbejde mellem de forskellige aktører i Danmark. Da de nævnte teorier er udviklet internationalt, må det internationale samarbejde, der allerede er i gang, fortsættes.

4.2.4 Dansk oversættelsesmaskine For at sikre, at danskerne kan bruge deres modersmål i fremtidens sprogteknologiske løsninger, må der bl.a. udvikles en generel dansk oversættelsesmaskine, der kan oversætte mellem dansk og udvalgte fremmedsprog – i første instans inden for udvalgte anvendelsesområder. Sådanne programmer vil også give udlændinge mulighed for at læse, hvad der er skrevet på dansk, og således give bedre mulighed for at udbrede danske tanker og løsninger. Der er ganske vist kommercielle programmer på vej til dansk, men for det første vil de være begrænsede, og for det andet er det vigtigt, at danske forskere med den store flersproglige baggrund og med den eksisterende baggrund i maskinoversættelse fortsat bidrager til verdensforskningen på dette område.

Danske forskere var blandt de mest aktive i EUROTRA-projektet, og danske forskere har som de eneste formået at videreudvikle EUROTRA-resultaterne. Imidlertid er der nu behov for en nyudvikling af kernen i oversættelsesmaskinen, dvs. der er behov for at trække ny viden om parsing og grammatikker ind, herunder evt. at forske i metoder til hybride systemer, hvor både regelbaserede og statistikbaserede metoder anvendes. Ny forskning i flerordsenheder, semantik og ontologier kan også inddrages.

Engelsk er et meget vigtigt sprog at arbejde med, men også tysk er et meget vigtigt samhandelssprog. Fokusering på tysk vil give mulighed for at inddrage nye samarbejdspartnere blandt fremmedsprogsforskerne på universiteterne.

4.3 FOKUS PÅ VISIONÆR INTEGRERET TEKNOLOGI Nedenfor følger en række eksempler på integrerede systemer, der hver fokuserer på bestemte problemstillinger, hvis løsning kræver en stor forskningsindsats, og hvor man i flere tilfælde vil kunne konstruere nyttige samarbejder med de ovenfor foreslåede fokusområder. Listen er langtfra udtømmende, og de mange muligheder inden for fremtidens integrerede sprogteknologi kender vi slet ikke endnu.

4.3.1 Informationssøgning Udfordringen i informationssøgning ligger i at finde al den relevante information og kun den. Sprogteknologi kan hjælpe med til denne udvælgelse, f.eks. ved anvendelse af ontologisk viden, ved at understøtte tværsproglige søgninger, ved at producere resumeer mv. Det er et område, der allerede nu har stor betydning internationalt, og det vil fortsætte i de kommende år. Danske forskere kan være med til at påvirke udviklingen og være med til at sikre, at der også kan søges i dansk information, og at danskere også kan få svar, selvom det er skrevet på fremmede sprog.

Der er en endnu større udfordring, hvis man også vil anvende inferens i informationssøgningen, således at søgesystemet ændrer karakter og bliver et system, der svarer på spørgsmål, i stedet for blot at finde relevante tekster frem.

Informationssøgning kan kombineres med talestyring, således at man indtaler sit spørgsmål i stedet for at skrive det, og evt. også med talesyntese, så svaret læses op.

4.3.2 Dialogsystemer, grænseflader, menneske-maskine-interaktion Et af de områder, hvor der vil ske en stor udvikling, er inden for interaktive systemer. Efterhånden som talegenkendelse og talesyntese bliver bedre, bliver det muligt at arbejde mere grundigt med dialoghåndtering og med inkorporering af andre modaliteter (gestus, mimik, se nedenfor). Grænsefladerne vil ikke blot være med talt sprog, men også grænse-

flader med forskellige blandinger af talt og skrevet sprog til systemer som f.eks. billetsalg, informationssøgning mv. Forskning i interaktion er et tværfagligt forskningsemne, hvor sprogteknologi er et vigtigt element.

4.3.3 Sprogteknologi i undervisningen E-læring får stor betydning fremover, både i det almindelige undervisningssystem og for specielle grupper, bl.a. indvandrere. Sprogteknologi har en helt speciel rolle at spille i edb-støttet undervisning. Sprogteknologi kan anvendes til at skabe gode grænseflader til undervisningsprogrammer i hvilket som helst fag – dette er den generelle grænsefladeproblematik. Men hertil kommer, at sprogteknologi kan anvendes til at undervise i sprog (CALL – Computer Assisted Language Learning). VISL på Syddansk Universitet er et godt eksempel herpå. Men en systematisk udnyttelse af alle sprogteknologiske forskningsresultater ville give en meget bred vifte af programmer for både dansk og fremmedsprog. Endelig er forskningen i målingen af sprogindlæreres (lørneres) indlæringsniveau nu så fremskreden, at den udfordring, der ligger i at automatisere processen, kan tages op, gerne i internationalt samarbejde. Alle disse emner er forskningsmæssige udfordringer både sprogpædagogisk og sprogteknologisk og skal tages op i et tværfagligt samarbejde.

4.3.4 Tolkemaskine Et interessant forskningsprojekt ville være arbejdet med at udvikle en tolkemaskine, der kunne modtage talt input på dansk, oversætte det og give talt output på engelsk – samt selvfølgelig behandle svaret på samme måde. Dette omfatter talegenkendelse, talesyntese, maskinoversættelse og informationssøgning eller opslag i en database. Det er kun realistisk at udvikle en sådan tolkemaskine for et meget begrænset domæne.

4.3.5 Sprogteknologi og multimodalitet En anden interessant tværfaglig problemstilling er integration af sprogteknologi i interaktive multimedier. Brugeren skal både kunne tale og bruge andre modaliteter (f.eks. håndbevægelser) for at interagere med computeren. Systemets sprogteknologiske komponent skal sørge for at oversætte de sproglige udsagn og tilhørende håndbevægelser til abstrakte meningsrepræsentationer, som computeren kan reagere på – vel at mærke i en sammenhængende dialog. Et sådant forskningsprojekt omfatter talegenkendelse og -syntese, dialogorganisering, parallel og integreret parsing af de sproglige ytringer og af gestus. Interaktive multimedier med sprogteknologi kan anvendes i alle former for brugergrænseflader, i undervisningsprogrammer og spil.

5. En stor strategisk satsning

Danskerne har færre og dårligere sprogteknologiske hjælpemidler til rådighed end folk, der taler engelsk og tysk. Årsagen er først og fremmest, at dansk tales af så få mennesker, at der ikke er et økonomisk bæredygtigt marked for private firmaer til at udvikle programmerne, men også at der ikke har været nær så store offentlige investeringer i sprogteknologi som f.eks. i England, Tyskland og Finland. Situationen fremover vil være yderligere skærpet af, at adgangen til EU-forskningsmidler på dette felt er indskrænket.

Vejen ud af denne situation er, at der investeres offentlige midler i grundforskning i de basale teorier og teknikker, som er grundlaget for alle de mere eller mindre kommercielle anvendelser, der kræver integreret sprogteknologi. Nogle af disse teorier og teknikker er der særlig grund til at fokusere på. Desuden er der grund til forskning i mere avanceret integreret teknologi.

Der arbejdes allerede med sprogteknologi flere steder på landets universiteter, og der er velfungerende miljøer og tilstrækkelig viden til, at arbejdet kan sættes i gang, styrkes og samordnes, hvis de økonomiske rammer udvides, og samarbejdet mellem aktørerne styrkes. Med en stor strategisk satsning på dansk sprogteknologi kan man få en effektiv udnyttelse af det allerede oparbejdede forskningspotentiale. Man kan sikre, at der udvikles god sprogteknologi for dansk nu og i fremtiden, og man kan sikre, at den danske forskning i sprogteknologi fortsat bidrager til verdensforskningen og til vigtig ny erkendelse inden for og uden for humaniora.

En stor strategisk satsning kan bestå i en samlet offentlig investering på 70 mio. kr.

5.1 INVESTERINGENS FORDELING

Tekst- og talekorpusser	10,0 mio. kr.
Talegenkendelse	10,0 mio. kr.
Grammatik, parsing, formel sprogbeskrivelse	7,5 mio. kr.
Maskinoversættelse	20,0 mio. kr.
Visionær, avanceret integreret teknologi	20,0 mio. kr.
Netværk og konsortium	2,5 mio. kr.
I alt	**70,0 mio. kr.**

5.1.1 Satsning på tekst- og talekorpusser En satsning på at sikre tilgængelige og anvendelige tekst- og talekorpusser for dansk kan udmønte sig i en egentlig sprogteknologisk databank, der kan koordinere indsamling, opmærkning og udnyttelse af korpusser for dansk. Forskningsministeriets arbejdsgruppe for IT på dansk har i 2001 foreslået oprettelse af en sådan Dansk Sprogbank, og Ministeriet for Videnskab, Teknologi og Udvikling arbejder for tiden med konkrete planer for sprogbankens formål, opgaver osv. Uafhængigt af disse planer kan man foreslå, at: Sprogbanken skal støtte, igangsætte og koordinere initiativer til indsamling af talt og skrevet sprog, mens hovedparten af indsamlingen foretages af de miljøer, som ønsker at anvende de indsamlede data. Sprogbanken bidrager med specialiseret viden om registrering, lagring og dokumentation, mens de respektive miljøer spiller en central rolle ved udvælgelsen af de relevante tekster og lydoptagelser.

Sprogbanken skal herudover forske i metoder og teknikker til berigelse af de indsamlede korpusser, f.eks. ved at videreudvikle og tilpasse udenlandske opmærkningsprogrammer til dansk.

Sprogbanken skal desuden forske i metoder og teknikker til udnyttelse af det indsamlede datamateriale, herunder værktøjer til statistisk bearbejdning af store tekstmængder og taledata. De juridiske forhold vedrørende brugsretten til sprogbankens indhold må imidlertid afklares, før man evt. lægger data på internettet efter open source-princippet eller lader virksomhederne købe data. I den forbin-

delse kan man evt. lade sig inspirere af de tanker, man har gjort sig i Norge i forbindelse med Norsk Språkbank.

Som første led i oprettelsen af en Dansk Sprogbank foreslås det at investere 10,0 mio. kr., hvilket ca. svarer til 4 stillinger i 5 år.

5.1.2 Satsning på talegenkendelse Hvis Danmark skal være med fremme med hensyn til talegenkendelse, må der investeres i forskningen. For at bringe forskningen videre er det væsentligt, at der etableres forskning på multidisciplinær basis, dvs. teknisk-videnskab, fonetik, lingvistik og probabilistisk mønsterdannelse.

Et talegenkendelsessystems anvendelighed er ikke alene bestemt af genkendelsesraten, idet f.eks. også brugervenlighed med mulighed for online fejlretning indgår. Dette peger på, at forskning i talestyret menneske/maskine-interaktion tilsvarende er en del af det fremtidige kompleks af multiple discipliner, som er nødvendige for at kunne skabe velfungerende og robuste talegenkendere til markedet.

En styrkelse af forskningen i talegenkendelse, både grundforskning og anvendelsesorienteret forskning, kan bestå i oprettelsen af 4 forsker- og ph.d.-stillinger i 5 år til en samlet pris af 10,0 mio. kr.

5.1.3 Satsning på grammatik, parsing, formel sprogbeskrivelse Grammatikker er en del af næsten enhver integreret sprogteknologisk anvendelse. Man kan anvende enkle grammatikker til enkle dialogsystemer, mens hele sproget må dækkes, hvis man skal oversætte f.eks. EU-tekster. Man er altså nødt til at have gode, omfattende danske grammatikker, hvis visionerne om dansk sprogteknologi på højt niveau skal opfyldes.

En koordinering og styrkelse af forskningen i grammatik, parsing og formel sprogbeskrivelse kan bestå i oprettelsen af 3 forsker- og ph.d.-stillinger i 5 år til en samlet pris af 7,5 mio. kr.

5.1.4 Satsning på maskinoversættelse

For at sikre, at danskerne kan bruge deres modersmål i fremtidens sprogteknologiske løsninger, må der bl.a. udvikles en generel dansk oversættelsesmaskine, der kan oversætte mellem dansk og udvalgte fremmedsprog – i første instans inden for udvalgte anvendelsesområder. Samtidig med at der er tale om et satsningsområde, hvor resultaterne kan anvendes i samfundet, er der også tale om et område, hvor en stor forskningsindsats er påkrævet.

Satsning på forskning i maskinoversættelse dansk-engelsk, engelsk-dansk, dansk-tysk, tysk-dansk (med systemdesign, flersprogede ordbøger og transferregler) kan bestå i oprettelsen af 8 forsker- og ph.d.-stillinger i 5 år til en samlet pris af 20,0 mio. kr.

5.1.5 Satsning på visionær, avanceret integreret teknologi En satsning på visionær og avanceret integreret teknologi kan f.eks. fokusere på informationssøgning, dialogsystemer, grænseflader, menneske-maskineinteraktion, tolkemaskine, sprogteknologi i undervisningen eller multimodale systemer. Satsningen kan bestå i oprettelsen af 8 forsker- og ph.d.-stillinger i 5 år til en samlet pris af 20,0 mio. kr.

5.1.6 Oprettelse af netværk og konsortium Det samarbejde, der allerede findes for forskning i sprogteknologi, må styrkes og udvikles, og til dette formål kan der oprettes et netværk for alle sprogteknologiske forskningsmiljøer i Danmark. Gennem netværket kan man påbegynde nye initiativer og udvikling af nye produkter.

Herudover må der etableres brede samarbejdskanaler med det private erhvervsliv, så man sikrer en effektiv teknologioverførsel. Samarbejdsformer som centerkontrakter, Center for IT-forskning, Aalborg Universitets samfinansieringsmodel og lignende foranstaltninger kan være med til at styrke samspillet mellem forskning og erhverv, men man kan med fordel oprette et konsortium for samarbejde mel-

lem universitetsforskningen og de private virksomheder, som har interesser i feltet. Konsortiet kan evt. knytte an til det forskningskonsortium, der i øjeblikket står over for etablering gennem Center for Sprogteknologi.

Oprettelsen af netværk og konsortium kan koste 2,5 mio. kr.

Ordliste

anafor: *ord, der bruges til reference (f.eks. han eller den)*

applikation: *edb-program, som regel en anvendelse (f.eks. dialogprogram)*

begrebsnet: *ontologi formet som net i modsætning til et hierarki*

difon: *lydmodel, der beskriver (modellerer) et akustisk signal, dækkende fra ca. midten af et fonem til midten af det efterfølgende, og på den måde indeholder information om overgangen mellem to fonemer*

diskursoplysning: *om taleture i en dialog m.m.*

domæne: *den sprogbrug, der knytter sig til et bestemt fagområde*

domænemodel: *begrebsnet eller ontologi for et bestemt domæne*

domænetab: *udvikling i retning af, at et sprog bliver fortrængt af et eller flere sprog på et bestemt domæne, f.eks. i den videnskabelige verden. Domænetab fører til, at der ikke længere udvikles nyt ordforråd og ny terminologi for domænet*

e-læring: *undervisningsprogrammer på computer, f.eks. til fjernundervisning*

fonem: *mindste udskillelige lydmodel i et sprog*

fonetik: *udtalelære*

fonologi: *udtalelære*

grammatik: *morfologi og syntaks*

indekseringsprogram: *program til elektronisk indeksering af dokumenter*

integreret sprogteknologi: *avancerede anvendelser af sprogteknologi, f.eks. i applikationer*

intelligent søgeprogram: *program, der kan gennemsøge data og håndtere homografer, synonymer m.m.*

interaktivt dialogsystem: *brugergrænseflade, der genkender, forstår og producerer tale*

korpus: *afgrænset mængde af talte eller skrevne data, som er indsamlet efter klart definerede kriterier, og som er tilgængelige i elektronisk form*

kunstig/indlært intelligens: *computers evne til at efterligne menneskelig intelligens*

leksikalsk viden: *viden om ordenes individuelle egenskaber, f.eks. deres semantik, valens og selektionsrestriktioner*

lemmatisering: *analyse af ord, således at det leksikalske opslagsord fremkommer*
lydmønster/lydmodel: *karakteristisk mønster for et bestemt fonem, difon, trifon mv.*
maskinoversættelse: *automatisk oversættelse mellem sprog*
monolingval: *etsproget*
morfologi: *læren om ordenes kategorier og bøjningsformer*
multimedial: *en applikation, som inddrager flere medier*
multimodal: *en applikation, som inddrager flere udtryksformer, f.eks. sprog, gestus og øjenbevægelser*
multilingval: *flersproglig*
natursprogsbehandling: *maskinel behandling af skrevet sprog*
ontologi: *begrebssystem*
open source: *princip om distribution af kildekode for programmer, ofte knyttet til gratis distribution*
ordbog, sprogteknologisk: *orddatabase, der systematisk beskriver et sprogs ord f.eks. hvad angår morfologi, syntaks, semantik, udtale*
ordmønster/ordmodel: *mønster for en mulig ordudtale*
ordprædiktion: *det at f.eks. software i mobiltelefonen fremsætter gæt på hele ord efter ganske få indtastninger*
parallelsproglighed: *flere sprogs sameksistens i samme domæne*
parser: *program, der sætter computeren i stand til at analysere og strukturere løbende datainput, f.eks. tekst eller tale*
pervasive computing: *computeres indlejring i dagligdags genstande, f.eks. vægge, køleskabe og kaffemaskiner, og deres interaktion med mennesker og hinanden*
pragmatik: *læren om ordenes/enhedernes betydning i kommunikationssituationen*
probabilistisk: *metode, der sætter computeren i stand til at inddrage den beregnede sandsynlighed for forekomsten af et givet fænomen*
prosodi: *tryk og rytme i talt sprog*
relationel struktur: *struktur udtrykt vha. tabeller og sammenkædninger*
semantik: *læren om ordenes/enhedernes betydning*
semantisk web: *The Semantic Web er et projekt under World Wide Web Consortium og vedrører en udvidelse af det nuværende internet, i hvilken al sproglig information er klart defineret og dermed i højere grad tilgængelig for avancerede sprogteknologiske applikationer og pervasive computing*
skærmlæser: *talesyntese for f.eks. navigationsmuligheder på skærmen*
stokastisk: *statistisk*
syntaks: *ordenes grammatiske funktion i større helheder*
sætningsmønster/sprogmodel/delsprog: *mønster for tale på sætningsniveau inden for et bestemt domæne*
sætningstræ: *sætningers struktur af bestanddele, beskrevet i en træstruktur*
talegrænseflade: *brugergrænseflade i interaktivt dialogsystem*
taleteknologi: *maskinel behandling af talt sprog: at genkende og producere talt sprog*
talesyntese: *computerproduceret tale*
talegenkendelse: *computers genkendelse af tale*
termbase: *systematisk beskrivelse af faglige begrebers indhold, afgrænsning og indbyrdes relationer*
tesaurus: *ordbog, hvor ordene er ordnet efter betydning*
transfer: *computerprogram, som kan angive, hvorledes leksikalske størrelser eller grammatiske strukturer oversættes fra et sprog til et andet*
transskribere: *lydskrive, dvs. nedskrive talt sprog som tekst ved benyttelse af fonetisk alfabet*
trifon: *speciel lydmodel dækkende tre fonemer, dele af det akustiske signal fra det centrale fonem samt dele af de to nabofonemer. Trifonen indeholder på den måde information om et fonem i dets udtalekontekst*
udtaleordbog, sprogteknologisk: *ordbog med oplysninger om, hvordan ordene udtales, specielt beregnet til sprogteknologiske applikationer*

Litteratur

Da mange af nedennævnte rapporter ikke har forfattere, har vi valgt at angive disse alfabetisk efter værkets titel.

A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch, D. Binnenpoorte o.a., i: Proceedings LREC 2002, (Third International Conference on Language Resources and Evaluation), Las Palmas de Gran Canaria, Spanien maj-juni 2002.

Anbefalinger fra arbejdsgruppen IT på dansk, Ministeriet for Videnskab, Teknologi og Udvikling, Danmark 2001.

Befolkningens brug af internet 2. halvår 2003, Danmarks Statistik 2003.

Benchmarking HLT progress in Europe. EUROMAP Language Technologies, 2003. Den danske regerings regeringsgrundlag fra 26. november 2001.

Dyrberg, Gunhild og Dorrit Faber og Steffen Leo Hansen og Joan Tournay, *Oprettelse af fagsproglige tekstkorpora – engelsk, fransk, dansk juridisk sprog – aftaleret-* ARK, nr.60, HHK, København 1991.

Forskningsrådenes tværvidenskabelige samarbejde, Supplement til forskningsrådenes strategiplaner 1993-97, 1991.

Frihed til at vælge – handlingsplan for handicappedes IT-brug (1996).

Global Information Technology Report 2003-2004, Suomitra Dutta, INSEAD, Bruno Lanvin, infoDev og Fiona Paua, World Economic Forum.

Handicap ingen Hindring, Handlingsplan for handicappedes it- og telebrug 2002, Ministeriet for Videnskab, Teknologi og Udvikling 2003.

Humanistisk forskning, Strategiplan 1998-2002, Statens Humanistiske Forskningsråd 1996.

It for alle – Danmarks fremtid, it- og telepolitisk redegørelse og handlingsplan 2002, Ministeriet for Videnskab, Teknologi og Udvikling 2002.

KUNSTI – kunnskapsutvikling for norsk språkteknologi. Programplan, Norges Forskningsråd, Norge 2001.

Lauridsen, O., Theis Riiber & Henning Søndergaard: *Erstellung eines dänischen und eines deutschen Textkorpus - Fachsprache Genetik*. Hermes 6, Handelshøjskolen i Århus 1991.

Mikkelsen, Brian, kulturminister: *Sprogpolitisk redegørelse*, 18.12.2003.

National Profile and LE Opportunity Map. Denmark. Afslutningsrapport. Euromap, Center for Sprogteknologi, København 1998.

Nye veje mellem forskning og erhverv - fra tanke til faktura, Ministeriet for Videnskab, Teknologi og Udvikling 2003.

Ny milliardfond skal styrke dansk højteknologi, pressemeddelelse, Ministeriet for Videnskab, Teknologi og Udvikling, 16. januar 2004.

Oplæg til dansk it-forskningsstrategi, Ministeriet for Videnskab, Teknologi og Udvikling 2002.

Maegaard & Ruus: *Hyppige Ord i Danske Romaner*, Gyldendal 1981, samt tilsvarende for andre tekstarter.

Samling og tilgjengeliggjering av norske språkteknologiressursar. Kultur- og kyrkjedepartementet, Norge 2002.

Sprog på spil. Et udspil til en dansk sprogpolitik, Kulturministeriet, Danmark 2003.

Teknisk-Videnskabelig forskning, Strategi- og Handlingsplan 1998-2002, Statens Teknisk-Videnskabelige Forskningsråd 1996.

Tværvidenskabelig forskning, Strategiplan 1998-2002, The Danish Research Councils 1997.