META NET

White Paper Series Hvidbogsserie

THE DANISH LANGUAGE IN THE DIGITAL AGE

DET DANSKE SPROG I DEN DIGITALE TIDSALDER

Bolette Sandford Pedersen
Jürgen Wedekind
Steen Bøhm-Andersen
Peter Juel Henrichsen
Sanne Hoffensetz-Andresen
Sabine Kirchmeier-Andersen
Jens Otto Kjærum
Louise Bie Larsen
Bente Maegaard
Sanni Nimb
Jens-Erik Rasmussen
Peter Revsbech
Hanne Erdman Thomsen

Springer

White Paper Series Hvidbogsserie

THE DANISH LANGUAGE IN THE DIGITAL AGE

DET DANSKE SPROG I DEN DIGITALE TIDSALDER

Bolette S. Pedersen Københavns Universitet
Jürgen Wedekind Københavns Universitet
Steen Bøhm-Andersen Ankiro
Peter J. Henrichsen Copenhagen Business School
Sanne Hoffensetz-Andresen ordbogen.com
Sabine Kirchmeier-Andersen Dansk Sprognævn
Jens Otto Kjærum Prolog Development Center
Louise Bie Larsen Ankiro
Bente Maegaard Københavns Universitet
Sanni Nimb Det Danske Sprog- og Litteraturselskab
Jens-Erik Rasmussen Mikro Værkstedet
Peter Revsbech ordbogen.com
Hanne E. Thomsen Copenhagen Business School

Georg Rehm, Hans Uszkoreit
(udgivere, editors)

FORORD PREFACE

Denne rapport er en del af en hvidbogsserie som omhandler sprogteknologi og dets potentiale. Den henvender sig til undervisere, journalister, politikere og til sprogsamfundet generelt. Det varierer fra sprog til sprog hvor meget sprogteknologi der er tilgængeligt, og hvor meget det bliver brugt. Derfor er det også forskelligt fra land til land hvilken indsats der er behov for. De nødvendige tiltag afhænger af mange forskellige faktorer, såsom et givent sprogs kompleksitet og størrelsen af det samfund hvor det tales.

META-NET, som er et Network of Excellence finansieret af EU-Kommissionen, har gennemført en undersøgelse af sprogresurser og sprogteknologier i de europæiske lande. Undersøgelsen har fokuseret på de 23 officielle europæiske sprog så vel som andre vigtige nationale og regionale sprog i Europa. Undersøgelsen viser at der stadig er lang vej igen for de fleste sprogs vedkommende. En grundig ekspertanalyse og vurdering af den nuværende situation kan hjælpe med til at højne effekten af mere forskning og minimere risikoen for fejlinvesteringer.

META-NET består af 54 forskningscentre fra 33 lande (se s. 70) som arbejder med interessenter fra virksomheder, ministerier, forskningsinstitutioner og europæiske universiteter. Ved at udvikle en strategisk dagsorden for forskning inden for det sprogteknologiske område, arbejdes der mod en fælles vision for hvordan manglerne kan udbedres inden år 2020.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others. The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 74). The analysis focuses on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of future research.

As of November 2011, META-NET consists of 54 research centres in 33 European countries (p. 70). META-NET is working with stakeholders from economy (software companies, technology providers and users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

Forfatterne af dette dokument er forfatterne til sprograpporten for tysk taknemmelige for tilladelsen til at genbruge udvalgt sproguafhængigt materiale fra deres dokument [1].

Udarbejdelsen af denne sprograpport er blevet finansieret af EU's 7. rammeprogram og ICT Policy Support Programme under kontrakterne T4ME (kontrakt nr. 249119), CESAR (kontrakt nr. 271022), METANET4U (kontrakt nr. 270893) og META-NORD (kontrakt nr. 270899). Vi takker Lina Henriksen og Sussi Olsen for den danske oversættelse af den engelske vsersion.

The authors of this document are grateful to the authors of the white paper on German for permission to re-use selected language-independent materials from their document [1].

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899). We thank Lina Henriksen and Sussi Olsen for the Danish translation of the English version.

INDHOLD CONTENTS

DET DANSKE SPROG I DEN DIGITALE TIDSALDER

THE DANISH LANGUAGE IN THE DIGITAL AGE

1

RESUMÉ

Informationsteknologien forandrer vores hverdag. Vi bruger computeren når vi skriver, læser, hører musik og ser billeder og film. Vi har computere i lommestørrelse som vi bruger til telefonopkald, e-mails, informationssøgning og underholdning, uanset hvor vi er. Men hvordan påvirkes sproget af denne massive digitalisering af information, viden og kommunikation? Vil vores sprog forandre sig eller måske endda forsvinde?

Alle vores computere er forbundet i et globalt netværk som hele tiden bliver stærkere. Pigen i Ipanema, toldofficeren i Padborg og ingeniøren i Katmandu kan chatte med venner på Facebook, men det er ikke særligt sandsynligt at de nogensinde mødes i online fora. Hvis de gerne vil vide hvordan man behandler ørepine, vil de alle tjekke Wikipedia for at lære mere om emnet, men ikke engang dér vil de læse den samme artikel. Når Europas internetbrugere i forskellige chatrum diskuterer Fukushima-atomulykkens indvirkning på europæisk energipolitik, gør de det i klart adskilte sprogfællesskaber. Hvad internettet forbinder, holdes stadig adskilt af brugernes sprog. Vil det altid være sådan?

Mange af verdens 6000 sprog vil ikke overleve i et globaliseret, digitaliseret informationssamfund. Man regner med at mindst 2000 sprog vil uddø i de kommende årtier. Andre vil stadig spille en rolle i familier og inden for mindre geografiske områder, men måske ikke i finansverdenen og i den akademiske verden. Hvad er chancerne for det danske sprogs overlevelse?

Ca. 5 mio. har dansk som modersmål, så dansk må anses for at være et relativt lille sprog, i hvert fald sammenlignet med flere andre EU-sprog. I lighed med andre industrialiserede lande er vores hverdag i høj grad påvirket af det engelske sprog. Store internationale virksomheder bruger i stigende grad engelsk som deres virksomhedssprog, og engelsk er ved at blive *lingua franca* inden for højere uddannelser, ligesom det er inden for videnskab og teknologi hvor det har haft den rolle i lang tid.

Man hører ofte kritik af den støt stigende brug af anglicismer, og nogle mennesker frygter ligefrem at det danske sprog er ved at blive gennemsyret af engelske ord og udtryk. Men danske ord og udtryk kan man kun bevare ved rent faktisk ved at bruge dem – ofte og bevidst; lingvistisk polemik om udenlandsk indflydelse og statslig regulering hjælper som regel ikke. Vores største bekymring bør dog ikke være den gradvise anglisering af vores sprog, men snarere at dansk kan forsvinde ud af store dele af vores liv. Videnskab, luftfart og de globale finansmarkeder har reelt brug for et verdensomspændende *lingua franca*, men vi bør værne om vores eget sprog inden for områder som primært angår landets borgere, fx national politik, administrative procedurer, love, kultur og handel.

Et sprogs status afhænger ikke kun af det antal af mennesker, bøger, film og tv-stationer, der bruger det, men også af at det findes i det digitale informationsrum og bruges i softwareprogrammer. Her er det danske sprog temmelig godt placeret: mange internationale softwareprodukter findes i danske versioner, det danske Wikipedia er i vækst, og med mere end 1 million internetdomæner registreret i 2011 er dansk godt repræsenteret på webben set i forhold til befolkningens størrelse.

Men inden for sprogteknologien mangler det danske sprog både værktøjer, teknologier og resurser for at

kunne leve op til morgendagens krav. Der findes en række programmer til talesyntese, talegenkendelse, stavekontrol og grammatikkontrol, men der kræves væsentlige forbedringer hvis man vil sikre en ordentlig funktionalitet i alle relevante sammenhænge. Der findes også programmer til automatisk oversættelse af sprog som dog ofte producerer oversættelser der hverken er sprogligt eller idiomatisk korrekte, hvilket til en vis grad kan forklares med mangel på træningsmateriale i form af parallelle tekstkorpusser som inkluderer dansk. Mere avancerede programtyper som tekstforståelse, sproggenerering og dialogstyring er stadig på et meget tidligt prototypestadie da de typisk kræver resurser med et rigt semantisk indhold i stor skala som slet ikke findes for dansk i dag.

Informations- og kommunikationsteknologien forbereder nu den næste revolution. Efter personlige computere, netværk, multimedier, mobile enheder og 'cloud computing', vil den næste generation af teknologi byde på software som forstår ikke blot talte og skrevne bogstaver og lyde, men hele ord og sætninger, og den vil støtte brugerne langt bedre fordi den taler, kender og forstår deres sprog. Frontløbere for denne udvikling er gratis online tjenester som Google Translate, som oversætter mellem 57 sprog, IBM's supercomputer Watson, som var i stand til at overvinde den amerikanske mester i spillet Jeopardy, og Apples mobile assistent Siri til iPhone, som kan reagere på stemmekommandoer og besvare spørgsmål på engelsk, tysk, fransk og japansk.

Den næste generation af informationsteknologi vil beherske sprog i et sådant omfang at mennesker vil være i stand til at kommunikere ved at bruge teknologi på deres eget sprog. En enkelt stemmekommando vil være nok til at finde de vigtigste nyheder og den vigtigste information fra verdens digitale videnbase. Sprogaktiveret teknologi vil kunne oversætte automatisk eller assistere ved tolkning, resumere samtaler og dokumenter samt understøtte brugere i indlæringssammenhænge. Fx vil den hjælpe immigranter til at lære dansk og dermed til at blive bedre integreret i vores lands kultur.

Den næste generation af informations- og kommunikationsteknologi vil sætte industri- og servicerobotter (som pt. er under udvikling i forskningslaboratorier) i stand til præcist at forstå hvad deres brugere vil have dem til at gøre og derpå stolt rapportere om deres resultater. Sådan et præstationsniveau kræver at vi skal langt videre end de simple leksika, stavekontrolprogrammer og udtaleregler som vi har i dag. Teknologien må bevæge sig fra overforenklede fremgangsmåder og begynde at modellere sproget på en altomfattende måde ved at tage både syntaks og semantik i betragtning for at forstå meningen bag spørgsmål og generere fyldestgørende, relevante svar.

Der er desværre en kæmpe teknologisk kløft mellem engelsk og dansk, og den vokser hele tiden. Hver eneste internationale teknologikonkurrence viser at resultaterne for automatisk analyse af engelsk er langt bedre end for de mere resursesvage sprog som dansk, skønt (eller måske netop fordi) analysemetoderne ligner hinanden eller er identiske. Dette gælder både for videnudtræk fra tekster, grammatikkontrol, maskinoversættelse og en hel række andre anvendelsesområder. Mange forskere regner med at denne tilbagegang skyldes det faktum at metoderne og algoritmerne inden for datalingvistik og sprogteknologi i de sidste 50 år først og fremmest har fokuseret på engelsk. Andre forskere mener imidlertid at det engelske sprog i sig selv er bedre egnet til computerprocessering. I al fald er der ingen tvivl om at vi har brug for en dedikeret, konsekvent og vedvarende forskningsindsats hvis vi vil kunne bruge næste generation af informations- og kommunikationsteknologi inden for de områder af vores privatliv og arbejdsliv hvor vi lever, taler og skriver på dansk.

Efter en relativt succesrig forskningsindsats med adskillige nationale og nordiske projekter inden for sprogteknologi i perioden 1985-2001, er dansk nu begyndt at

halte bagefter, også i det nordiske felt. I de sidste ti år er der ikke blevet givet nogen væsentlig støtte til at fremme og udvikle dansk sprogteknologi, og den uddannelsesmæssige situation er lige så kritisk. Som rapporten her viser, kan vi ikke tillade os at gå i stå. Danmark ligger lavt på den europæiske liste når det drejer sig om tilgængelighed og udvikling af sprogteknologi, og der er et uomgængeligt behov for programmer der kan genoplive og styrke forskningen samt resurse- og teknologiudviklingen på området. Ellers vil vi ikke kunne følge med når en ny generation af teknologi for alvor begynder at beherske de menneskelige sprog. Gennem forbedringer af maskinoversættelse vil sprogteknologien fremover hjælpe med at overvinde sprogbarriererne, men det vil kun fungere mellem de sprog som har evnet at overleve i den digitale verden. Hvis den rigtige sprogteknologi er til rådighed, vil den kunne sikre overlevelsen af selv sprog med et meget lille antal indfødte sprogbrugere. Hvis ikke, vil selv 'større' sprog komme under hårdt pres.

2

EN RISIKO FOR SPROGET OG EN UDFORDRING FOR SPROGTEKNOLOGIEN

Vi er midt i en digital revolution som har markant indflydelse på den måde vi kommunikerer på og på samfundet som helhed. Den seneste udvikling inden for digitale informations- og kommunikationsteknologier bliver undertiden sammenlignet med Gutenbergs opfindelse af trykpressen. Hvad kan denne parallel så fortælle os om fremtiden for EU's informationssamfund og om vores sprog?

Den digitale revolution er sammenlignelig med Gutenbergs opfindelse af den moderne trykpresse.

Gutenbergs opfindelse betød nye gennembrud for kommunikationen og videnudvekslingen; Luthers oversættelse af Biblen til tysk er et godt eksempel herpå. I de efterfølgende århundreder har vi videreudviklet både kommunikative og tekniske færdigheder til bedre at kunne håndtere sprogbehandling og videnudveksling:

- ortografisk og grammatisk standardisering af de store sprog har muliggjort hurtig udbredelse af nye forskningsmæssige og intellektuelle ideer;
- udvikling af de officielle sprog har givet borgerne mulighed for at kommunikere inden for visse (ofte politiske) grænser;
- undervisning i og oversættelse af sprog har muliggjort udveksling på tværs af sprogene;
- opbygning af redaktionelle og bibliografiske vejledninger har givet os kvalitetssikring samt givet os adgang til trykt materiale;
- udvikling af de forskellige medier som fx aviser, radio, fjernsyn og bøger har tilfredsstillet forskellige kommunikationsbehov.

I løbet af de seneste 20 år har informationsteknologien hjulpet os til at automatisere og lette mange af processerne:

- software til desktoppublishing har erstattet maskinskrivning og typografisk opsætning;
- Microsoft PowerPoint har erstattet overhead-transparenter;
- med e-mail kan man afsende og modtage dokumenter hurtigere end med en faxmaskine;
- med Skype kan man få billig internet-telefoni, og man kan opsætte virtuelle møder;
- audio- og videoformater gør det nemt at udveksle multimedie-indhold;
- søgemaskiner giver søgeordsbaseret adgang til websider;
- online tjenester som Google Translate giver hurtige råoversættelser;
- socale medier som fx Facebook, Twitter og Google+ gør det nemmere at kommunikere, samarbejde og dele information.

Selv om disse værktøjer og programmer er nyttige, kan de endnu ikke understøtte et flersprogligt samfund for alle, hvor information og varer kan flyde frit.

2.1 SPROGGRÆNSERNE HÆMMER DET EUROPÆISKE INFORMATIONSSAMFUND

Man kan ikke med sikkerhed forudsige hvordan informationssamfundet vil se ud i fremtiden. Men meget taler for at kommunikationsteknologiens fremskridt vil samle folk med forskellig sproglig baggrund på nye måder. Den enkelte vil blive motiveret til at lære nye sprog, og især vil udviklerne motiveres til at skabe nye sprogteknologiske anvendelser som understøtter en fælles forståelse og fælles adgang til viden. I et globalt informationsrum interagerer flere mennesker på flere sprog med mere indhold og ved hjælp af nye medier. Sociale mediers aktuelle popularitet er kun toppen af isbjerget (fx Wikipedia, Facebook, Twitter, YouTube, og Google+).

> Det globale informationsrum vil betyde flere sprog og mere indhold.

Vi kan i dag hente kolossale tekstmængder fra den ene ende af verden til den anden på ganske få sekunder, og sommetider indser vi først bagefter at en fremsøgt tekst er skrevet på et andet sprog. Ifølge en ny rapport fra EU-Kommissionen køber 57% af EU's internetbrugere varer og tjenester hvor det anvendte sprog ikke er deres modersmål. (Engelsk er det mest almindelige fremmedsprog, fulgt af fransk, tysk og spansk). 55% af brugerne læser tekster på fremmedsprog, mens kun 35% anvender et fremmedsprog til at skrive e-mails eller indlæg på nettet [2]. For nogle få år siden kunne engelsk være blevet internettets lingua franca (fællessprog) – langt størstedelen af teksterne på nettet var nemlig på engelsk – men situationen har ændret sig markant. Mængden af online tekster på andre EU-sprog (såvel som asiatiske og mellemøstlige sprog) er eksploderet.

Denne digitale kløft som hænger nøje sammen med sproggrænserne, har overraskende nok ikke tiltrukket offentlighedens opmærksomhed i særlig høj grad. Men den rejser et meget presserende spørgsmål: hvilke EU-sprog vil trives i det netværksbaserede informations- og videnssamfund, og hvilke er dømt til at forsvinde?

2.2 EU-SPROG I FARE

Trykpressen bidrog til at øge informationsudvekslingen i Europa, men den bidrog også til udryddelsen af mange EU-sprog. Regionale sprog og minoritetssprog blev sjældent trykt, og sprog som cornisk og dalmatisk blev kun overleveret mundtligt, og det har begrænset disse sprogs anvendelsesmuligheder. Vil internettet få samme indflydelse på vores sprog?

> Sprogrigdommen er en af EU's største kulturelle aktiver.

EU's ca. 80 sprog er blandt vore største kulturelle rigdomme, og de er også en vital del af EU's enestående velfærdsmodel [3]. Sprog som engelsk og spansk vil sandsynligvis overleve i det nye digitale verdensbillede under alle omstændigheder, mens andre EU-sprog kunne blive overflødige i et netværksbaseret samfund hvis vi ikke passer på. Denne situation ville svække EU's globale position, og det ville være i modstrid med det strategiske mål som handler om at sikre lige deltagelse for alle EU-borgere uanset sprog.

En UNESCO-rapport om flersproglighed viser at sprog er en væsentlig forudsætning for at kunne gøre brug af grundlæggende rettigheder som fx deltagelse i politiske debatter, i uddannelse og i samfundet generelt [4].

2.3 SPROGTEKNOLOGI ER EN NØGLETEKNOLOGI

Investeringer i sprogbevarende tiltag bestod tidligere primært i sproguddannelse og oversættelse. Ifølge et estimat har EU i 2008 anvendt 8,4 milliarder € på oversættelse, tolkning, software-lokalisering og internationalisering af websider, og det tal ventes at stige med 10% om året [5]. Alligevel dækker dette beløb kun en lille delmængde af hvad der faktisk er brug for til kommunikation mellem sprogene nu og i fremtiden. Den ultimative løsning, som vil sikre både bredden og dybden i morgendagens EU-sprog, er inddragelse af alle relevante teknologier; ligesom vi fx anvender teknologier i forbindelse med transport og udnyttelse af energi.

Sprogteknologien (med fokus på alle former for talt og skrevet sprog) bidrager til at folk kan samarbejde, drive forretning, dele viden og deltage i sociale og politiske debatter, uanset sprog og it-færdigheder. Sprogteknologien indgår ofte i komplekse softwaresystemer og understøtter:

- informationssøgning med en søgemaskine;
- stave- og grammatikkontrol i et tekstbehandlingssystem;
- visning af produktanbefalinger i en online butik;
- talebaseret kørselsvejledning i et navigationssystem til bilen;
- oversættelse af websider ved hjælp af en online tjeneste.

Sprogteknologi består af et antal centrale teknologier som muliggør forskellige former for sprogbehandling i meget store softwaresystemer. Formålet med META-NETs sprog-rapporter er at afdække hvor parate disse kerneteknologier er for hvert enkelt EU-sprog.

> Europa har brug for robust sprogteknologi for alle EU-sprog.

For at fastholde vores position i frontlinjen, skal EU bruge robust sprogteknologi for alle EU-sprog, den skal være til at betale, og den skal være integreret i de vigtigste softwaremiljøer. Uden sprogteknologi vil vi ikke for alvor kunne give brugere oplevelsen af interaktiv, flersproglig og multimediebaseret kommunikation i den nærmeste fremtid.

2.4 SPROGTEKNOLOGIENS MULIGHEDER

I det trykte ords verden var trykpressen det teknologiske gennembrud som betød hurtig kopiering af en tekst. Det besværlige arbejde som bestod i opslag, læsning, oversættelse og sammenfatning af viden, skulle stadig gøres af mennesker. Først med Edison kunne vi optage det talte sprog – og hans teknologi kunne endda kun optage analoge kopier.

Sprogteknologien kan nu automatisere visse processer forbundet med oversættelse, produktion af indhold og håndtering af viden for alle EU-sprog. Sprogteknologien kan også styrke intuitive sprog-/talebaserede grænseflader i hjemmets elektroniske udstyr som fx computere og robotter. Rigtige kommercielle og erhvervsrettede anvendelser er stadig mere eller mindre i støbeskeen, men de nyeste landvindinger peger på helt nye muligheder. Som eksempel kan nævnes at maskinoversættelse allerede nu fungerer ganske godt inden for specifikke domæner, og at nye eksperimentelle applikationer bidrager med flersproglig information, videnhåndtering og generering af indhold på mange EU-sprog.

De første sprogprogrammer, som fx stemmestyrede brugergrænseflader og dialogsystemer, blev udviklet til højt specialiserede domæner, og de havde i reglen begrænset ydeevne. Men der er enorme markedsmuligheder inden for uddannelses- og underholdningsbranchen for at integrere sprogteknologi i spil, på web-

steder om vores kulturarv, i edutainment-pakker, på biblioteker, i simuleringsmiljøer og uddannelsesprogrammer. Mobile informationstjenester, software til computerbaseret sprogundervisning, e-læringsmiljøer, selvevalueringsværktøjer og software til plagiatafsløring er blot nogle af de anvendelsestyper hvor sprogteknologien kan gøre en væsentlig forskel. Facebooks, Twitters og andre sociale mediers popularitet peger endvidere på behov for avanceret sprogteknologi der kan monitorere indlæg, resumere diskussioner, pege på tendenser, afsløre følelsesladede reaktioner, identificere krænkelser af ophavsretten og spore misbrug.

Sprogteknologien kompenserer for vanskeligheder forbundet med sproglig mangfoldighed.

Sprogteknologien repræsenterer enorme muligheder for EU. Den kan på afgørende vis bidrage til at løse de problemstillinger som er forbundet med sproglig mangfoldighed – det faktum, at forskellige sprog eksisterer side om side i virksomheder, organisationer og skoler. EU-borgerne skal nemlig kommunikere på tværs af både sproggrænserne og det indre marked. Sprogteknologien kan både bidrage til at fjerne sprogbarrieren og samtidig understøtte brugen af alle EU-sprog. På længere sigt vil EU's innovative sprogteknologi kunne udgøre et benchmark for vore globale partnere når de på et tidspunkt vil tage de sproglige udfordringer op i deres forskellige sprogsamfund. Sprogteknologi kan betragtes som en form for hjælpeteknologi som udligner de ulemper der er forbundet med sproglig mangfoldighed, og som gør sprogsamfundene mere tilgængelige for hinanden. Endelig er et aktivt forskningsområde anvendelsen af sprogteknologi i forbindelse med redningsaktioner i katastrofeområder hvor effektiv kommunikation kan redde liv. Fremtidens intelligente robotter med tværsproglige kompetencer vil have potentialet til at redde liv.

2.5 SPROGTEKNOLOGIENS UDFORDRINGER

Sprogteknologien har gjort store fremskridt i de senere år, og alligevel går den teknologiske udvikling for langsomt. Almindelige teknologier som stave- og grammatiktjekkere i tekstbehandlingssystemer er typisk monolingvale og findes kun for ganske få sprog. Online maskinoversættelsestjenester er ganske vist velegnede til at generere råoversættelser, men de er utilstrækkelige når der kræves færdige og meget præcise oversættelser. Sproget er en så kompleks størrelse at modellering og afprøvning af sproglig software er en langvarig og dyr affære. EU skal derfor fastholde sin rolle som pioner i mødet med alle de teknologiske udfordringer som er forbundet med et flersprogligt samfund, ved at finde nye metoder til at fremskynde udviklingen på tværs af landene. Disse metoder kan omfatte både nyeste datalogiske fremskridt og teknikker som fx crowdsourcing.

Der skal sættes ekstra skub i den teknologiske udvikling.

2.6 MENNESKERS OG MASKINERS INDLÆRING AF SPROG

For at illustrere hvordan computere håndterer sprog, og hvorfor det er så svært at programmere dem til at bruge det, vil vi kort kigge på den måde mennesker lærer første og andet sprog, og derefter se på hvordan sprogteknologiske systemer fungerer.

Mennesker lærer sprog på to forskellige måder: ved at høre eksempler og ved at forstå de bagvedliggende regler. Et lille barn lærer et sprog ved at lytte til samtaler mellem forældre, søskende og andre familiemedlemmer. I omkring to-års alderen siger barnet de første ord og korte

sætninger. Dette er kun muligt fordi mennesket har en genetisk evne til at imitere, analysere og forstå.

Skal man lære endnu et sprog i en senere alder kræves en større indsats, især fordi barnet så ikke indgår i en sammenhæng med indfødte sprogbrugere. I skolen lærer man i reglen fremmedsprog ved hjælp af øvelser i grammatik, ordforråd og stavning, og disse øvelser tager udgangspunkt i sproglig viden som er udtrykt i abstrakte regler, tabeller og eksempler. Jo ældre man er, jo sværere bliver det at lære et nyt sprog.

Mennesker lærer sprog på to forskellige måder: ved at høre eksempler og ved at lære de sproglige regler.

Der findes to overordnede tilgange til opbygning af sprogteknologiske systemer som begge tager udgangspunkt i "indlæring" af sprog på tilsvarende måder. Den statistiske tilgang indhenter lingvistisk viden fra kolossale tekstsamlinger der fungerer som eksempelmateriale. Man har kun brug for tekst på et enkelt sprog til træning af fx en stavekontrol, men til træning af et maskinoversættelsessystem skal parallelle tekster på to (eller flere) sprog være til rådighed. Maskinlæringsalgoritmen "lærer" på denne måde mønstre for hvordan ord, udtryk og hele sætninger skal oversættes.

Den statistiske tilgang vil i reglen kræve millioner af sætninger, og jo flere analyserede tekster systemet råder over, jo bedre vil oversættelsernes kvalitet blive. Det er en af grundene til at udbydere af søgemaskiner gerne indsamler så meget tekstmateriale som muligt. Stavekontrol i tekstbehandling og i tjenester som fx Google og Google Translate er baseret på statistiske tilgange. Den store fordel ved statistik er at maskinen "lærer" hurtigt hvis bare træningsmaterialet er stort nok, selv om kvaliteten af forskellige årsager kan variere.

Den anden tilgang til sprogteknologi og især til maskinoversættelse er opbygning af regelbaserede systemer. Denne tilgang kræver at eksperter inden for lingvistik, datalingvistik og datalogi først indkoder grammatiske analyser (oversættelsesregler) og kompilerer lister over ordforråd (leksika). Dette kræver både masser af tid og en stor arbejdsindsats. Nogle af de bedste regelbaserede maskinoversættelsessystemer har været under konstant udvikling i mere end tyve år. Fordelen ved de regelbaserede systemer er at udvikleren har større kontrol over sprogbehandlingen. Det er således muligt at korrigere software-fejl systematisk og give detaljeret feedback til brugeren, især i de tilfælde hvor det regelbaserede system anvendes til sprogindlæring. Regelbaseret sprogteknologi findes endnu kun for de store sprog eftersom det er særdeles dyrt at udvikle.

Statistiske og regelbaserede systemers styrker og svagheder komplementerer ofte hinanden, og derfor koncentrerer forskningen sig i dag om hybride tilgange som kombinerer de to metoder. Disse hybride systemer ser lovende ud, men indtil videre har de været mindre vellykkede i erhvervsorienterede anvendelser.

Som ovenfor beskrevet er en stor del af den software som vi bruger i dagens informationssamfund, baseret på sprogteknologi. Selvom sprogteknologien har gjort store fremskridt i løbet af de senere år, er der stadig et enormt potentiale i kvalitetsforbedringer af sprogteknologiske systemer. I det følgende vil vi beskrive det danske sprogs rolle i EU's informationssamfund, og vi vil vurdere sprogteknologiens *state-of-the-art* for det danske sprog.

3

DANSK I DET EUROPÆISKE INFORMATIONSSAMFUND

3.1 GENERELLE FAKTA

Damarks officielle sprog er dansk, og landet har ca. 5.500.000 indbyggere. 90% af disse er etniske danskere med dansk som modersmål. For de sidste 10% findes kun ét officielt etableret minoritetssprog, tysk. Byerne Sønderborg, Åbenrå, Tønder og Haderslev giver officielt mindretalsrettigheder til deres indbyggere; det samlede antal indbyggere med tysk som modersmål udgør 20.000 alene i Sønderjylland (jf. fx [6]).

Udover de dansktalende der er bosat i Danmark, er dansk også modersmål og kultursprog for ca. 50.000 tysk-danske borgere der lever i det sydlige Slesvig. Desuden bevarer danskere som er emigreret til Amerika og Australien, til en vis udstrækning deres modersmål.

En lov fra marts 2006 opstiller betingelserne for den sproglige integration af indvandrere. Indvandrere i besiddelse af en opholdstilladelse og et personnummer kan tilmelde sig tre års danskundervisning. Det er ikke obligatorisk at lære dansk, men hvis man ønsker at få permanent opholdstilladelse eller få dansk statsborgerskab, er det nødvendigt at bestå en danskprøve.

På Færøerne og Grønland garanterer selvstyreloven officiel lighed mellem dansk og færøsk eller grønlandsk, og dansk er et obligatorisk fag i skolen. I Island har dansk været en del af skolernes pensum siden sidst i 1990'erne, og dansk bruges stadig som hjælp til at kommunikere med andre nordiske lande.

Danmark har underskrevet Nordisk Sprogkonvention (1987) som sikrer nordiske statsborgere ret til at anvende deres eget sprog i kontakten med myndighederne i hele Norden. Danmark har også underskrevet den nordiske sprogdeklaration (2006) som er en fælles betænkning fra Nordisk Ministerråd. I den står der at både nationalsprog og mindretalssprog skal støttes og beskyttes, at universiteterne skal have en parallelsprogsstrategi som sikrer brugen af engelsk ved siden af brugen af de nationale sprog, og at statsborgere i de nordiske lande skal have mulighed for at lære deres modersmål såvel som mindst to fremmedsprog. Hvad angår sprogteknologi, understreger deklarationen behovet for maskinoversættelsessystemer, informationssøgningssystemer og avancerede terminologidatabaser for de nordiske sprog.

3.2 SÆRLIGE KARAKTERISTIKA FOR DANSK

Dansk stammer fra gruppen af østnordiske sprog. Ifølge en nyere klassifikation baseret på gensidig sprogforståelse (mutual inteligibility) adskilles moderne talt dansk, norsk og svensk fra de andre nordiske sprog i en skandinavisk sproggruppe.

Dansk udviser flere særlige karakteristika både hvad angår fonologi, ordforråd og syntaks, og alle disse udgør særlige muligheder og/eller udfordringer for sprogteknologien (jf. fx [7, 8, 9]).

Specifikke karakteristika for dansk udgør særlige udfordringer for sprogteknologien.

For taleteknologi kan følgende karakteristika nævnes som relevante:

- et meget stort antal vokaler (29) i talt dansk [10];
- anvendelsen af enhedstryk til at indikere proceslæsningen af et verbum: *læse a'vis*, *spejle 'æg*, *spille 'skak*;
- stød som et betydningsadskillende træk: *stien* (stød) vs. *stigen* (ikke stød).

Desuden udviser det danske ordforråd:

- en stor fleksibilitet til at danne komposita dynamisk, såsom *skiinstruktørsammenslutningssekretærsaspirant*;
- en udstrakt brug af partikler med delvist leksikaliseret betydning: *skrive op*, *skrive ned*, *skrive af*, *skrive ud* etc.

På det syntaktiske niveau tillader dansk sammen med de andre skandinaviske sprog et betydeligt antal flytninger, såsom:

- *Hvem* troede du han sagde at hun kendte _?
- *Denne bog* ved vi hvem der har skrevet _.
- *Denne bog* går der rygter om at du har læst _.
- *Peter* ved jeg ikke om _ vil komme.

Dansk tillader et betydeligt antal syntaktiske flytninger.

3.3 DEN SENESTE UDVIKLING

I de sidste 50 år har det danske sprog været domineret af:

- en tendens til mindre dialektal variation;
- en mindre distinkt udtale af visse lyde i talesproget;
- en vis indflydelse fra engelsk både på grammatik (syntaks og morfologi) og ordforråd;
- en tendens til at foretrække engelsk frem for andre fremmedsprog som tysk og fransk.

Tendensen til mindre dialektal variation favoriserer den københavnske dialekt som standardudtalen brugt over hele landet. Nogle forskere har erklæret de danske dialekter for uddøde allerede, mens andre fastslår at man stadig kan spore regionale variationer. Denne udvikling er blevet styrket af en stærk standardisering af sproget i medierne siden ca. 1950 og en meget lav tolerance i skolesystemet over for dialekter.

Efterhånden som den københavnske dialekt er blevet den dominerende variant af talt dansk, påvirkes hele landet af ændringerne i denne dialekt hen mod en mindre tydelig udtale af især visse vokaler som *a* og *e*. Nogle unge danskere er fx ikke længere i stand til at udtale forskellen mellem ord som *ret* og *rat*. Denne tendens startede imidlertid allerede i middelalderen.

Siden slutningen af Anden Verdenskrig har det engelske sprog haft stigende indflydelse på danske sprogbrugere. Mere end 25% af de kurser der udbydes på de danske universiteter, udbydes på engelsk, og ca. 25% af de store og mellemstore danske virksomheder har valgt engelsk som deres virksomhedssprog. Dette betyder at nye ord ofte er låneord fra engelsk, som fx *governance*, og at danske ord i nogle tilfælde er i konkurrence med deres engelske ækvivalent, fx *deadline* i stedet for *tidsfrist*, *bodyguard* i stedet for *livvagt*. I mange tilfælde kan det dog observeres at de engelske ord bruges til noget andet end de danske, fx bliver ordet *at booke* brugt som ækvivalent til *at bestille*, men *bestille* kan også bruges i betydningen *at afgive bestilling på*. Så den semantiske rækkevidde af *at booke* er smallere end det danske nærsynonym *at bestille*.

I nogle få tilfælde kan man se at engelsk også påvirker den danske syntaks. Fx er ordstillingen i imperativsætninger ved at skifte. En af de mest markante forskelle mellem dansk og engelsk er placeringen af sætningsadverbialer. På engelsk står sætningsadverbialer altid foran

hovedverbet hvor det på dansk i umarkerede sætninger og i imperativer altid står efter verbet. *Please, close the door* svarer til *Luk venligst døren*. Men i løbet af de sidste 15 år er engelsk ordstilling som i *Venligst luk døren* blevet stadig mere almindelig.

Endelig kan det ses at dansk låner nye betydninger af eksisterende ord fra engelsk. Således kunne udtrykket *hænge ud* for 20 år siden kun anvendes i betydningen *hænge tøj ud*, men nu kan det også anvendes i betydningen *hænge ud med vennerne*. Sådanne forandringer er typisk også forbundet med en ændring af verbets valens- eller argumentstruktur, fx i tilfældet *hænge ud* med brug af et præpositionsled i stedet for et objekt.

Pga. det engelske sprogs påvirkning er andre fremmedsprog blevet mindre attraktive for unge mennesker, og antallet af tysk-, fransk-, italiensk- og russiskstuderende er blevet reduceret markant i løbet af de sidste 10 år.

Antallet af tysk-, fransk-, italiensk- og russiskstuderende er blevet reduceret markant i løbet af de sidste 10 år.

3.4 SPROGPLEJE I DANMARK

Det centrale omdrejningspunkt for sprogplejen i Danmark er Dansk Sprognævn som hører under Kulturministeriet. Dansk Sprognævn har tre hovedopgaver:

- at følge det danske sprogs udvikling samt rådgive og informere om det danske sprog. Nævnet fastlægger den danske retskrivning;
- at udgive skrifter om dansk sprog, navnlig vejledninger i brugen af dansk, samt at samarbejde med terminologiorganer, ordbogsredaktioner og offentlige institutioner, der autoriserer eller registrerer stednavne, personnavne og varenavne;
- at samarbejde med sprognævn og tilsvarende organer i de øvrige nordiske lande.

Foruden Dansk Sprognævn redigerer og udgiver Det Danske Sprog- og Litteraturselskab, som er en uafhængig institution delvist finansieret af Kulturministeriet, videnskabelige ordbøger og videnskabelige udgaver af danske tekster. Institutionen udvikler også korpussamlinger for dansk.

Dansk Sprognævn er ansvarlig for sprogplejen i Danmark.

Endvidere har den private institution Modersmål-Selskabet som sin vision at arbejde for at bevare og udvikle dansk. Foreningen udgiver medlemsblade, årbøger og arrangerer foredrag og workshopper om det danske sprog.

Center for Sprogteknologi er det nationale center for sprogteknologi.

Endelig har Center for Sprogteknologi på Københavns Universitet, som er det nationale center for sprogteknologi, det formål at støtte sprogplejen fra den teknologiske vinkel. Centrets mission er at udføre og fremme strategisk forskning og udvikling af anvendelser inden for dansk sprogteknologi. Udover missionen om at sikre god sprogteknologi til danske brugere – og andre brugere af dansk sprog, har centret til formål at skaffe ny viden til Danmark gennem internationalt samarbejde.

Dansk Sprognævn og Det Danske Sprog- og Litteraturselskab har etableret en hjemmeside for dansk sprog *Sproget.dk* som samler information om det danske sprog og betingelserne for dets brug. Hjemmesiden har til formål at tilbyde professionel hjælp ved at informere om lingvistiske emner, og den giver mulighed for at samtidig søgning i flere danske ordbøger, giver adgang til svar på hyppigt stillede spørgsmål og til artikler om forskellige sproglige problemstillinger.

Desuden blev et fælles initiativ om sproglig bevidsthed, den såkaldte Gang-i-Sproget-kampagne, søsat i september 2010 af Dansk Sprognævn og Det Danske Sprog- og Litteraturselskab for den danske regering, og den vil fortsætte i de næste to år. Kampagnen inkluderer en hjemmeside (med en sprogtest), seminarer og tv-programmer om emner inden for dansk sprog.

Andre parametre som angiver niveauet af sprogpleje i Danmark, angår antallet af bøger og aviser der udgives på dansk, samt antallet af tv-kanaler som sender på dansk. Dansk Bibliotekstjeneste skriver i deres årlige statistisk at 7707 titler (inkl. både fiktion og faglitteratur) blev udgivet på dansk i 2010, og 220 danske titler blev oversat til andre sprog samme år. Hvad angår antallet af udkomne aviser, skriver Dansk Oplagskontrol at der i 2010 blev udgivet 34 dagblade på dansk. Ti af disse er nationale dagblade, og de udkommer dagligt i ca. 584.000 eksemplarer [11].

Danmark har seks nationale tv-kanaler, hvoraf tre (DR1, DR2, TV2) bliver betalt via medielicensen. Desuden har lokale tv-kanaler daglig sendetid. Ifølge en lov fra december 2002 om radio- og tv-virksomhed skal "befolkningen ved programlægningen sikres adgang til væsentlig samfundsinformation og debat. Der skal endvidere lægges særlig vægt på dansk sprog og dansk kultur […]". I DR's sprogpolitik står at en betydelig del af programmerne skal være på dansk og skabt til et dansk publikum.

3.5 SPROGET I UDDANNELSEN

Dansk er et obligatorisk fag i danske skoler samt i de danske selvstyrende regioner, Færøerne (hvor dansk også er et officielt sprog) og Grønland. I Island som tidligere var en del af rigsfællesskabet, udbydes dansk som andetsprog parallelt med andre skandinaviske sprog.

Ifølge en kendelse defineres dansk som et nødvendigt skolefag da det giver borgerne mulighed for at deltage i den demokratiske proces. Faget står helt centralt i undervisningen da det lader de studerende blive integreret i det danske samfund. Immigrantbørn er blevet undervist i dansk som fremmedsprog i grundskolen og på gymnasieniveau siden 1993. Desuden findes der kandidatuddannelser i dansk på fem danske universiteter.

På det videnskabelige område bliver engelsk mere og mere toneangivende i Danmark. Mere end 25% af alle universitetskurser bliver afholdt på engelsk, og inden for naturvidenskab bliver kandidatkurser næsten udelukkende afholdt på engelsk. Desuden bliver langt den største del af alle videnskabelige artikler skrevet på engelsk. Med andre ord falder antallet af videnskabelige tidsskrifter på dansk, og disse tidsskrifter har generelt ikke samme videnskabelige status som de internationale.

> Det skal sikres at dansk bevares som et funktionsdygtigt sprog på alle niveauer af uddannelse og forskning.

I sin sprogpolitik understreger Københavns Universitet – det største universitet i Danmark – princippet om parallelisme mellem sprogene engelsk og dansk og skriver at det forudses at engelsk bliver "lingua franca" for forskningen i fremtiden og derved også bliver mere toneangivende inden for undervisning og uddannelse. Det skal dog samtidig sikres at dansk bevares som et funktionsdygtigt sprog på alle niveauer af uddannelse og forskning.

3.6 INTERNATIONALE ASPEKTER

Dansk har siden 1973 været et af de officielle EU-sprog. Udover at have underskrevet Nordisk Sprogkonvention som nævnt ovenfor, er Danmark også med i Nordisk Ministerråds plan for sproglig bevidsthed fra 2007. Her blev sprogteknologi udpeget som en central faktor til beskyttelse og bevaring af vores sprog og kultur. Rådet har nedsat et ekspertpanel til at udarbejde en rapport med en ti-årsplan for hvordan de nordiske lande kan

blive en førende region inden for sprogteknologi. Som en del af planen er flere danske virksomheder og forskere medlemmer af NEALT (Northern European Association for Language Technology), som er en organisation der koordinerer forskellige initiativer og netværker angående uddannelse, forskning og viden om det sprogteknologiske område. Endelig har Københavns Universitet været medlem af ELRA (European Language Resources Association) siden det blev oprettet.

Nordisk Ministerråd har udpeget sprogteknologi som en central faktor til beskyttelse og bevaring af de nordiske sprog og kultur.

3.7 DANMARK PÅ INTERNETTET

Ifølge de seneste statistikker fra 2010 findes der 4.750.000 internetbrugere i Danmark, hvilket udgør 86% af befolkningen [12]. Blandt unge mennesker er andelen af brugere endda højere. Dette er en meget høj procentdel sammenlignet med de øvrige EU-lande, og det viser at danskerne generelt er dygtige teknologibrugere. Hvad angår danske internetdomæner findes der mere en 1 mio. registrerede domæner i 2011 [13]. Danskere foretrækker at bruge internetsider på dansk; de fleste offentlige tjenester findes dog både på dansk og engelsk.

Danskerne er generelt dygtige teknologibrugere.

Den udbredte brug af internettet i Danmark er vigtig for sprogteknologien på to måder. For det første udgør den store mængde digitale sproglige data en rig kilde til analyse af naturlig sprogbrug, især ved indsamling af statistisk viden. For det andet tilbyder internettet en bred vifte af anvendelsesområder for sprogteknologi.

Internettet anvendes fortrinsvis til søgning, hvilket involverer automatisk behandling af sprog på flere niveauer. Dette aspekt involverer sofistikeret sprogteknologi – også for dansk – og vi vil se nærmere på det i anden halvdel af denne rapport.

Internetbrugere og leverandører af indhold til hjemmesider kan have gavn af sprogteknologi på mindre åbenlyse måder, fx når sprogteknologi anvendes til automatisk oversættelse af en hjemmeside fra et sprog til et andet. Når man tager i betragtning, hvor høje omkostningerne er ved manuel oversættelse af sådanne sider, udvikles og anvendes der forholdsvis lidt brugbar sprogteknologi sammenlignet med det forventede behov.

4

SPROGTEKNOLOGISK STØTTE TIL DANSK

Sprogteknologi er den teknologi der ligger bag softwareudvikling til at håndtere naturligt sprog. Derfor går denne type teknologi ofte under betegnelsen "natursprogsbehandling". Naturligt sprog findes i mundtlig og skriftlig form. Hvor tale er den ældste og naturligste form for menneskelig kommunikation, bliver kompleks information og størstedelen af den menneskelige viden registreret og formidlet i skreven form. Tale- og teksteknologi behandler og producerer disse to forskellige slags sprog skønt de begge gør brug af ordbøger, grammatikregler og semantik. Det betyder at sprogteknologi forbinder sprog med flere former for viden uafhængigt af det medie (tale eller tekst) som det bliver udtrykt i (se figur 1).

I vores kommunikation kombinerer vi sprog med andre former for kommunikation og andre informationskanaler. Fx kan tale inddrage gestik og ansigtsudtryk. Digitale tekster er forbundet med billeder og lyd. Film kan indeholde både skrevet og talt sprog. Tale- og teksteknologi overlapper og interagerer altså med megen anden teknologi som indgår i behandlingen af multimodal kommunikation og multimediefiler.

I det følgende vil vi diskutere sprogteknologiens vigtigste anvendelsesområder, dvs. sprogkontrol, informationssøgning på nettet, taleteknologi og maskinoversættelse. Dette inkluderer softwaresystemer og basisteknologier som

- stavekontrol
- forfatterstøtte
- computerstøttet sprogindlæring
- informationssøgning
- videnuddragelse
- tekstresumering
- spørgsmål/svar
- talegenkendelse
- talesyntese

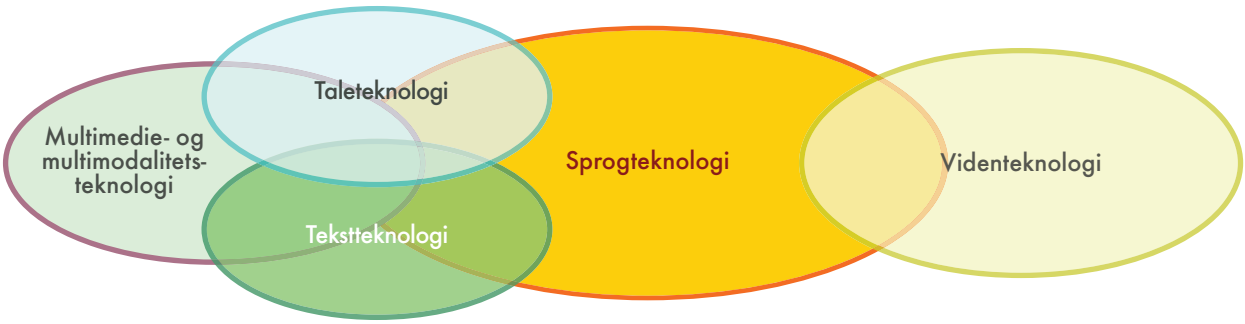
Sprogteknologi er et etableret forskningsområde med en omfattende mængde af introducerende litteratur. Den interesserende læser henvises til følgende referencer: [9, 14, 15, 16, 17].

Før vi diskuterer disse anvendelsesområder, vil vi kort beskrive arkitekturen i et typisk sprogteknologisk system.

4.1 SYSTEMARKITEKTUR

Software til natursprogsbehandling består typisk af flere komponenter der afspejler forskellige aspekter af sproget. Figur 2 viser en meget forenklet arkitektur for et typisk system til behandling af tekst. De første tre moduler beskæftiger sig med tekstinputtets struktur og betydning:

1. Præprocessering: rydder op i data, fjerner formatering, genkender inputsproget osv.
2. Grammatisk analyse: finder verbet og dets objekter, attributive led og andre ordklasser samt afdækker sætningsstrukturen.
3. Semantisk analyse: entydiggør (beregner den korrekte betydning af et ord i en given kontekst); afgør hvad anaforer og refererende udtryk som *hun*, *bilen*



1: Sprogteknologi i kontekst

osv. refererer til; og udtrykker sætningens betydning i en maskinlæsbar form.

Når teksten er analyseret, kan opgavespecifikke moduler udføre fx automatisk resumering og databaseopslag. Dette er en forsimplet beskrivelse af systemarkitekturen og kompleksiteten ved sprogteknologiske systemer.

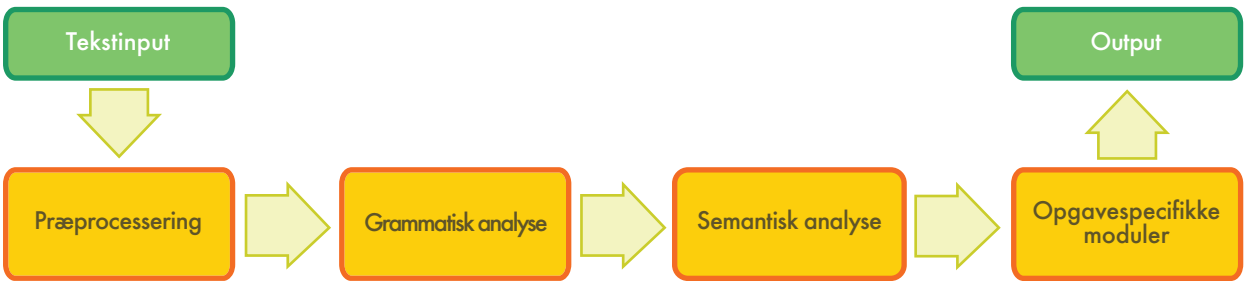
Introduktionen til sprogteknologiens centrale anvendelsesområder efterfølges af et kort overblik over sprogteknologisk forskning og uddannelse i dag og afsluttes med en oversigt over tidligere og nuværende forskningsprogrammer. Derpå vil vi præsentere en ekspertbedømmelse af vigtige sprogteknologiske værktøjer og resurser målt på flere dimensioner såsom tilgængelighed, modenhed og kvalitet. Den generelle situation for sprogteknologi for dansk bliver opsummeret i en tabel (figur 8) på side 28. Værktøjer og resurser som er skrevet med fed i teksten, optræder også i figur 8 i slutningen af dette kapitel. Her sammenlignes også sprogteknologistøtte til dansk med de øvrige europæiske sprog som beskrives i denne hvidbogsserie.

4.2 CENTRALE UDVIKLINGSOMRÅDER

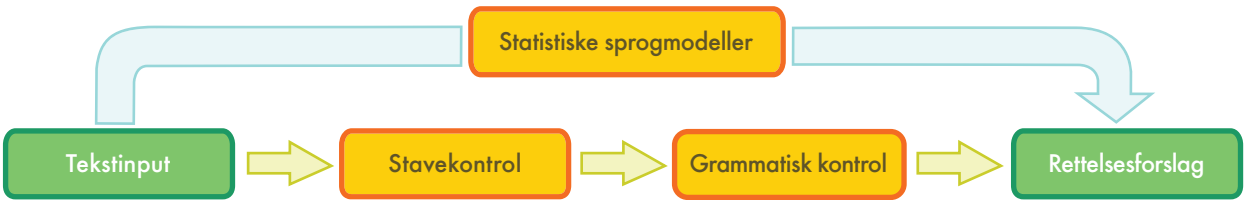
I dette afsnit sætter vi fokus på de vigtigste sprogteknologiske værktøjer og resurser og giver en oversigt over sprogteknologiske aktiviteter i Danmark.

4.2.1 Sprogkontrol

Alle der har anvendt et tekstbehandlingssystem som Microsoft Word, ved at det har en stavekontrol som mar-



2: Arkitekturen for et typisk system til behandling af tekst



3: Sprogkontrol (statistisk; regelbaseret)

kerer stavefejl og foreslår rettelser. De første stavekontrolprogrammer sammenlignede dokumentets ord med en ordbogs korrekt stavede ord. I dag er disse programmer langt mere avancerede. Ved at anvende sprogafhængige algoritmer til **grammatisk analyse** finder de både fejl der har relation til morfologi (fx flertalsdannelse) og til syntaks såsom et manglende verbum eller kongruensfejl mellem subjekt og verbal (fx *hun *skrive et brev*), men mange stavekontroller vil fx ikke finde fejl i følgende engelske tekst [18]:

> *I have a spelling checker,*
> *It came with my PC.*
> *It plane lee marks four my revue*
> *Miss steaks aye can knot sea.*

Sådanne fejl kræver normalt en analyse af konteksten. Denne type analyse bygger enten på sprogspecifikke **grammatikker** som eksperter har kodet ind i softwaren, eller på en statistisk sprogmodel. I sidstnævnte model beregnes sandsynligheden for at et bestemt ord optræder i en bestemt position. Fx er *stort hus* en langt mere sandsynlig ordsekvens end *stor hus*. En statistisk sprogmodel kan laves automatisk ved at anvende en stor mængde (korrekte) sproglige data (et såkaldt **tekstkorpus**). Hovedparten af de systemer som anvender disse to fremgangsmåder, er blevet udviklet med engelske data.

Ingen af de to modeller kan let overføres til det danske sprog med dets særlige karakteristika som kompositumdannelse og et rigere bøjningssystem. Derfor var det en alvorlig fejl ved de første danske stavekontroller at de ukorrekt fejlmarkerede produktive komposita som fx *pasningsordning*. Hvis et sådant kompositum ikke var leksikaliseret i en ordbog eller på en ordliste (hvilket produktive komposita normalt ikke er), medførte det en fejlmarkering. Disse fejlagtige markeringer skyldtes at der endnu ikke fandtes gode kompositumopdelingsprogrammer som kunne tjekke hvert enkelt ord i kompositummet. Desværre har denne mangel i de tidlige stavekontroller ført til en stigning i antallet af stavefejl; folk er under indflydelse dels af at sammensatte ord på engelsk i reglen skrives i to ord, dels af det faktum at det ikke fører til fejlmarkering med en dansk stavekontrol hvis et dansk sammensat ord skrives i to ord. De seneste stavekontroller for dansk er blevet forbedret på dette punkt.

Danske grammatiktjekkere er derimod stadig på et temmelig tidligt stadie. De er generelt i stand til at finde simple grammatiske fejl såsom konkordansfejl inden for kort afstand som i **den røde hus*, mens andre grammatiske fejl ikke identificeres, som fx **jeg var kede af du ikke kom*. OpenOffice er også begyndt at levere værktøjer til sprogkontrol af dansk; Magenta har integreret flere danske open-source leksikalske resurser i dokumentbehandlingsværktøjet, bl.a. synonymopslag.

Med særligt henblik på undervisning er Mikro Værkstedet en af hovedaktørerne på markedet for digitale undervisningshjælpemidler, heriblandt læseværktøjer for ordblinde og skrivehjælp. Desuden bør Ordbogen.com nævnes fordi de giver adgang til mange forskellige online ordbøger.

Sprogkontrol er ikke begrænset til tekstbehandlingssystemer; det bruges også i forfatterværktøjer.

Brugen af sprogkontrol er ikke begrænset til tekstbehandlingssystemer; det bruges også i "forfatterværktøjer", som især anvendes til skrivning af manualer og anden dokumentation for avanceret it, sundhed, teknik og lignende som skrives efter særlige standarder. Kundeklager over forkert brug og erstatningskrav som et resultat af dårligt forstået instruktion har motiveret virksomhederne til at prioritere kvaliteten af den tekniske dokumentation samtidig med at de henvender sig til det internationale marked (via oversættelse eller lokalisering). Fremskridtene inden for natursprogsbehandling åbner mulighed for udvikling af forfatterværktøjer som hjælper forfattere af teknisk dokumentation med at bruge det ordforråd og den sætningsstruktur der er i overensstemmelse med branchens retningslinjer og (virksomhedens) terminologi.

Udover stavekontrol og forfatterhjælp er sprogkontrol også vigtig inden for computerstøttet sprogindlæring. Og sprogkontrolprogrammer anvendes også til automatisk korrektion af forespørgsler i browsere, som man ser det i Googles *Did you mean ...*-forslag.

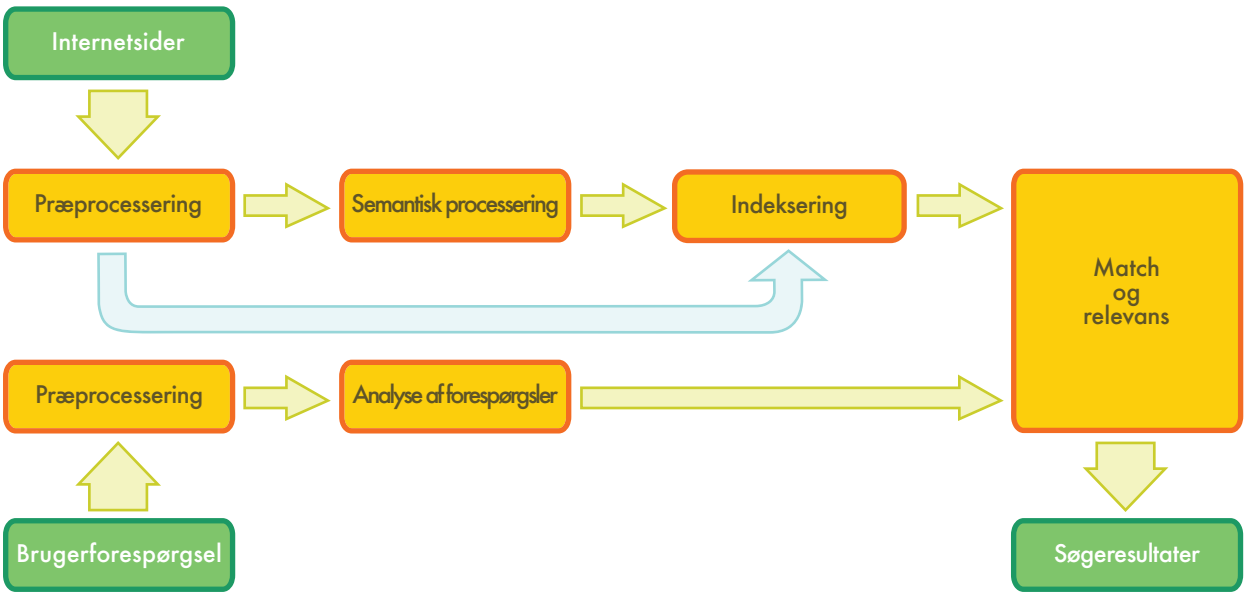
4.2.2 Internetsøgning

Søgning på internettet, på intranet og i digitale biblioteker udgør i dag sandsynligvis den hyppigste anvendelse af sprogteknologi, og dog er denne anvendelse i det store og hele underudviklet. Googles søgemaskine som blev lanceret i 1998, behandler nu 80% af alle forespørgsler [19]. Verbet *at google* har endda en indgang i Den Danske Ordbog. Google-søgegrænsefladen og dens resultatsider er ikke blevet ændret væsentligt siden den første version. Men den nuværende version giver mulighed for korrektion af stavefejl og har inkorporeret grundlæggende semantiske søgemuligheder som kan forbedre nøjagtigheden af søgningen ved at analysere betydningen af termer i en søgekontekst [20]. Googles succeshistorie viser at en stor mængde tilgængelige data og effektive indekseringsteknikker kan give tilfredsstillende resultater ved at bruge en statistisk baseret fremgangsmåde.

Avanceret informationssøgning kræver at man integrerer mere lingvistisk viden for at kunne udføre tekstforståelse. Forsøg med systemer der anvender **leksikalske resurser** som maskinlæsbare thesauri eller ontologiske sprogresurser (fx Princeton WordNet for engelsk eller det tilsvarende danske wordnet, DanNet), har forbedret deres søgeresultater ved at bruge synonymer for de originale søgetermer (fx *autoforsikring*, *bilforsikring* og *kaskoforsikring* eller endnu mere løst relaterede termer).

Den næste generation af søgemaskiner skal indeholde meget mere avanceret sprogteknologi.

Den næste generation af søgemaskiner skal indeholde meget mere avanceret sprogteknologi, især for at kunne håndtere søgninger formuleret som spørgsmål eller andre sætningstyper i stedet for en liste af nøgleord. Søgningen *Giv mig en liste over alle virksomheder som er blevet overtaget af andre virksomheder inden for de seneste fem år* kræver en syntaktisk og **semantisk analyse**. Systemet skal også levere et indeks for at kunne fremsøge de relevante dokumenter i en fart. Et tilfredsstillende svar vil kræve syntaktisk parsing for at kunne analysere sætningens grammatiske struktur og afgøre at brugeren er interesseret i virksomheder der er blevet opkøbt, ikke virksomheder der har opkøbt andre virksomheder. Hvad angår udtrykket *de seneste fem år* skal systemet afgøre hvilke år der er relevante. Og søgningen skal matches mod et vældigt stort antal ustrukturerede data for at finde frem til den relevante information som brugeren er interesseret i. Dette kaldes informationssøg-



4: Internetsøgning

ning og involverer fremsøgning og prioritering af relevante dokumenter. For at kunne generere en liste af virksomheder skal systemet desuden kunne identificere et virksomhedsnavn i et dokument, en proces der kaldes "navnegenkendelse".

En mere krævende udfordring er at matche en forespørgsel i ét sprog med dokumenter i et andet. Tværsproglig informationssøgning involverer automatisk oversættelse af forespørgslen til alle mulige sprog og derpå oversættelse af resultaterne tilbage til det sprog brugeren anvender.

Nu da data i stadig højere grad findes i ikke-tekstlige formater, er der behov for tjenester der kan håndtere multimedie-informationssøgning i billeder, lydfiler og videodata. Lyd- og videofiler kræver et talegenkendelsesmodul for at omdanne tale til tekst (eller til en fonetisk repræsentation), som så kan matches med en brugerforespørgsel.

Små og mellemstore virksomheder (SMV'er) i Danmark som Ankiro, ScanJour, LAT Consulting, Findwise, RD-Fined og andre udvikler og anvender med succes søgeteknologier der er skræddersyet til særlige virksomhedsbehov. Udviklingsarbejdet hos fx Ankiro har fokus på avancerede søgemaskiner til emnespecifikke portaler ved at udnytte emnerelevant semantik. Pga. de store krav til computerkraft er disse søgemaskiner kun økonomisk anvendelige på relativt små korpusser. Processeringstiden kan let overstige den tid en almindelig statistisk søgemaskine, som fx Google, ville bruge. Disse søgemaskiner stiller også høje krav til domænespecifik modellering hvilket gør det umuligt at anvende disse mekanismer på hele internettet. Desuden er de teknologier der udvikles i disse sammenhænge, generelt ikke offentligt tilgængelige for yderligere forskning og udvikling. Mange danske hjemmesider linker til en Google-søgemaskine som den eneste søgefacilitet pga. praktiske forhindringer af denne art.

Eksperimentelle, ontologibaserede søgemaskiner er blevet udviklet af flere danske universiteter, bl.a. Roskilde Universitet. OntoQuery og SIABO prototyperne er eksempler på sådanne eksperimentelle søgemaskiner som arbejder på mindre domæner med en rig ontologisk re-

præsentation. Men disse kan ikke umiddelbart skaleres op til større domæner.

4.2.3 Taleteknologi

Taleteknologi handler om at skabe grænseflader som giver brugere mulighed for at arbejde interaktivt i talt sprog i stedet for at anvende grafisk display, tastatur og mus. I dag anvender virksomheder stemmestyrede grænseflader (VUI) som regel til halv- eller fuldautomatisk telefonservice for deres kunder, ansatte og partnere. Forretningsområder der er meget afhængige af VUI-systemer, indbefatter bankverdenen, leverandørkæder, offentlig transport og telekommunikation. Anden brug af taleteknologi omfatter grænseflader til bilnavigationssystemer og brug af talt sprog som alternativ til grafiske grænseflader eller touch-screen-grænseflader i smartphones.

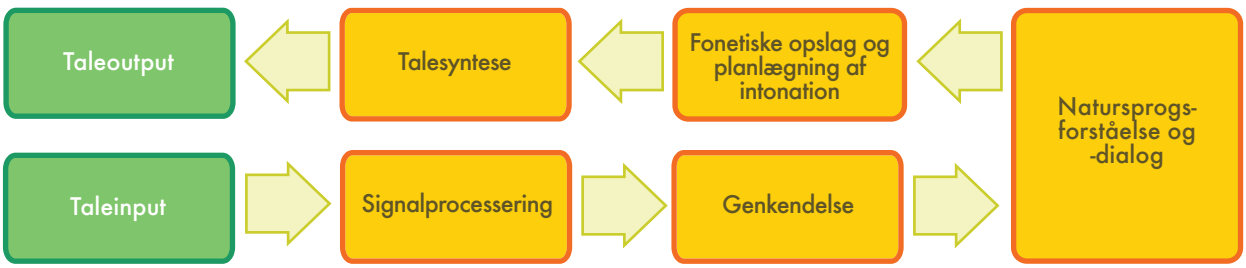
Taleteknologi omfatter fire teknologier:

1. Automatisk **talegenkendelse** afgør hvilke ord der bliver sagt i en given sekvens af lyde udtalt af en bruger.
2. Natursprogsforståelse analyserer den syntaktiske struktur i en brugerytring og fortolker den i henhold til det pågældende system.
3. Dialogstyring afgør hvilken handling der skal udføres, afhængig af brugerinputtet og systemfunktionaliteten.
4. **Talesyntese** (tekst til tale eller TTS) omdanner systemets svar til lyde for brugeren.

En af de største udfordringer for automatiske talegenkendelsessystemer er præcis genkendelse af brugerens ord. Det betyder at man må begrænse omfanget af brugerens ytringer til en liste af nøgleord eller manuelt udvikle sprogmodeller som dækker et stort udvalg af natursprogsytringer. Ved at anvende maskinlæringsteknikker kan sprogmodeller også genereres automatisk ud fra **talesprogskorpusser**, dvs. store samlinger af tale-lydfiler og teksttranskriptioner. Begrænsning af mulige ytringer vil normalt betyde en ufleksibel anvendelse af grænsefladen og kan gå ud over brugernes accept; men at skabe, tune og vedligeholde detaljerede sprogmodeller vil øge omkostningerne betydeligt. VUI'er som anvender sprogmodeller, og som tillader brugeren at udtrykke sit ønske på noget der minder om natursprog – tilskyndet af en *Hvordan kan jeg hjælpe dig?*-hilsen – bliver ofte bedre accepteret af brugeren.

Taleteknologi udgør grundlaget for grænseflader som giver brugere mulighed for at arbejde interaktivt i talt sprog.

Virksomheder anvender ofte svar som er indtalt af professionelle talere, til at generere output til den stemmestyrede grænseflade. Ved faste ytringer hvor ordvalget



5: Talebaseret dialogarkitektur

ikke afhænger af en særlig brugskontekst eller af personlige brugerdata, kan dette give en rigtig god brugeroplevelse. Men et mere dynamisk indhold i en ytring vil lide under den unaturlige intonation fordi stumper af lydfiler simpelthen er blevet klippet sammen. Talesyntesesystemer i dag bliver bedre og bedre (skønt de kan blive endnu bedre) til at producere naturligt lydende dynamiske ytringer.

På markedet for taleinteraktionsteknologi er grænsefladerne i løbet af de seneste ti år blevet stærkt standardiserede hvad angår de forskellige teknologiske komponenter. Der er også sket en kraftig markedskonsolidering inden for talegenkendelse og talesyntese. De nationale markeder i G20-landene (økonomisk robuste lande med store befolkninger) er blevet domineret af bare fem globale aktører med Nuance (USA) og Loquendo (Italien) som de største aktører i Europa. I 2011 annoncerede Nuance overtagelsen af Loquendo hvilket udgør endnu et skridt i markedskonsolideringen.

På det danske marked for taleteknologiske løsninger er der et antal nationale firmaer (Mikro Værkstedet, Prolog Development Center, Max Manus samt IBM og Siemens Danmark) der har specialiseret sig i udvikling af løsninger baseret på taleteknologier fra de internationale teknologileverandører. Nuance er den altdominerende internationale taleteknologileverandør med dansksproget taleteknologi. Så godt som samtlige taleteknologier der kan genkende og udtale dansk, er opkøbt af Nuance over de sidste 5-6 år, fx Philips SpeechMagic, Loquendo og SVOX.

Af basisteknologier der kan genkende dansk sprog, findes pt. Nuance SpeechMagic og Nuance Dragon Development Platform (cloud baseret). På disse teknologier har Max Manus udviklet en løsning primært til sygehussektoren, IBM en løsning til den kommunale sektor og Prolog Development Center standardsystemet Dictus samt skræddersyede løsninger til Folketingstidende og de to nationale TV-stationer. Nuance selv leverer gratis App's Dragon Dictation og Dragon Search til iPhone/iPad mens Prolog Development Center leverer Dictus til Android. Talegenkendere til telefoni er reduceret til én basisleverandør, Nuance. Siemens Danmark og Prolog Development Center er de danske firmaer der har leveret talestyret telefontjeneste på Nuance Recognizer v9. Der udbydes over 10 forskellige danske stemmer til talesyntese fra firmaerne Nuance (inkl. Loquendo og SVOX), Acapela og det danske firma Mikro Værkstedet. Desuden planlægger det succesfulde polske firma IVONA at lancere et antal danske stemmer i starten af 2012.

Et stigende antal af de store internationale leverandører af forbrugerelektronik (Garmin, Navigon, Samsung) tilbyder produkter hvor der er indbygget dansk talegenkendelse så produkterne kan styres med stemmen. Et endnu større antal produkter, fx iRobot, har også indbygget dansk talesyntese. Et lille men stigende antal danske producenter af forbrugerprodukter har indbygget dansk taleteknologi, fx 6th Sense Solution til oplæsning af busstoppesteder og CIM Interconn til oplæsning af skærminformation på plejecentre.

Hvad angår anvendelsen af stemmestyrede grænseflader, er efterspørgslen steget voldsomt de seneste fem år. Denne tendens er drevet af kundernes stigende efterspørgsel på selvbetjening og den betydelige udgiftsoptimering som automatisk telefonservice giver, sammen med en væsentligt øget accept af talt sprog som en mulighed for menneske-maskine-interaktion.

I fremtiden vil der komme store forandringer med smartphones som en ny platform til adminstration af kunderelationer foruden fastnettelefoner, internettet og e-mail. Det vil også påvirke brugen af taleinteraktionsteknologien. På langt sigt vil der blive færre telefonbaserede stemmestyrede brugergrænseflader, og talt sprog vil spille en langt mere central rolle som et brugervenligt input til smartphones. Det vil i det store og hele dreje sig om trin for trin-forbedringer af nøjagtigheden

i den taleruafhængige talegenkendelse via dikteringssystemer som allerede nu tilbydes som centrale tjenester for smartphonebrugere.

4.2.4 Maskinoversættelse

Ideen med at anvende digitale computere til oversættelse af natursprog går tilbage til 1946 og blev fulgt op af betydelige forskningsbevillinger i 1950'erne og igen i 1980'erne. **Maskinoversættelse** (MT) kan dog stadig ikke leve op til det indledende løfte om fuldstændig automatiseret oversættelse.

> Den mest basale fremgangsmåde til maskinoversættelse er at erstatte ordene i en tekst på ét sprog med ord fra et andet sprog.

Den mest basale fremgangsmåde til maskinoversættelse er at erstatte ordene i en tekst på ét sprog med ord fra et andet sprog. Denne fremgangsmåde kan stadig anvendes inden for emneområder som har et meget begrænset, formelagtigt sprog, som fx vejrudsigter. En god oversættelse af tekster som er knap så standardiserede, kræver at større tekstenheder (fraser, sætninger og endda hele afsnit) matches med deres nærmeste ækvivalenter på målsproget. Den største vanskelighed ligger i at menneskesprog er flertydigt. Flertydighed skaber udfordringer på flere niveauer, som entydiggørelse af ord på det leksikalske plan (en *jaguar* er et bilmærke eller et dyr) eller afgørelse af hvortil et led hører, som i eksemplet nedenfor hvor *med en kikkert* kan referere til enten *jeg* eller *mand*:

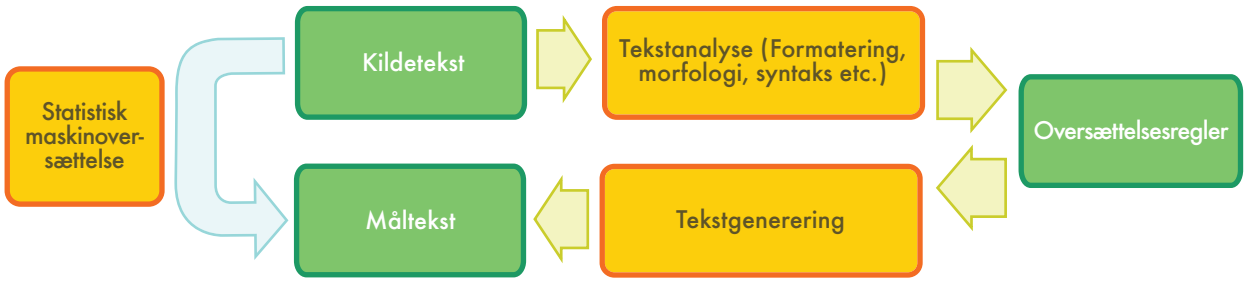
- *Jeg så en mand med en kikkert.*

En måde at bygge MT-systemer på er at anvende lingvistiske regler. Når man oversætter mellem tæt beslægtede sprog, kan direkte udskiftning af ordene være en fornuftig metode. Men regelbaserede maskinoversættelsessystemer analyserer som oftest inputteksten og danner en midlertidig symbolsk repræsentation, hvorfra teksten kan genereres på målsproget. Om denne metode giver succes, afhænger helt af om der er store ordbøger med morfologisk, syntaktisk og semantisk information til disposition samt store mængder af omhyggeligt opbyggede grammatiske regler udarbejdet af dygtige lingvister. Det er en meget lang og derfor meget kostbar proces.

I slutningen af 1980'erne da computerkraften øgedes og blev billigere, var der stigende interesse for statistiske modeller til maskinoversættelse. Statistiske modeller stammer fra analyser af **parallelle korpusser** som fx det parallelle korpus Europarl som indeholder udskrifter fra Europaparlamentets møder på 21 EU-sprog. Hvis der er data nok, er statistisk MT godt nok til at udlede en omtrentlig betydning af en tekst på et fremmedsprog ved at behandle parallelle versioner og finde sandsynlige mønstre af ord. Men i modsætning til regelbaserede systemer genererer statistisk MT ofte et ugrammatisk output. Statistisk MT har den fordel at det kræver færre menneskelige resurser, og det kan også tage højde for særlige karakteristika i et sprog (fx idiomatiske udtryk) som ofte bliver overset i regelbaserede systemer.

Styrkerne og svaghederne ved regelbaseret og statistisk maskinoversættelse har en tilbøjelighed til at være komplementære, så forskere nu til dags fokuserer på hybride fremgangsmåder der kombinerer de to metoder. Én metode anvender både regelbaserede og statistiske systemer sammen med et udvælgelsesmodul som afgør hvad der er det bedste output for hver sætning. Men resultaterne for sætninger som er længere end fx 12 ord, er som regel ikke særligt gode. Det er en bedre løsning at kombinere de bedste dele af hver sætning fra mange forskellige output; dette er en ret kompliceret proces da det ikke altid er indlysende hvilke dele der passer sammen i de forskellige alternative sætninger. De skal først aligneres.

Maskinoversættelse er en særlig udfordring for dansk. Muligheden for at danne nye komposita gør ordbogsa-



6: Maskinoversættelse (statistisk; regelbaseret)

nalyse og -dækning vanskelig; de mange partikelverber udgør et problem for analysen, og udbredt leksikalsk flertydighed vanskeliggør entydiggørelsen af ord.

Bortset fra PaTrans (et engelsk-dansk patentoversættelsessystem) blev de tidlige MT-systemer for dansk, som SYSTRAN-prototypen, primært udviklet af udenlandske virksomheder. Alle disse systemer var regelbaserede. Der foregår en ganske betydelig forskning i MT-teknologi i både nationale og internationale sammenhænge, og der findes nogle nye danske virksomheder (som fx Grammar Soft og Languagelens) der leverer regelbaserede og statistiske maskinoversættelsessystemer for dansk. Størstedelen af de lettilgængelige systemer for dansk, som fx Google Translate og ESTeam Translator, bliver dog stadig udviklet i udlandet.

Maskinoversættelse er en særlig udfordring for dansk.

Der er stadig et kæmpestort potentiale i at forbedre MT-systemernes kvalitet. Udfordringerne indebærer at man tilpasser de sproglige resurser til det aktuelle emne- eller brugerområde og integrerer teknologien i arbejdsgange der allerede omfatter termbaser og oversættelseshukommelser. Et andet problem er at de fleste nuværende systemer er centreret om engelsk og kun støtter nogle få sprog fra og til dansk. Det skaber hindringer for oversættelsesprocessen og tvinger brugerne af MT til at lære forskellige leksikalske kodningsværktøjer for de forskellige systemer.

Evalueringskampagner hjælper med at sammenligne kvaliteten af MT-systemerne, de forskellige fremgangsmåder og systemernes status for forskellige sprogpar. Tabel 7 som blev lavet i EC Euromatrix+-projektet, viser resultaterne for 22 af de 23 officielle EU-sprog parvis. (Irsk blev ikke sammenlignet.) Resultaterne er ordnet efter BLEU-pointtal [21] som giver højere tal for bedre oversættelser. (En menneskelig oversætter ville score omkring 80 point.)

De bedste resultater (med grønt og blåt) blev opnået af sprog som nyder godt af en stor forskningsindsats i koordinerede forskningsprogrammer, og af at der findes mange parallelle korpusser for disse sprog (fx engelsk, fransk, hollandsk, spansk og tysk). Sprog med dårligere resultater er vist med rødt. Disse sprog mangler enten forsknings- og udviklingsindsatser af denne art, eller også er de strukturelt set meget anderledes end andre sprog (fx ungarsk, maltesisk og finsk).

4.3 ANDRE ANVENDELSESOMRÅDER

Opbygningen af sprogteknologiske systemer indebærer en mængde mindre opgaver som ikke altid kan ses på systemets overflade, men som leverer vigtige servicefunktionaliteter "under kølerhjelmen" til det pågældende sy-

	Målsprog – Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

7: Maskinoversættelse mellem 22 EU-sprog – Machine translation between 22 EU-languages [22]

stem. De udgør alle vigtige forskningsemner og har nu udviklet sig til individuelle subdiscipliner inden for datalingvistik.

Spørgsmål/svar-systemer er fx et aktivt forskningsområde hvor der er blevet opbygget opmærkede korpusser og igangsat videnskabelige konkurrencer. Begrebet spørgsmål/svar-systemer går ud over nøgleordsbaserede søgninger (hvor søgemaskinen svarer ved at levere en samling potentielt relevante dokumenter) og sætter brugeren i stand til at stille et konkret spørgsmål som systemet kan give et enkelt svar på. Fx

> *Spørgsmål: Hvor gammel var Neil Armstrong da han trådte ud på månen?*
>
> *Svar: 38.*

Spørgsmål/svar-systemer er tydeligt nok tæt relateret til internetsøgningens kerneområde, men nu til dags er det en paraplyterm for forskningsemner som: hvilke forskellige typer spørgsmål findes der, og hvordan skal de behandles; hvordan kan en samling dokumenter som muligvis indeholder svaret, analyseres og sammenlignes (og indeholder de modstridende svar?); samt hvor specifik og pålidelig information (svaret) kan man uddrage fra et dokument uden at kende dets indhold.

Sprogteknologiske applikationer leverer ofte vigtige servicefunktionaliteter under kølerhjelmen til større softwaresystemer.

Dette hænger igen sammen med videnuddragelse (information extraction, IE), et felt som var utroligt populært og toneangivende da datalingvistik tog en statistisk drejning i starten af 1990'erne. IE har til hensigt at identificere særlige stykker information i særlige typer af dokumenter, som fx at finde nøglepersonerne i virksomhedsovertagelser ud fra hvad der rapporteres i aviserne.

Et andet almindeligt scenarie som er blevet studeret, er rapporter om terrorhændelser. Problemet her er at parre teksten med en skabelon som fastslår gerningsmanden, målet, tidspunktet, stedet og udfaldet af hændelsen. Udfyldning af domænespecifikke skabeloner er det centrale karaktertræk ved informationssøgning hvilket gør den til endnu et eksempel på teknologi "bag scenen", en teknologi som udgør et velafgrænset forskningsområde der til praktisk brug skal indlejres i et passende system.

Tekstresumering og **tekstgenerering** er to grænseområder som enten kan optræde som selvstændige anvendelser eller spille en støttende rolle "under kølerhjelmen". Tekstresumering forsøger at give essensen af en lang tekst i kort form og er en af funktionaliteterne i Microsoft Word. Som oftest anvendes en statistisk fremgangsmåde til at identificere de "vigtige" ord i en tekst (dvs. ord der optræder meget hyppigt i den aktuelle tekst, men mindre hyppigt i almindelig sprogbrug) og afgøre hvilke sætninger der indeholder de "vigtigste" ord. Disse sætninger trækkes ud og sættes sammen for at danne resumeet. I dette helt almindelige kommercielle scenario er resumering ganske enkelt en form for sætningsudtrækning, og teksten reduceres til en delmængde af dens sætninger. En alternativ fremgangsmåde, som der er blevet forsket en del i, er at generere helt nye sætninger der ikke eksisterer i kildeteksten.

Dansk forskning inden for teksteknologi er ikke nær så udviklet som den tilsvarende engelske.

Det kræver en langt dybere forståelse af teksten, hvilket er ensbetydende med at metoden er meget mindre robust. Alt i alt bliver tekstgenerering sjældent brugt som selvstændigt system, men indbygges i større softwaresystemer, fx et klinisk informationssystem der samler, gemmer og behandler patientdata. At lave rapporter er bare en af tekstresumeringens mange anvendelsesmuligheder.

For dansk er alle disse forskningsområder meget mindre udviklede end for engelsk hvor spørgsmål/svarsystemer, informationssøgning og resumering har været genstand for mange åbne konkurrencer siden 1990'erne (fortrinsvis organiseret af DARPA/NIST i USA). Disse har forbedret *state of the art* betydeligt, men det har altid været med fokus på engelsk; nogle konkurrencer har tilføjet flersproglige spor, men dansk har kun deltaget i ganske få af disse. Som følge heraf findes der knap nok opmærkede korpusser eller andre resurser til disse opgaver.

Som det var tilfældet med søgning, som nævnt ovenfor, er der sat flere projekter i gang af SMV'er (såsom RDFined og Ankiro) og visse medier (Information, InfoMedia, DR) inden for områderne navnegenkendelse og videnuddragelse. Nogle af disse trækker på værktøjer udviklet på Center for Sprogteknologi ved Københavns Universitet, såsom ordklassetagger, lemmatiser og en nøgleords-udtrækker, og nogle af projekterne anvender den leksikalske database STO og/eller ordnettet DanNet, som begge er resurser der er blevet udviklet i samarbejde med andre institutioner. Der er online adgang til et resumeringssystem, DanSum på centrets hjemmeside. Hvad angår tekstgenerering har genbrugelige komponenter traditionelt været begrænset til overfladerealiseringen (genereringsgrammatikker); og igen er den software der findes, mest for engelsk.

4.4 SPROGTEKNOLOGI I FORSKNING OG UDDANNELSE

Dansk forskning i sprogteknologi udføres ved flere forskningsinstitutioner og gennem flere samarbejdsprojekter vedrørende infrastruktur og forskning. I det følgende nævnes primært drivkræfterne inden for området, men det skal understreges at relateret forskning også udføres ved andre institutioner og i andre projekter end dem der nævnes her.

Som tidligere nævnt er Center for Sprogteknologi ved Københavns Universitet det nationale center for sprogteknologi. Centrets primære forskningsemner er sprogresurser og værktøjer som fx fagsproglige korpusser og ordnet, flersproglighed (maskinoversættelse, kontrollerede sprog osv.), multimodalitet, informationssøgning, ontologier og sprogteknologi inden for andre anvendelsesområder som fx e-læring.

Institut for Internationale Sprogstudier og Vidensteknologi på Copenhagen Business School forsker i sprogteknologi og datalingvistik, herunder træbanker, statistisk maskinoversættelse, terminologi samt taleteknologi. Instituttets forskning tager med udgangspunkt i sprog, tekst og oversættelsesstudier fat på udfordringerne omkring virksomhedernes professionelle arbejde med sprog i en globaliseret verden og omkring problemstillinger relateret til fagsprog (LSP). På Copenhagen Business School findes også DANTERMcentret for terminologi og terminologiske værktøjer.

På Institut for Sprog og Kommunikation, Syddansk Universitet, arbejdes bl.a. med forskning inden for sprogteknologi og datalingvistik, et eksempel herpå er VISL-projektet (Visual Interactive Syntax Learning). Institut for Elektroniske Systemer på Aalborg Universitet er en drivende kraft inden for taleteknologien. Dansk Sprognævn samt Det Danske Sprog- og Litteraturselskab arbejder også med forskning relateret til sprogteknologien, især vedrørende udvikling af ordbøger, korpusser og korpusbaseret identifikation af nye danske ord.

Der findes flere nyligt afsluttede eller igangværende danske forskningsinfrastrukturprojekter i Danmark som det er relevant at nævne. Et eksempel er DK-CLARIN, hvis formål var at opbygge en dansk forskningsinfrastruktur for humaniora som samlede billeder samt talt og skrevet sprog i et sammenhængende og systematisk digitalt arkiv. Endvidere har LARM til formål at opbygge en national, digital og brugerdreven forskningsinfrastruktur som vil give mulighed for at bevare den danske radiofoniske kulturarv.

Der findes også flere samarbejdsprojekter som fx ESICT, NOMCO og DanTermBank. ESICT forsker i udvikling af metoder og teknologier der vil udmunde i et innovativt it-system som giver borgerne adgang til information om sundhed og sygdom, og som er baseret på informationsteknologi, sprogteknologi og formaliseret medicinsk viden. NOMCO er et nordisk samarbejdsprojekt som handler om multimodal korpusanalyse. DanTermBank er et nystartet projekt om udvikling af det teknologiske grundlag for en national termbank for fagsprog. Desuden indgår danske forskningscentre i flere EU-projekter som er relateret til sprogteknologi, fx LetsMT!, CLARIN, CLARA og META-NORD/META-NET.

I modsætning til forskningsaktiviterne i sprogteknologi, som er på et ret høj niveau i Danmark, kan uddannelsessituationen kun betegnes som kritisk. I øjeblikket er der ikke mulighed for at tage en bachelor- eller kandidateksamen i sprogteknologi i Danmark. Der undervises kun i sprogteknologi som en del af andre bachelor- og kandidatuddannelser. Eksempler på sådanne uddannelser er *It and Cognition* som er en kandidatuddannelse på Københavns Universitet og som indbefatter et obligatorisk kursus i sprogteknologi. Et andet eksempel er Syddansk Universitets bachelor i *Kommunikation og it* som har et valgfrit forløb om sprogteknologi. Manglen på uddannelser i grundlæggende sprogteknologi er problematisk og udgør en forholdsvis ny situation i landet eftersom der indtil for nylig blev tilbudt undervisning i datalingvistik på både Copenhagen Business School og på Københavns Universitet. Situationen bør give anledning til politisk indgriben eftersom fremtidens forskning og udvikling i dansk sprogteknologi kan være truet. Man kan undre sig over at manglen på uddannelse på området sker samtidig med en stigende interesse for og efterspørgsel efter sprogteknologi i dansk erhvervsliv.

4.5 NATIONALE PROGRAMMER OG TILTAG

De danske forskningsråd har støttet flere projekter om sprogteknologi gennem årene og har tidligere også oprettet specielle programmer rettet mod forskningsområdet.

Det europæiske EUROTRA-program udgjorde nok den første massive støtte til sprogteknologien i Danmark. EUROTRA sigtede mod opbygning af flersproglig maskinoversættelse for alle de officielle EU-sprog. Projektet startede sidst i 1970'erne og fortsatte til starten af 1990'erne. Formanden for EUROTRAs samarbejdsudvalg 1986-1992 var en dansker, hvilket betød at danske interessenter rettede en del opmærksomhed mod projektet. EUROTRAs slutprodukt var ikke et færdigt maskinoversættelsessystem, men projektet leverede en prototype og dermed det nødvendige grundlag for at kunne udvikle et kommercielt engelsk-dansk oversættelsessystem til patenttekster; et system som blev brugt i over 10 år.

Et af de tidlige programmer som blev lanceret af Statens Humanistiske Forskningsråd var ETTO (Edb for Tekst, Tale og Ordbøger), 1982-85. Det næste program var Eksperimentel Sprogvidenskab som startede i 1991. Statens Humanistiske Forskningsråd har ikke haft specielle programmer for sprogteknologi siden da. Forskningsrådet for de Tekniske Videnskaber havde taleteknologi som et af de primære temaer i strategiplanen for 1998-2002, og de støttede flere sprogteknologiprojekter.

De tværfaglige it-forskningsprogrammer som løb fra midten af 1990'erne til 2003 har kunnet støtte nogle få projekter relateret til sprogteknologi (som fx OntoQuery og SIABO), men programmerne havde ikke en specifik målsætning om støtte til sprogteknologi.

Forskningsstyrelsen under Ministeriet for Videnskab, Teknologi og Udvikling besluttede i 2001 at fremme sprogteknologien ved at give midler til et samarbejdsprojekt. Dette projekt handlede om at udvide det danske sprogteknologiske leksikon som oprindeligt blev udviklet i PAROLE-projektet. Det nye leksikon kom til at hedde den SprogTeknologiske Ordbase (STO), og den kan nu erhverves gennem ELRA og anvendes til forskning og kommercielle formål. Det var forskningsstyrelsens klare mål at understøtte det danske sprog ved at tildele midler til udvikling af sprogteknologiens grundlæggende byggesten.

Siden tildeling af disse midler har der så vidt vi ved, ikke været andre programmer specielt til understøttelse af sprogteknologien, men som altid kan sprogteknologiprojekter støttes af forskningsrådene og andre i konkurrence med andre emner.

Parallelt med ESFRI-initiativet vedrørende forskningsinfrastrukturer, har Danmark påbegyndt udviklingen af et dansk roadmap for forskningsinfrastrukturer. En af de forskningsinfrastrukturer der vil blive finansieret er Digital Humaniora Laboratorium (DigHumLab) til understøttelse af forskning i humaniora gennem brug af bl.a. sprogteknologi.

Som nævnt ovenfor har tidligere programmer givet anledning til et antal sprogteknologiske værktøjer og resurser for det danske sprog. I det følgende opsummeres den aktuelle situation vedrørende dansk sprogteknologi.

4.6 VÆRKTØJER OG RESURSER

Tabel 8 nedenfor sammenfatter fakta vedrørende sprogteknologiens niveau for dansk. Vurderingen af eksisterende værktøjer og resurser er baseret på eksperters kvalificerede skøn, og pointskalaen spænder fra 0 (meget lavt) til 6 (meget højt). Resultaterne for det danske sprog kan opsummeres som følger:

- Inden for taleteknologi findes der en del værktøjer, som dog næsten udelukkende er tilgængelige på kommerciel basis. Hvad angår kvaliteten af disse, er uenigheden stor blandt kommercielle udviklere og forskere. Hvor kommercielle udviklere fx fremfører

	Kvantitet	Tilgængelighed	Kvalitet	Dækningsgrad	Modenhed	Bæredygtighed	Tilpasningsevne
Sprogteknologi: værtøjer, teknologier og applikationer							
Talegenkendelse	4	2	–*	4	4	3	3
Talesyntese	5	2	–*	3	3	2	3
Grammatisk analyse	3	2	4	4	3	2	3
Semantisk analyse	1	0	1	1	1	1	1
Tekstgenerering	0	0	0	0	0	0	0
Maskinoversættelse	3	2	2	3	3	1	2
Sprogresurser: resurcer, data og videnbaser							
Tekstkorpusser	4	3	4	3	4	4	3
Talesprogskorpusser	3	3	3	1	2	2	2
Parallelle korpusser	3	1	3	2	2	2	2
Leksikalske resurcer	3	3	4	4	3	3	3
Grammatikker	2	1	4	1	2	2	2

8: Sprogteknologiens niveau for dansk (*der er ikke afsat værdier for kvaliteten; se forklaring i teksten)

at genkendelsesprocenten er mindst 95%, pointerer forskere at denne høje genkendelsesprocent kun gør sig gældende hvis man taler meget tydeligt, og hvis systemet kender brugerens stemme i forvejen. På grund af disse divergerende opfattelser af taleteknologiens kvalitet er der ikke afsat værdier i dette felt i figur 8.

- For tekstanalyseværktøjer der relaterer sig til morfologi, såsom tokenisers, ordklassetaggere og morfologiske analyseværktøjer, er situationen i Danmark forholdsvis god. Til sammenligning findes der kun ganske få parsere for dansk med en rimelig dækningsgrad selvom en del arbejde er gjort inden for dependensparsing og constraint grammar.
- Tekstforståelse og semantik er vanskeligere at håndtere end syntaks, og tekstsemantik er vanskeligere at håndtere end ord- og sætningssemantik. Tekstforståelse opnår således en meget lav score for dansk.
- Der forskes i maskinoversættelse ved flere danske forskningsinstitutioner, især i statistisk maskinoversættelse. Kvaliteten er dog stadig dårlig, og antallet af sprogpar er begrænset.
- I de senere år er der opbygget ganske betydelige resurser, og det betyder at situationen for dansk er temmelig gunstig hvad angår referencekorpusser, leksika, wordnets og terminologisamlinger. Problemstillinger omkring IPR betyder dog at korpusserne ikke nødvendigvis er gratis. Der findes referencekorpusser af høj kvalitet, mens dette ikke er tilfældet for korpusser opmærket med syntaks og semantik. Vedrørende multimodale korpusser (fx videoer) er en del tiltag i gang, men korpusserne er endnu ikke

alment tilgængelige. Parallelle korpusser er også stadig en mangelvare.

- For flere resursetyper findes ingen standardisering så selvom de eksisterer nu, er det ikke sikkert de findes så længe.

På en række specifikke områder af dansk sprogforskning har vi software med begrænset funktionalitet til rådighed i dag. Men der er behov for yderligere forskning for at udbedre de nuværende mangler inden for tekstanalyse på et dybere semantisk niveau og for at videreudvikle ressourcer som parallelle korpusser for maskinoversættelse. Det er således påkrævet at udarbejde fælles programmer og initiativer for at standardisere data og udvekslingsformater.

4.7 SAMMENLIGNING PÅ TVÆRS AF SPROG

Det sprogteknologiske niveau varierer meget fra land til land. For at sammenligne situationen mellem sprogene fokuseres der i dette afsnit på to eksempelområder (maskinoversættelse og taleteknologi) og en underliggende teknologi (tekstanalyse), så vel som på de basisresurser der er nødvendige for at bygge sprogteknologiske anvendelser. Sprog er blevet kategoriseret ved hjælp af følgende fem-trins-skala:

1. Fuldstændig støtte
2. God støtte
3. Medium støtte
4. Fragmentarisk støtte
5. Ringe eller ingen støtte

Støtten til sprogteknologi blev målt i henhold til følgende kriterier:

Taleteknologi: Kvaliteten af eksisterende talegenkendelse, kvaliteten af eksisterende talesyntese, domænedækning, antal og størrelse af eksisterende talekorpusser, mængden af og variationen i tilgængelige talebaserede systemer.

Maskinoversættelse: Kvaliteten af eksisterende maskinoversættelsesteknologi, antal dækkede sprogpar, dækningsgraden af lingvistiske fænomener og domæner, kvaliteten af og størrelsen på eksisterende parallelle korpusser, mængden af og variationen i tilgængelige MT-systemer.

Tekstanalyse: Kvaliteten og dækningsgraden af eksisterende tekstanalyseteknologi (morfologi, syntaks, semantik), dækningsgraden af lingvistiske fænomener og domæner, mængden af og variationen i tilgængelige systemer, kvaliteten og størrelsen af eksisterende (annoterede) tekstkorpusser, kvaliteten og dækningsgraden af eksisterende leksikalske resurser (fx WordNet) og grammatikker.

Resurser: Kvaliteten og størrelsen af eksisterende tekstkorpusser, talekorpusser og parallelle korpusser, kvaliteten og dækningsgraden af leksikalske resurser og grammatikker.

Tabellerne 9 – 12 viser at Danmark vurderes til at være på omtrent samme niveau som dets skandinaviske naboer (Sverige, Norge og Finland). Dog ligger Danmark i den næstdårligste kategori for taleteknologiområdet (skønt der som nævnt ovenfor er nogen uenighed blandt forskere og kommercielle udviklere om kvaliteten på dette område). Ligeledes ligger vi i den næstdårligste kategori for både tekstanalyse og resurser. For maskinoversættelse er tallene endnu lavere. Her er Danmark i den dårligste kategori sammen med mange andre lande.

For at kunne udvikle mere sofistikerede systemer som fx maskinoversættelse, er der derfor et klart behov for resurser og teknologier som dækker en bred vifte af lingvistiske aspekter og muliggør en dyb semantisk analyse af inputteksten. Ved at forbedre kvaliteten og dækningen af de grundlæggende resurser og teknologier vil der åbne sig nye chancer for en række avancerede anvendelsesområder inklusiv maskinoversættelse af høj kvalitet.

4.8 KONKLUSIONER

Sprogteknologirapporterne indeholder en vurdering af det sprogteknologiske niveau for 30 EU-sprog. Rapporterne giver os dermed mulighed for at sammenligne situationen på tværs af sprogene og identificere væsentlige mangler og behov. Dermed vil det sprogteknologiske miljø og dets interessenter være i stand til på et langt bedre grundlag at udstikke retningslinjerne for fremtidige, storstilede forsknings- og udviklingsprogrammer med det formål at fremme et teknologistøttet og flersprogligt Europa.

Vi har set at der er store forskelle mellem de europæiske sprog. Mens nogle sprog ligger på et rimelig højt niveau inden for visse anvendelsesområder, har andre sprog (som oftest de "mindre" sprog) store huller. Mange sprog mangler basisteknologier for tekstanalyse og ikke mindst de basisresurser der skal anvendes til at forbedre teknologierne. Andre sprog har basisresurserne klar, men har endnu ikke investeret i semantisk analyse. Der er derfor stadig brug for en storstilet indsats hvis vi fx skal sikre maskinoversættelse af høj kvalitet mellem alle de europæiske sprog.

Resultaterne for dansk viser at situationen kun er rimelig god med hensyn til de mest basale værktøjer og resurser. Vi har set at der findes værktøjer til informationssøgning, maskinoversættelse, talegenkendelse og -syntese samt parallelle korpusser og talekorpusser i en vis udstrækning. Disse værktøjer og resurser er dog temmelig simple og har en begrænset funktionalitet. Parallelle korpusser findes fx kun for ganske få sprogpar hvor dansk udgør det ene sprog.

Når det drejer sig om mere avancerede områder som semantisk analyse, intelligent informationssøgning og sproggenerering, mangler dansk helt klart basale værktøjer og resurser selvom nogle i øjeblikket er under udvikling. De mest avancerede værktøjer til fx diskursbehandling og dialogstyring har meget en begrænset funktionalitet, og semantik- og diskurskorpusser findes kun i særdeles begrænset udstrækning.

En anden betragtning, som ikke helt afspejles i tabellen ovenfor, er at sprogteknologiske resurser og værktøjer til håndtering af kun det danske sprog ikke nødvendigvis fremmer globaliseringen og det internationale samarbejde. "Tvær- og flersproglige" teknologier som forbinder dansk med andre sprog (fx ordnet, parallelle korpusser, maskinoversættelse, flersproglig søgning osv.) er en forudsætning for avanceret teknologiunderstøttet interaktion med vore omgivelser.

Set i lyset af at sprogteknologien spiller en afgørende rolle i bestræbelserne på at beskytte og fremme det danske sprog, peger disse resultater på et stort behov for nye programmer som især fokuserer på udvikling af sprogteknologiske værktøjer og resurser. I modsætning til flere andre nordiske lande, som har fx ingangsat sprogbankprojekter der omfatter resurser og værktøjer (en BLARK), har sprogteknologisk forskning og udvikling i Danmark i den senere tid været i fuld konkurrence med andre temaer. Så selvom dansk sprogteknologi har oplevet fremgang i de senere år, og selvom dansk erhvervsliv gør fremskridt inden for området, er der en overhængende fare for at det ikke går hurtigt nok. Kun en politisk beslutning vil for alvor kunne sætte skub i udviklingen.

Fuldstændig støtte	God støtte	Medium støtte	Fragmentarisk støtte	Ringe/ingen støtte
	engelsk	finsk fransk hollandsk italiensk portugisisk spansk tjekkisk tysk	baskisk bulgarsk **dansk** estisk galicisk græsk irsk katalansk norsk polsk serbisk slovakisk slovensk svensk ungarsk	islandsk kroatisk lettisk litauisk maltesisk rumænsk

9: Taleteknologi: den sprogteknologiske støttes tilstand for 30 europæiske sprog

Fuldstændig støtte	God støtte	Medium støtte	Fragmentarisk støtte	Ringe/ingen støtte
	engelsk	fransk spansk	hollandsk italiensk katalansk polsk rumænsk tysk ungarsk	baskisk bulgarsk **dansk** estisk finsk galicisk græsk irsk islandsk kroatisk lettisk litauisk maltesisk norsk portugisisk serbisk slovakisk slovensk svensk tjekkisk

10: Maskinoversættelse: den sprogteknologiske støttes tilstand for 30 europæiske sprog

Fuldstændig støtte	God støtte	Medium støtte	Fragmentarisk støtte	Ringe/ingen støtte
	engelsk	fransk hollandsk italiensk spansk tysk	baskisk bulgarsk **dansk** finsk galicisk græsk katalansk norsk polsk portugisisk rumænsk slovakisk slovensk svensk tjekkisk ungarsk	estisk irsk islandsk kroatisk lettisk litauisk maltesisk serbisk

11: Tekstanalyse: den sprogteknologiske støttes tilstand for 30 europæiske sprog

Fuldstændig støtte	God støtte	Medium støtte	Fragmentarisk støtte	Ringe/ingen støtte
	engelsk	fransk hollandsk italiensk polsk spansk svensk tjekkisk tysk ungarsk	baskisk bulgarsk **dansk** estisk finsk galicisk græsk katalansk kroatisk norsk portugisisk rumænsk serbisk slovakisk slovensk	irsk islandsk lettisk litauisk maltesisk

12: Sprog- og tekstresurser: den sprogteknologiske støttes tilstand for 30 europæiske sprog

OM META-NET

META-NET er et *Network of Excellence* delvist finansieret af EU-Kommissionen. Netværket består i øjeblikket af 54 medlemmer fra 33 EU-lande. META-NET støtter *Multilingual Europe Technology Alliance* (**META**), et voksende EU-fællesskab af fagfolk og organisationer som arbejder med sprogteknologi. META-NET understøtter det teknologiske grundlag for et flersprogligt informationssamfund som:

- muliggør kommunikation og samarbejde på tværs af sprogene;
- giver lige adgang til information og viden på ethvert sprog;
- bygger på og fremmer funktionaliteter i netværksbaseret informationsteknologi.

Netværket støtter visionen om et enkelt digitalt marked og informationsrum i Europa. META-NET stimulerer og fremmer flersproglige teknologier for alle EU-sprogene. Teknologierne giver mulighed for automatisk oversættelse, generering af indhold, informationsbehandling og vidensstyring inden for en bred vifte af anvendelser og emneområder. De gør det også muligt at lave intuitive sprogbaserede grænseflader til forskellig slags teknologi, lige fra husholdningsmaskiner og biler til computere og robotter.

META-NET startede 1. februar 2010 og har allerede gennemført flere aktiviteter inden for dets tre akser META-VISION, META-SHARE og META-RESEARCH.

META-VISION understøtter en dynamisk og inflydelsesrig interessentgruppe som er samlet om en fælles vision og en strategisk forskningsdagsorden. Denne aktivitets hovedmål er at opbygge et sammenhængende europæisk fællesskab bestående af repræsentanter fra yderst fragmenterede og forskelligartede interessentgrupper inden for sprogteknologien. Denne hvidbogsserie dækker 30 forskellige sprog. Den fælles teknologiske vision blev udviklet i tre visionsgrupper med hver sit emneområde. META Teknologirådet blev etableret for at diskutere og udarbejde en strategisk forskningsdagsorden baseret på denne vision i tæt samarbejde med forskere, udviklere og brugere af sprogteknologi.

META-SHARE står for en åben distribueret infrastruktur til udveksling og deling af resurser. Peer-to-peer netværket vil indeholde sprogdata, værktøjer og webtjenester som er dokumenteret med metadata og organiseret i standardiserede kategorier. Det er nemt at tilgå resurserne, og man kan anvende samme søgeteknik i dem alle. Resursesamlingen inkluderer gratis open-source materialer såvel som kommercielt tilgængelige resurser og værktøjer som kræver betaling.

META-RESEARCH bygger bro til relaterede teknologiske områder. Denne aktivitet forsøger at udnytte fremskridt inden for andre områder og at få det bedste ud af den innovative forskning som kan gavne sprogteknologien. Fokus for aktiviteten er især at lede førende forskning inden for maskinoversættelse, indsamle data, klargøre sæt af data og sørge for sprogresurser til evaluering. Aktiviteten inkluderer også udarbejdelse af oversigter over værktøjer og metoder samt organisering af workshopper og kurser for medlemmerne.

office@meta-net.eu – http://www.meta-net.eu

1

EXECUTIVE SUMMARY

Information technology changes our everyday lives. We typically use computers for writing, editing, calculating, and information searching, and increasingly for reading, listening to music, viewing photos and watching movies. We carry small computers in our pockets and use them to make phone calls, write emails, get information and entertain ourselves, wherever we are. How does this massive digitisation of information, knowledge and everyday communication affect our language? Will our language change or even disappear?

All our computers are linked together into an increasingly dense and powerful global network. The girl in Ipanema, the customs officer in Padborg and the engineer in Kathmandu can all chat with their friends on Facebook, but they are unlikely ever to meet one another in online communities and forums. If they are worried about how to treat earache, they will all check Wikipedia to find out all about it, but even then they won't read the same article. When Europe's netizens discuss the effects of the Fukushima nuclear accident on European energy policy in forums and chat rooms, they do so in cleanly-separated language communities. What the internet connects is still divided by the languages of its users. Will it always be like this?

Many of the world's 6,000 languages will not survive in a globalised digital information society. It is estimated that at least 2,000 languages are doomed to extinction in the decades ahead. Others will continue to play a role in families and neighbourhoods, but not in the wider business and academic world. What are the Danish language's chances of survival?

With approximately 5 million native speakers, Danish must be considered a relatively small language at least when compared to several of the other EU languages. Similar to other small industrialised countries, people's daily lives are greatly influenced by the English language: English movies and TV series are usually not dubbed, but shown with subtitles; big international companies increasingly use English as a "corporate language"; English is also becoming the *lingua franca* in higher education, similar to science and technology where it is playing this role for a long time.

There are plenty of complaints about the ever-increasing use of Anglicisms, and some even fear that the Danish language is becoming riddled with English words and expressions. But the only way to maintain Danish words and phrases is to actually use them – frequently and consciously; linguistic polemics about foreign influences and government regulations do not usually help. Our main concern should not be the gradual Anglicisation of our language, but its complete disappearance from major areas of our personal lives. Not science, aviation and the global financial markets, which actually need a world-wide *lingua franca*. We mean the many areas of life in which it is far more important to be close to a country's citizens than to international partners – domestic policies, for example, administrative procedures, the law, culture and shopping.

The status of a language depends not only on the number of speakers or books, films and TV stations that use it, but also on the presence of the language in the digital information space and software applications. Here

the Danish language is fairly well-placed: many international software products are available in Danish versions; the Danish Wikipedia is growing, and with more than 1 million internet domains registered in 2011, Danish is well represented on the Web relative to its population.

In the field of language technology, however, the Danish language is not sufficiently equipped with products, technologies and resources for meeting future demands. There are applications and tools for speech synthesis, speech recognition, spelling correction, and grammar checking, but substantial improvements are required to ensure proper functionality in all relevant contexts. There are also some applications for automatically translating language, even though these often fail to produce linguistically and idiomatically correct translations, some of which can be explained by the lack of training material in terms of parallel corpora which include Danish. More advanced applications like text understanding, language generation, and dialogue management, are still in very early prototype stage, requiring typically semantically rich resources at a larger scale which are not available for Danish today.

Information and communication technology are now preparing for the next revolution. After personal computers, networks, miniaturisation, multimedia, mobile devices and cloud-computing, the next generation of technology will feature software that understands not just spoken or written letters and sounds but entire words and sentences, and supports users far better because it speaks, knows and understands their language. Forerunners of such developments are the free online service Google Translate that translates between 57 languages, IBM's supercomputer Watson that was able to defeat the US-champion in the game of "Jeopardy", and Apple's mobile assistant Siri for the iPhone that can react to voice commands and answer questions in English, German, French and Japanese.

The next generation of information technology will master human language to such an extent that human users will be able to communicate using the technology in their own language. Devices will be able to automatically find the most important news and information from the world's digital knowledge store in reaction to easy-to-use voice commands. Language-enabled technology will be able to translate automatically or assist interpreters; summarise conversations and documents; and support users in learning scenarios. For example, it will help immigrants to learn the Danish language and integrate more fully into the country's culture.

The next generation of information and communication technologies will enable industrial and service robots (currently under development in research laboratories) to faithfully understand what their users want them to do and then proudly report on their achievements. This level of performance means going way beyond simple character sets and lexicons, spell checkers and pronunciation rules. The technology must move on from simplistic approaches and start modelling language in an all-encompassing way, taking syntax as well as semantics into account to understand the drift of questions and generate rich and relevant answers.

However, there is a yawning technological gap between English and Danish, and it is currently getting wider. Every international technology competition tends to show that results for the automatic analysis of English are far better than those for less-resourced languages such as Danish, even though (or precisely because) the methods of analysis are similar, if not identical. This holds true for extracting information from texts, grammar checking, machine translation and a whole range of other applications. Many researchers reckon that these setbacks are due to the fact that, for fifty years now, the methods and algorithms of computational linguistics and language technology application research have first and foremost focused on English. However, other researchers believe

that English is inherently better suited to computer processing. In any case, there is no doubt of the fact that we need a dedicated, consistent, and sustainable research effort if we want to be able to use the next generation of information and communication technology in those areas of our private and work life where we live, speak and write Danish.

After a relatively successful research record with several national and Nordic initiatives in the area of language technology in the period from 1985-2001, Danish is currently beginning to lack behind, also in the Nordic landscape. During the last decade, no substantial funding has been given to drive Danish language technology forward and the educational situation in the field is equally critical. As the present report will show, we cannot afford this stagnation. Denmark is ranked low on the European list when it comes to availability and development of language technology and there is an indispensable need for invigorating programs focusing on research and resource and technology development in the field. Otherwise we will fail to keep up when a new generation of technologies really starts to master human languages effectively. Through improvements in machine translation, language technology will help in overcoming language barriers, but it will only be able to operate between those languages that have managed to survive in the digital world. If there is adequate language technology available, then it will be able to ensure the survival of languages with very small populations of speakers. If not, even 'larger' languages will come under severe pressure.

2

LANGUAGES AT RISK: A CHALLENGE FOR LANGUAGE TECHNOLOGY

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in information and communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

The digital revolution is comparable to Gutenberg's invention of the printing press.

After Gutenberg's invention, real breakthroughs in communication were accomplished by efforts such as Luther's translation of the Bible into vernacular language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled exchanges across languages;
- the creation of editorial and bibliographic guidelines assured the quality of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology has helped to automate and facilitate many processes:

- desktop publishing software has replaced typewriting and typesetting;
- Microsoft PowerPoint has replaced overhead projector transparencies;
- e-mail allows documents to be sent and received more quickly than using a fax machine;
- Skype offers cheap Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- web search engines provide keyword-based access;
- online services like Google Translate produce quick, approximate translations;
- social media platforms such as Facebook, Twitter and Google+ facilitate communication, collaboration, and information sharing.

Although these tools and applications are helpful, they are not yet capable of supporting a fully-sustainable, multilingual European society in which information and goods can flow freely.

2.1 LANGUAGE BORDERS HOLD BACK THE EUROPEAN INFORMATION SOCIETY

We cannot predict exactly what the future information society will look like. However, there is a strong likelihood that the revolution in communication technology is bringing together people who speak different languages in new ways. This is putting pressure both on individuals to learn new languages and especially on developers to create new technology applications to ensure mutual understanding and access to shareable knowledge. In the global economic and information space, there is increasing interaction between different languages, speakers and content thanks to new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter, YouTube, and, recently, Google+) is only the tip of the iceberg.

> The global economy and information space confronts us with different languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognise that it is in a language that we do not understand. According to a recent report from the European Commission, 57% of Internet users in Europe purchase goods and services in non-native languages; English is the most common foreign language followed by French, German and Spanish. 55% of users read content in a foreign language while 35% use another language to write e-mails or post comments on the Web [2]. A few years ago, English might have been the lingua franca of the Web – the vast majority of content on the Web was in English – but the situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has exploded.

Surprisingly, this ubiquitous digital linguistic divide has not gained much public attention; yet, it raises a very pressing question: Which European languages will thrive in the networked information and knowledge society, and which are doomed to disappear?

2.2 OUR LANGUAGES AT RISK

While the printing press helped step up the exchange of information in Europe, it also led to the extinction of many European languages. Regional and minority languages were rarely printed and languages such as Cornish and Dalmatian were limited to oral forms of transmission, which in turn restricted their scope of use. Will the Internet have the same impact on our modern languages?

Europe's approximately 80 languages are one of our richest and most important cultural assets, and a vital part of this unique social model [3]. While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many European languages could become irrelevant in a networked society. This would weaken Europe's global standing, and run counter to the strategic goal of ensuring equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society [4].

> The wide variety of languages in Europe is one of its richest and most important cultural assets.

2.3 LANGUAGE TECHNOLOGY IS A KEY ENABLING TECHNOLOGY

In the past, investments in language preservation focused primarily on language education and translation. According to one estimate, the European market for translation, interpretation, software localisation and website globalisation was €8.4 billion in 2008 and is expected to grow by 10% per annum [5]. Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use appropriate technology, just as we use technology to solve our transport and energy needs among others.

Language technology targeting all forms of written text and spoken discourse can help people to collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills. It often operates invisibly inside complex software systems to help us already today to:

- find information with a search engine;
- check spelling and grammar in a word processor;
- view product recommendations in an online shop;
- follow the spoken directions of a navigation system;
- translate web pages via an online service.

Language technology consists of a number of core applications that enable processes within a larger application framework. The purpose of the META-NET language white papers is to focus on how ready these core enabling technologies are for each European language.

> Europe needs robust and affordable language technology for all European languages.

To maintain our position in the frontline of global innovation, Europe will need language technology, tailored to all European languages, that is robust and affordable and can be tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

2.4 OPPORTUNITIES FOR LANGUAGE TECHNOLOGY

In the world of print, the technology breakthrough was the rapid duplication of an image of a text using a suitably powered printing press. Human beings had to do the hard work of looking up, assessing, translating, and summarising knowledge. We had to wait until Edison to record spoken language – and again his technology simply made analogue copies.

Language technology can now simplify and automate the processes of translation, content production, and knowledge management for all European languages. It can also empower intuitive speech-based interfaces for household electronics, machinery, vehicles, computers and robots. Real-world commercial and industrial applications are still in the early stages of development, yet R&D achievements are creating a genuine window of opportunity. For example, machine translation is already reasonably accurate in specific domains, and experimental applications provide multilingual information and knowledge management, as well as content production, in many European languages.

As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for specialised domains, and often exhibit limited performance. However, there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, edutainment packages, libraries, simula-

tion environments and training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just some of the application areas in which language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology helps overcome the "disability" of linguistic diversity.

Language technology represents a tremendous opportunity for the European Union. It can help to address the complex issue of multilingualism in Europe – the fact that different languages coexist naturally in European businesses, organisations and schools. However, citizens need to communicate across the language borders of the European Common Market, and language technology can help overcome this final barrier, while supporting the free and open use of individual languages. Looking even further ahead, innovative European multilingual language technology will provide a benchmark for our global partners when they begin to support their own multilingual communities. Language technology can be seen as a form of "assistive" technology that helps overcome the "disability" of linguistic diversity and makes language communities more accessible to each other. Finally, one active field of research is the use of language technology for rescue operations in disaster areas, where performance can be a matter of life and death: Future intelligent robots with cross-lingual language capabilities have the potential to save lives.

2.5 CHALLENGES FACING LANGUAGE TECHNOLOGY

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. Widely-used technologies such as the spelling and grammar correctors in word processors are typically monolingual, and are only available for a handful of languages. Online machine translation services, although useful for quickly generating a reasonable approximation of a document's contents, are fraught with difficulties when highly accurate and complete translations are required. Due to the complexity of human language, modelling our tongues in software and testing them in the real world is a long, costly business that requires sustained funding commitments. Europe must therefore maintain its pioneering role in facing the technological challenges of a multiple-language community by inventing new methods to accelerate development right across the map. These could include both computational advances and techniques such as crowdsourcing.

Technological progress needs to be accelerated.

2.6 LANGUAGE ACQUISITION IN HUMANS AND MACHINES

To illustrate how computers handle language and why it is difficult to program them to process different tongues, let's look briefly at the way humans acquire first and second languages, and then see how language technology systems work.

Humans acquire language skills in two different ways. Babies acquire a language by listening to the real interactions between their parents, siblings and other family members. From the age of about two, children produce

their first words and short phrases. This is only possible because humans have a genetic disposition to imitate and then rationalise what they hear.

Learning a second language at an older age requires more cognitive effort, largely because the child is not immersed in a language community of native speakers. At school, foreign languages are usually acquired by learning grammatical structure, vocabulary and spelling using drills that describe linguistic knowledge in terms of abstract rules, tables and examples.

Humans acquire language skills in two different ways: learning from examples and learning the underlying language rules.

Moving now to language technology, the two main types of systems acquire language capabilities in a similar manner. Statistical (or "data-driven") approaches obtain linguistic knowledge from vast collections of concrete example texts. While it is sufficient to use text in a single language for training, e. g., a spell checker, parallel texts in two (or more) languages have to be available for training a machine translation system. The machine learning algorithm then "learns" patterns of how words, short phrases and complete sentences are translated.

This statistical approach usually requires millions of sentences to boost performance quality. This is one reason why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, and services such as Google Search and Google Translate, all rely on statistical approaches. The great advantage of statistics is that the machine learns quickly in a continuous series of training cycles, even though quality can vary randomly.

The second approach to language technology, and to machine translation in particular, is to build rule-based systems. Experts in the fields of linguistics, computational linguistics and computer science first have to encode grammatical analyses (translation rules) and compile vocabulary lists (lexicons). This is very time consuming and labour intensive. Some of the leading rule-based machine translation systems have been under constant development for more than 20 years. The great advantage of rule-based systems is that the experts have more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. However, due to the high cost of this work, rule-based language technology has so far only been developed for a few major languages.

As the strengths and weaknesses of statistical and rule-based systems tend to be complementary, current research focusses on hybrid approaches that combine the two methodologies. However, these approaches have so far been less successful in industrial applications than in the research lab.

As we have seen in this chapter, many applications widely used in today's information society rely heavily on language technology, particularly in Europe's economic and information space. Although this technology has made considerable progress in the last few years, there is still huge potential to improve the quality of language technology systems. In the next section, we describe the role of Danish in European information society and assess the current state of language technology for the Danish language.

3

THE DANISH LANGUAGE IN THE EUROPEAN INFORMATION SOCIETY

3.1 GENERAL FACTS

Danish is the official language of Denmark, which has approx. 5,500,000 inhabitants. 90% of these are ethnic Danes, with Danish as their mother tongue. For the last 10%, only one minority language, German, is established officially. Thus, the cities Sønderborg, Åbenrå, Tønder and Haderslev officially grant minority rights to their citizens. The overall number of native speakers of German amounts to approx. 20,000 in South Jutland alone (cf., e. g., [6]). Apart from the Danish speakers who live in Denmark, Danish is also the native or cultural language of around 50,000 Germano-Danish citizens living in the south of Schleswig. Furthermore, the Danes who emigrated to America and Australia preserve, to a certain extent, their native language.

A law of March 2006 fixes the conditions for the linguistic integration of migrants. Migrants in possession of a residence permit and national identity number can access three years of Danish language training. Learning Danish is not compulsory. However, if you want to become a permanent resident or obtain Danish citizenship it is necessary to pass a Danish language test.

In the Faroe Islands and Greenland, the law of autonomy guarantees official equality of Danish alongside the Faeroese and Greenlandic languages. Danish is an obligatory subject in schools. In Iceland, Danish has been a part of the school curriculum since the end of the 1990s and Danish is still used to facilitate communication with other Nordic countries.

Denmark has ratified the Nordic Language Convention (1987) which secures the rights of Nordic citizens to use their own language, in contact with authorities in all Nordic countries. Denmark has also ratified the Nordic Language Declaration (2006), which is a joint policy document of the Nordic Council of Ministers. It states that both national and minority languages should be supported and protected, that universities should use a parallel language strategy ensuring the use of English alongside the use of the national languages, and that the citizens of the Nordic countries should be given the opportunity to learn their mother tongue, as well as at least two foreign languages. Regarding language technology, the declaration emphasises the need for MT-systems, information retrieval systems and advanced terminology databases for the Nordic languages.

3.2 PARTICULARITIES OF THE DANISH LANGUAGE

Danish derives from the East Norse dialect group. A more recent classification based on mutual intelligibility separates modern spoken Danish, Norwegian and Swedish from the other Nordic languages into the *Mainland Scandinavian* group.

Danish exhibits a number of specific characteristics that contribute to the richness of the language but can also be a challenge for the computational processing of natural language (cf., e. g., [7, 8, 9]).

Certain linguistic characteristics of Danish are challenges for computational processing.

For speech technology, the following characteristics can be mentioned as relevant:

- a very large number of vowels (29) in spoken Danish [10];
- use of unit stress to indicate process reading of the verb: *læse a'vis* [lit. read 'newspaper], *spejle 'æg* [lit. fry 'egg], *spille 'skak* [play 'chess];
- glottal stop as meaning differentiating feature: *stien* (glottal stop) vs. *stigen* (no glottal stop) [the path vs. the ladder].

Furthermore, the Danish vocabulary exposes:

- a large flexibility regarding dynamic generation of compounds: such as *skiinstruktørsammenslutningssekretærsaspirant* [lit. ski-instructor-association-secretary-aspirant];
- an extensive use of particles with semi-lexicalised meanings: *skrive op*, *skrive ned*, *skrive af*, *skrive ud* etc. [lit. write up, write down, duplicate, write out (print out)].

Also at the syntactic level, Danish together with the other Scandinavian languages allows for a considerable number of movements, such as:

- *Hvem* troede du han sagde at hun kendte _?
  [Lit. *Who* thought you he said she knew _?]
- *Denne bog* ved vi hvem der har skrevet _.
  [Lit. *This book* know we who has written _.]
- *Denne bog* går der rygter om at du har læst _.
  [Lit. *This book* are there rumors that you know who has read _.]
- *Peter* ved jeg ikke om _ vil komme.
  [Lit. *Peter* know I not if _ will come.]

Danish exhibits various types of syntactic movement.

3.3 RECENT DEVELOPMENTS

During the last 50 years changes in the Danish language are dominated by:

- a tendency towards less dialectal variation;
- a less distinct pronunciation of certain sounds in the spoken language;
- some influence from English both on grammar (syntax and morphology) and lexis;
- a tendency to prefer English to foreign languages such as German and French.

The tendency towards less dialectal variation is favouring the Copenhagen dialect as the standard pronunciation used all over the country. Some researchers have declared that dialects in Denmark are already extinct, whereas others state that some regional variations are still traceable. This development has been enforced by a strong standardisation of the language in the media since around 1950 and little tolerance towards dialects in the school system.

As the Copenhagen dialect has become the dominant variety of spoken Danish, the changes in this dialect towards a less distinct pronunciation especially of certain vowels such as *a* and *e* affect the whole country. Some young speakers are for instance no longer able to pronounce a distinction between words like *ret* [right/court] and *rat* [steering wheel]. This tendency, however, already started in the Middle Ages.

Since the end of World War II the influence of the English language on Danish language users has increased. More than 25% of the courses taught at Danish universities are taught in English, and about 25% of large and

medium sized Danish companies have chosen English as their company language. This means that new words often are English loanwords such as *governance*, and that in some cases Danish words are competing with English equivalents, i. e., *deadline* instead of *tidsfrist*, *bodyguard* instead of *livvagt*. In many cases, however, it can be observed that the English words serve a different purpose than the Danish ones, i. e., the word *to book* is used equivalent to *bestille*, but *bestille* can also be used in the sense 'to order'. Thus, the semantic range of *booke* is narrower than the Danish near-synonym.

In a few cases it can be observed that English also affects Danish syntax. For instance word order in imperative clauses is changing. One of the most characteristic differences between Danish and English is the placement of the sentence adverbial in the clause. In English sentence adverbials always occur before the main verb, whereas in Danish unmarked clauses and imperatives they always occur after the verb. 'Please, close the door' corresponds to *luk venligst døren* [lit. close please the door]. However, during the last 15 years the English word order, i. e., *venligst luk døren* has become increasingly common.

Finally, it can be observed that Danish borrows new word senses of existing words from English. Thus 20 years ago the phrase *hænge ud* [hang out] could only be used in the sense 'to hang out your clothes', but now it can also be used in the sense 'to hang out with friends'. Such changes are typically also associated with a change in the valency structure or argument structure of the verb, i. e., in the case of *hænge ud* the use of a prepositional complement instead of an object.

The number of students of German, French, Italian and Russian has reduced dramatically over the last 10 years.

Due to the influence of English other foreign languages have become less attractive for young people, and the number of students of German, French, Italian and Russian has reduced dramatically over the last 10 years.

3.4 OFFICIAL LANGUAGE PROTECTION IN DENMARK

The central pivot of language cultivation in Denmark is the Ministry of Culture via the Danish Language Council. The purpose of the Danish Language Council is threefold:

- monitor the development of the Danish language and give advice and information on it. It determines the spelling of Danish;
- publish information on the Danish language, in particular those on the use of the native language, and co-operate with terminology centres, dictionary editors and public institutions involved in authorising or registering people's names, surnames and brand names;
- collaborate with language councils and institutions in other Nordic countries.

In addition to the Danish Language Council, The Society for Danish Language and Literature (an independent institution partially financed by the Ministry of Culture) edits and publishes Danish texts in scholarly editions as well as scholarly dictionaries. This is done both by means of book publications and on the Web. The institution also develops corpus collections of the Danish language.

The Danish Language Council is taking care of the Danish Language.

Furthermore, the private institution, Modersmål-Selskabet ('The Mother Tongue Association') has as vision to work towards preservation and development of

Danish. The society publishes member magazines, yearbooks and arranges talks and workshops on the Danish language.

Centre for Language Technology is the national centre for language technology.

Finally, with the aim of supporting language cultivation from the technological angle, Centre for Language Technology at the University of Copenhagen is the national centre for language technology with the mission of carrying out and promoting strategic research and application development in the areas of Danish language technology. Apart from the main aim of assuring good language technology for Danish users – and other users of the Danish language, the centre aims at bringing new knowledge to Denmark through international co-operation.

The Danish Language Council and the Society for Danish Language and Literature have established a Danish language website which collects information about the Danish language and its conditions of use, *Sproget.dk*. The site's aim is to provide professional help by informing about linguistic topics, and it offers simultaneous search in several Danish dictionaries as well as access to FAQ's and articles about various linguistic problems.

Furthermore, a joint language awareness effort, the so-called 'Gang i sproget' campaign was launched in September 2010 by the Danish Language Council and the Society for Danish Language and Literature for the Danish Government and will be running for the next two years. The aim of the campaign is to stimulate the interest in Danish and to develop more knowledge about the Danish language. The campaign includes a website (with a language test), seminars and television programs on issues of the Danish language.

Other parameters which indicate the level of language cultivation in Denmark relate to the number of book titles and newspapers published in Danish, as well as the number of television channels broadcasting in Danish. Danish Library Center states in its annual statistics that 7707 titles (including both fiction and technical language) were published in Danish in 2010, and 220 Danish titles were translated to other languages. 2261 foreign titles were translated to Danish the same year. With regard to the number of published newspapers, Dansk Oplagskontrol states that in 2010 34 daily newspapers were published in Danish. Ten of these are national newspapers, and they complete a daily print run of approx. 584,000 copies [11].

Denmark has six national television channels, three of which (DR1, DR2, TV2) are paid via a general license fee. In addition, several local television channels broadcast every day. According to a law of December 2002 on public radio and television services "programming must ensure public access to information and important social debates. It must also draw on Danish language and culture […]". DR's language policy states that a significant proportion of programs must be in Danish or designed for a Danish audience.

3.5 LANGUAGE IN EDUCATION

Danish is a mandatory subject in schools of Denmark as well as in the Danish territories of the Faroe Islands (where it is also an official language) and Greenland. In the former crown holding of Iceland, Danish is offered as a second language parallel to the other Scandinavian languages.

Following the ruling, Danish is defined as a necessary school subject since it enables citizens to participate in the democratic process. This subject is at the very centre of teaching, since it allows the student to integrate into Danish society. Immigrant children have been taught Danish as a Foreign Language in primary and secondary school since 1993. Danish is furthermore taught at masters level at five universities in Denmark.

In the scientific domain, English is dominating the scene in Denmark to an increasing extent. More than 25% of all university courses are taught in English, in sciences MA courses are almost exclusively taught in English. Furthermore, a vast majority of scientific papers are written in English. In sum, the number of scientific journals in Danish is decreasing, and these journals generally do not have the same scientific status as the international ones.

Danish has to be preserved as a functioning language at all levels of education and research.

In its language policy, University of Copenhagen – the largest university in Denmark – stresses the principle of parallelism between the languages English and Danish, and states that English is foreseen to become the "lingua franca" of research in the future and thereby also a more dominant language for teaching and education. However, it should at the same time be ensured that Danish is preserved as a functioning language at all levels of education and research.

3.6 INTERNATIONAL ASPECTS

Danish has been one of the official languages of the European Union since 1973. Apart from ratifying the Nordic Language Convention as mentioned above, Denmark is also included in the awareness plan established by The Nordic Council of Ministers in 2007. Here language technology has been identified as a central factor for protecting and maintaining our languages and our culture. This council has commissioned a ten-year plan, in form of an expert panel report, for making the Nordic countries a leading region in language technology. As a part of the plan, several Danish companies and researchers are members of NEALT (Northern European Association for Language Technology) which is an organisation for coordinating various initiatives and networks regarding education, research and awareness in the language technology field. Finally, University of Copenhagen is a member of ELRA (European Language Resources Association) since its creation.

The Nordic Council of Ministers identified language technology as a central factor for protecting and maintaining the Nordic languages and culture.

3.7 DANISH ON THE INTERNET

According to the latest statistics from 2010, 4,750,500 persons in Denmark are Internet users, amounting to 86% of the population [12]. Among young people, the proportion of users is considered to be even higher. This is a very high user percentage compared to European standards and shows that the Danes are generally technology advanced. With regard to Danish Internet domains, there are more than 1,000,000 domains registered in 2011 [13]. Danes tend to use Internet sites in Danish; most public service sites, however, are available in both Danish and English.

Danes are generally technology advanced.

For language technology, the extensive use of the Internet in Denmark is important in two ways. On the one hand, the large amount of digitally available language data represents a rich source for analysing the usage of natural language, in particular by collecting statistical information. On the other hand, the Internet offers a wide range of application areas for language technology. The most commonly used web application is certainly web search, which involves the automatic processing of language on multiple levels, as we will see in more detail

in the second part of this paper. It involves sophisticated language technology, also for Danish.

Internet users and providers of web content can also profit from language technology in less obvious ways, e. g., when language technology is used to automatically translate web contents from one language into another. Considering the high costs associated with manually translating these contents, comparatively little usable language technology is developed and applied, compared to the anticipated need.

The next chapter gives an introduction to language technology and its core application areas, together with an evaluation of current language technology support for Danish.

4

LANGUAGE TECHNOLOGY SUPPORT FOR DANISH

Language technology is used to develop software systems designed to handle human language and are therefore often called "human language technology". Human language comes in spoken and written forms. While speech is the oldest and in terms of human evolution the most natural form of language communication, complex information and most human knowledge is stored and transmitted through the written word. Speech and text technologies process or produce these different forms of language, using dictionaries, rules of grammar, and semantics. This means that language technology (LT) links language to various forms of knowledge, independently of the media (speech or text) in which it is expressed. Figure 1 illustrates the LT landscape.

When we communicate, we combine language with other modes of communication and information media – for example speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other multimodal communication and multimedia technologies.

In this section, we will discuss the main application areas of language technology, i. e., language checking, web search, speech interaction, and machine translation. These applications and basic technologies include

- spelling correction
- authoring support
- computer-assisted language learning
- information retrieval
- information extraction
- text summarisation
- question answering
- speech recognition
- speech synthesis

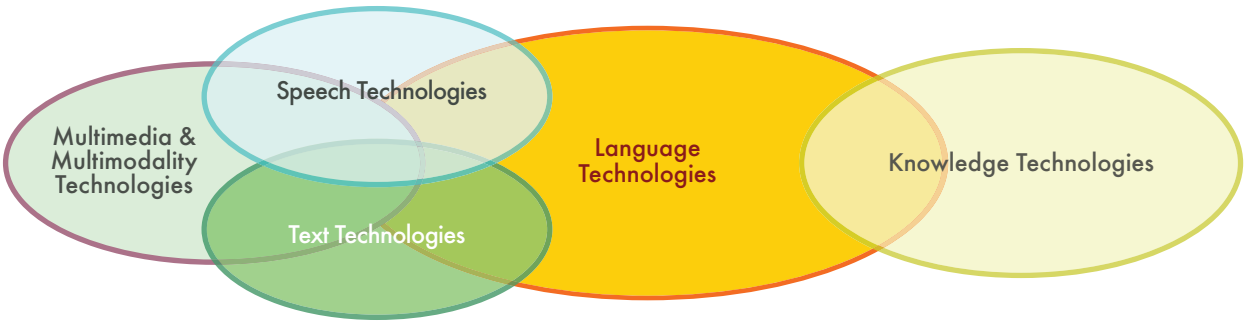
Language technology is an established area of research with an extensive set of introductory literature. The interested reader is referred to the following references: [9, 14, 15, 16, 17].

Before discussing the above application areas, we will briefly describe the architecture of a typical LT system.

4.1 APPLICATION ARCHITECTURES

Software applications for language processing typically consist of several components that mirror different aspects of language. While such applications tend to be very complex, figure 2 shows a highly simplified architecture of a typical text processing system. The first three modules handle the structure and meaning of the text input:

1. Pre-processing: cleans the data, analyses or removes formatting, detects the input languages, and so on.
2. Grammatical analysis: finds the verb, its objects, modifiers and other sentence elements; detects the sentence structure.



1: Language technology in context

3. Semantic analysis: performs disambiguation (i. e., computes the appropriate meaning of words in a given context); resolves anaphora (i. e., which pronouns refer to which nouns); represents the meaning of the sentence in a machine-readable way.

After analysing the text, task-specific modules can perform other operations, such as automatic summarisation and database look-ups.

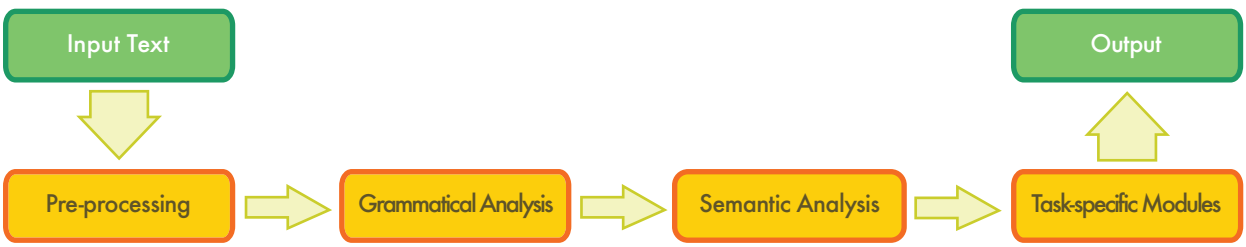
In the remainder of this section, we firstly introduce the core application areas for language technology, and follow this with a brief overview of the state of LT research and education today, and a description of past and present research programmes. Finally, we present an expert estimate of core LT tools and resources for Danish in terms of various dimensions such as availability, maturity and quality. The general situation of LT for the Danish language is summarised in a matrix (figure 7, p. 61) at the end of this chapter. Tools and resources that are boldfaced in the text can also be found in figure 7. LT support for Danish is also compared to other languages that are part of this series.

4.2 CORE APPLICATION AREAS

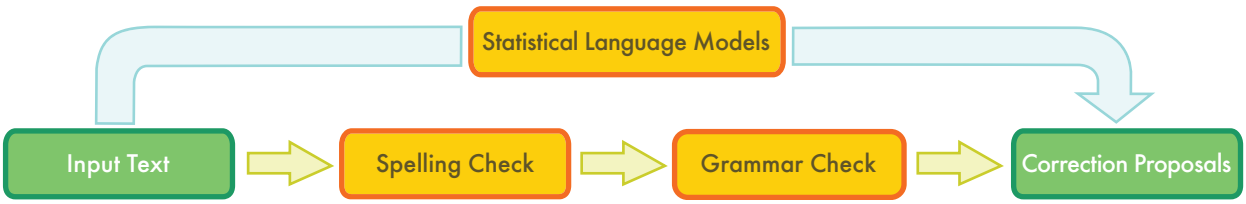
In this section, we focus on the most important LT tools and resources, and provide an overview of LT activities in Denmark.

4.2.1 Language Checking

Anyone who has used a word processor such as Microsoft Word knows that it has a spell checker that highlights spelling mistakes and proposes corrections. The first spelling correction programs compared a list of extracted words against a dictionary of correctly spelled words. Today these programs are far more sophisticated.



2: A typical text processing architecture



3: Language checking (statistical; rule-based)

Using language-dependent algorithms for **grammatical analysis**, they detect errors related to morphology (e. g., plural formation) as well as syntax–related errors, such as a missing verb or a conflict of verb-subject agreement (e. g., *she *write a letter*). However, most spell checkers will not find any errors in the following text [18]:

> *I have a spelling checker,*
> *It came with my PC.*
> *It plane lee marks four my revue*
> *Miss steaks aye can knot sea.*

Handling these kinds of errors usually requires an analysis of the context. This type of analysis either needs to draw on language-specific **grammars** laboriously coded into the software by experts, or on a statistical language model. In this case, a model calculates the probability of a particular word as it occurs in a specific position (e. g., between the words that precede and follow it). For example: *stort hus* [big house] is a much more probable word sequence than *stor hus* [big house]. A statistical language model can be automatically created by using a large amount of (correct) language data, a **text corpus**. Most of these two approaches have been developed around data from English. Neither approach can transfer easily to Danish because the language has unlimited compound building and a richer inflection system.

Thus, a serious defect of early stage Danish spell checkers was the incorrect markings of productive compounds, for instance, *pasningsordning* [childcare arrangement]. If these were not lexicalised in dictionaries or word lists (where the productive ones are generally not), they resulted in an error marking. Such wrong markings appeared due to lack of high-quality compound splitters for Danish that could check each component of the compound per se. Unfortunately, this flaw in early stage spell checkers for Danish has lead to an increase in spelling errors. People are influenced partly by the split compound tendency in English, and partly by the fact that a split compound does not result in an error marking with a Danish spell checker. Recent spell checkers for Danish as provided for instance by Microsoft Office products have now improved with regards to this phenomenon.

Danish grammar checkers, however, are still at a rather initial level. They are generally able to identify some simple grammatical errors such as lack of short dependency concordance as in **den røde hus* [red house; wrong gender of *den* and *red*], whereas other grammatical errors are not identified, such as **jeg var kede af at du ikke kom* [I was sorry that you didn't come; wrong number on *ked* [sorry]].

Recently, also OpenOffice provides Danish language checking tools to a certain extent. Magenta has integrated several open source Danish lexical resources into the document processing tool, including writing aids such as synonymy look-up.

Specifically related to teaching, Mikro Værkstedet is one of the prime players in the market for digital teaching facilities, including reading tools for dyslexia clients as well as writing aids. Further, Ordbogen.com should be

mentioned for facilitating the use of online dictionaries. Language checking is not limited to word processors; it is also used in "authoring support systems", i. e., software environments in which manuals and other types of technical documentation for complex IT, healthcare, engineering and other products, are written. To offset customer complaints about incorrect use and damage claims resulting from poorly understood instructions, companies are increasingly focusing on the quality of technical documentation while targeting the international market (via translation or localisation) at the same time. Advances in natural language processing have led to the development of authoring support software, which helps the writer of technical documentation to use vocabulary and sentence structures that are consistent with industry rules and (corporate) terminology restrictions.

Language checking is not limited to word processors but also applies to authoring systems.

Besides spell checkers and authoring support, language checking is also important in the field of computer-assisted language learning. Language checking applications also automatically correct search engine queries, as found in Google's *Did you mean...* suggestions.

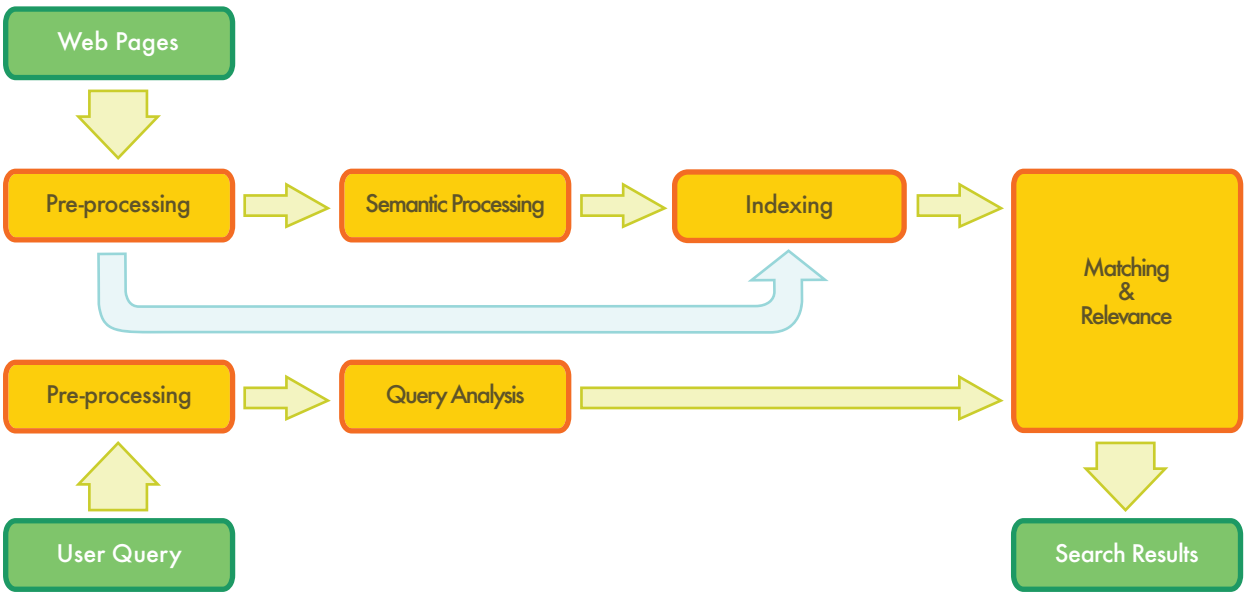
4.2.2 Web Search

Searching the Web, intranets or digital libraries is probably the most widely used yet largely underdeveloped language technology application today. The Google search engine, which started in 1998, now handles about 80% of all search queries [19]. The verb *google* even has an entry in the Danish dictionary 'Den Danske Ordbog'. The Google search interface and results page display has not significantly changed since the first version. However, in the current version, Google offers spelling correction for misspelled words and incorporates basic semantic search capabilities that can improve search accuracy by analysing the meaning of terms in a search query context [20]. The Google success story shows that a large volume of data and efficient indexing techniques can deliver satisfactory results using a statistical approach to language processing.

For more sophisticated information requests, it is essential to integrate deeper linguistic knowledge to facilitate text interpretation. Experiments using **lexical resources** such as machine-readable thesauri or ontological language resources (e. g., WordNet for English or DanNet for Danish) have demonstrated improvements in finding pages using synonyms of the original search terms, such as *autoforsikring* [auto insurance], *bilforsikring* [motor car insurance] and *kaskoforsikring* [insurance covering loss of or damage to the car], or even more loosely related terms.

The next generation of search engines will have to include much more sophisticated language technology.

The next generation of search engines will have to include much more sophisticated language technology, especially to deal with search queries consisting of a question or other sentence type rather than a list of keywords. For the query, *Give me a list of all companies that were taken over by other companies in the last five years*, a syntactic as well as **semantic analysis** is required. The system also needs to provide an index to quickly retrieve relevant documents. A satisfactory answer will require syntactic parsing to analyse the grammatical structure of the sentence and determine that the user wants companies that have been acquired, rather than companies that have acquired other companies. For the expression *last five years*, the system needs to determine the relevant range of years, taking into account the present year. The query then needs to be matched against a huge



4: Web search

amount of unstructured data to find the pieces of information that are relevant to the user's request. This process is called information retrieval, and involves searching and ranking relevant documents. To generate a list of companies, the system also needs to recognise a particular string of words in a document represents a company name, using a process called "named entity recognition".

A more demanding challenge is matching a query in one language with documents in another language. Cross-lingual information retrieval involves automatically translating the query into all possible source languages and then translating the results back into the user's target language.

Now that data is increasingly found in non-textual formats, there is a need for services that deliver multimedia information retrieval by searching images, audio files and video data. In the case of audio and video files, a speech recognition module must convert the speech content into text (or into a phonetic representation) that can then be matched against a user query.

In Denmark, SMEs like Ankiro, ScanJour, LAT Consulting, Findwise, RDFined and others successfully develop and apply search technologies that are tailored to specific company needs.

These companies focus their development on providing add-ons and advanced search engines for special interest portals by using topic-relevant semantics. Due to the constant high demand for processing power, such search engines are only cost-effective when handling relatively small text corpora. The processing time is several thousand times higher than that needed by a standard statistical search engine, like Google. These search engines are in high demand for topic-specific domain modelling, but they cannot be used on the Web with its billions and billions of documents. Furthermore, the technologies developed in these contexts are generally not available to the public for further research or development. Due to such practical obstacles, many Danish websites link to a Google search engine as their only search facility.

Experimental, ontology-based search engines have been developed at several Danish universities such as

Roskilde Universitet. The OntoQuery and SIABO prototypes are examples of such experimental search engines that work on smaller domains with a rich ontological representation. Again, however, such prototypes are not easily scalable to larger domains.

4.2.3 Speech Interaction

Speech interaction is one of many application areas that depend on speech technology, i. e., technologies for processing spoken language. Speech interaction technology is used to create interfaces that enable users to interact in spoken language instead of using a graphical display, keyboard and mouse. Today, these voice user interfaces (VUI) are used for partially or fully automated telephone services provided by companies to customers, employees or partners. Business domains that rely heavily on VUIs include banking, supply chain, public transportation, and telecommunications. Other uses of speech interaction technology include interfaces to car navigation systems and the use of spoken language as an alternative to the graphical or touchscreen interfaces in smartphones.

Speech interaction technology comprises four technologies:

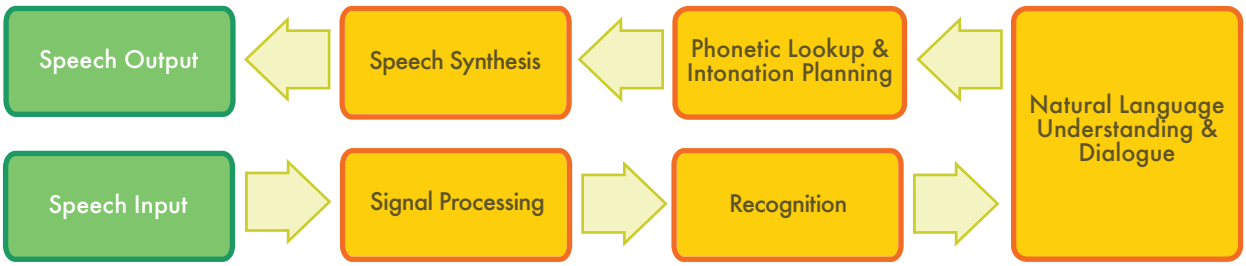
1. Automatic **speech recognition** (ASR) determines which words are actually spoken in a given sequence of sounds uttered by a user.
2. Natural language understanding analyses the syntactic structure of a user's utterance and interprets it according to the system in question.
3. Dialogue management determines which action to take given the user input and system functionality.
4. **Speech synthesis** (text-to-speech or TTS) transforms the system's reply into sounds for the user.

One of the major challenges of ASR systems is to accurately recognise the words a user utters. This means restricting the range of possible user utterances to a limited set of keywords, or manually creating language models that cover a large range of natural language utterances. Using machine learning techniques, language models can also be generated automatically from **speech corpora**, i. e., large collections of speech audio files and text transcriptions. Restricting utterances usually forces people to use the voice user interface in a rigid way and can damage user acceptance; but the creation, tuning and maintenance of rich language models will significantly increase costs. VUIs that employ language models and initially allow a user to express their intent more flexibly – prompted by a *How may I help you?* greeting – tend to be automated and are better accepted by users.

Speech interaction is the basis for interfaces that allow a user to interact with spoken language.

Companies tend to use utterances pre-recorded by professional speakers for generating the output of the voice user interface. For static utterances where the wording does not depend on particular contexts of use or personal user data, this can deliver a rich user experience. But more dynamic content in an utterance may suffer from unnatural intonation because different parts of audio files have simply been strung together. Through optimisation, today's TTS systems are getting better at producing natural-sounding dynamic utterances.

Interfaces in speech interaction have been considerably standardised during the last decade in terms of their various technological components. There has also been strong market consolidation in speech recognition and speech synthesis. The national markets in the G20 countries (economically resilient countries with high populations) have been dominated by just five global players, with Nuance (USA) and Loquendo (Italy) being the most prominent players in Europe. In 2011, Nuance announced the acquisition of Loquendo, which represents a further step in market consolidation.



5: Speech-based dialogue system

On the Danish market for speech technology solutions there are a number of national companies (Mikro Værkstedet, Prolog Development Center, Max Manus, as well as IBM and Siemens Denmark) that have specialised in providing speech-based solutions powered by technology developed by the leading international technology suppliers. Nuance is the predominant international speech technology provider of Danish speech technology. Virtually all other providers of Danish speech technology have been acquired by Nuance over the last 5-6 years, for example, Philips Speech Magic, Loquendo and SVOX.

Currently there are two software packages that offer Danish speech recognition, Nuance SpeechMagic and Nuance Dragon Development Platform (cloud-based). Based on these technologies Max Manus has developed an application for the health care sector, IBM an application for the municipal sector and Prolog Development Center the standard system Dictus and customised solutions for Folketingstidende (The Parliament Hansard) and two national broadcasters. Nuance developed the free Apps Dragon Dictation and Dragon Search for the iPhone/iPad while Prolog Development Center delivers Dictus to Android on the same cloud-based platform. Nuance is the only supplier of speech recognition telephone technology. Siemens Denmark and Prolog Development Center are the Danish companies that have supplied voice-controlled telephone service on the basis of Nuance Recognizer v9. Currently there are more than 10 different Danish voices for speech synthesis developed by Nuance (including Loquendo and SVOX), Acapela, and the Danish company Mikro Værkstedet. The successful Polish company IVONA is also planning to launch a number of Danish voices in early 2012.

A growing number of major international suppliers of consumer electronics (Garmin, Navigon, Samsung) offer products that incorporate Danish speech recognition so that the products can be controlled by voice. An even larger number of products, such as iRobot, have also built-in Danish speech synthesis. A small but growing number of Danish producers of consumer products utilise Danish speech technology, for example, 6th Sense Solution for reading bus stops and CIM Interconn for screen reading in nursing homes.

The demand for voice user interfaces in Denmark has grown fast in the last five years, driven by increasing demand for customer self-service, cost optimisation for automated telephone services, and the increasing acceptance of spoken language as a media for human-machine interaction.

Looking ahead, there will be significant changes, due to the spread of smartphones as a new platform for managing customer relationships, in addition to fixed telephones, the Internet and e-mail. This will also affect how speech interaction technology is used. In the long

term, there will be fewer telephone-based VUIs, and spoken language apps will play a far more central role as a user-friendly input for smartphones. This will be largely driven by stepwise improvements in the accuracy of speaker-independent speech recognition via the speech dictation services already offered as centralised services to smartphone users.

4.2.4 Machine Translation

The idea of using digital computers to translate natural languages can be traced back to 1946 and was followed by substantial funding for research during the 1950s and again in the 1980s. Yet **machine translation** (MT) still cannot deliver on its initial promise of providing across-the-board automated translation.

At its basic level, Machine Translation simply substitutes words in one natural language with words in another language.

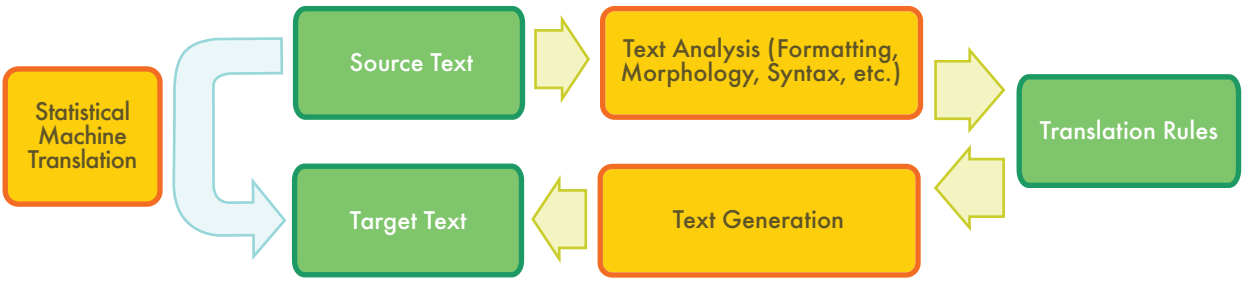
The most basic approach to machine translation is the automatic replacement of the words in a text written in one natural language with the equivalent words of another language. This can be useful in subject domains that have a very restricted, formulaic language such as weather reports. However, in order to produce a good translation of less restricted texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty is that human language is ambiguous. Ambiguity creates challenges on multiple levels, such as word sense disambiguation at the lexical level (a *jaguar* is a brand of car or an animal) or the assignment of case on the syntactic level, for example:

- *The woman saw the car and her husband, too.*

One way to build an MT system is to use linguistic rules. For translations between closely related languages, a translation using direct substitution may be feasible in cases such as the above example. However, rule-based (or linguistic knowledge-driven) systems often analyse the input text and create an intermediary symbolic representation from which the target language text can be generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by skilled linguists. This is a very long and therefore costly process. In the late 1980s when computational power increased and became cheaper, interest in statistical models for machine translation began to grow. Statistical models are derived from analysing bilingual text corpora, **parallel corpora**, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 21 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text by processing parallel versions and finding plausible patterns of words. Unlike knowledge-driven systems, however, statistical (or data-driven) MT systems often generate ungrammatical output. Data-driven MT is advantageous because less human effort is required, and it can also cover special particularities of the language (e. g., idiomatic expressions) that are often ignored in knowledge-driven systems.

Machine Translation is particularly challenging for the Danish language.

The strengths and weaknesses of knowledge-driven and data-driven machine translation tend to be complementary, so that nowadays researchers focus on hybrid approaches that combine both methodologies. One such approach uses both knowledge-driven and data-driven systems, together with a selection module that decides on the best output for each sentence. However, results for sentences longer than, say, 12 words, will often be



6: Machine translation (statistical; rule-based)

far from perfect. A more effective solution is to combine the best parts of each sentence from multiple outputs; this can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

The potential for creating arbitrary new words by compounding makes dictionary analysis and dictionary coverage difficult; split verb constructions pose problems for analysis; and extensive lexical ambiguity is a challenge for word sense disambiguation. Apart from the PaTrans system (a patent translation system for English-Danish) early MT systems for Danish, like the SYSTRAN prototype, were primarily developed by foreign companies. All of these systems are rule-based. Although significant research in this technology exists in national and international contexts, this situation has not substantially changed, even if there are some Danish start-ups such as Grammar Soft and Languagelens, providing rule-based and statistical machine translation systems for Danish. A majority of the available systems, like, e. g., Google Translate and ESTeam Translator, is still developed abroad.

There is still a huge potential for improving the quality of MT systems. The challenges involve adapting language resources to a given subject domain or user area, and integrating the technology into workflows that already have term bases and translation memories. Another problem is that most of the current systems are English-centred and only support a few languages from and into Danish. This leads to friction in the translation workflow and forces MT users to learn different lexicon coding tools for different systems.

Evaluation campaigns help to compare the quality of MT systems, the different approaches and the status of the systems for different language pairs. Figure 7 (p. 24), which was prepared during the EC Euromatrix+ project, shows the pair-wise performances obtained for 22 of the 23 official EU languages (Irish was not compared). The results are ranked according to a BLEU score, which indicates higher scores for better translations [21]. A human translator would normally achieve a score of around 80 points.

The best results (in green and blue) were achieved by languages that benefit from a considerable research effort in coordinated programmes and the existence of many parallel corpora (e. g., English, French, Dutch, Spanish and German). The languages with poorer results are shown in red. These languages either lack such development efforts or are structurally very different from other languages (e. g., Hungarian, Maltese and Finnish).

4.3 OTHER APPLICATION AREAS

Building language technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but they provide significant

service functionalities "behind the scenes" of the system in question. They all form important research issues that have now evolved into individual sub-disciplines of computational linguistics. Question answering, for example, is an active area of research for which annotated corpora have been built and scientific competitions have been initiated. The concept of question answering goes beyond keyword-based searches (in which the search engine responds by delivering a collection of potentially relevant documents) and enables users to ask a concrete question to which the system provides a single answer. For example:

> *Question: How old was Neil Armstrong when he stepped on the moon?*
> *Answer: 38.*

While question answering is obviously related to the core area of web search, it is nowadays an umbrella term for such research issues as which different types of questions exist, and how they should be handled; how a set of documents that potentially contain the answer can be analysed and compared (do they provide conflicting answers?); and how specific information (the answer) can be reliably extracted from a document without ignoring the context.

Language technology applications often provide significant service functionalities behind the scenes of larger software systems.

Question answering is in turn related to information extraction (IE), an area that was extremely popular and influential when computational linguistics took a statistical turn in the early 1990s. IE aims to identify specific pieces of information in specific classes of documents, such as the key players in company takeovers as reported in newspaper stories. Another common scenario that has been studied is reports on terrorist incidents. The task here consists of mapping appropriate parts of the text to a template that specifies the perpetrator, target, time, location and results of the incident. Domain-specific template-filling is the central characteristic of IE, which makes it another example of a "behind the scenes" technology that forms a well-demarcated research area, which in practice needs to be embedded into a suitable application environment.

Text summarisation and **text generation** are two borderline areas that can act either as standalone applications or play a supporting role. Summarisation attempts to give the essentials of a long text in a short form, and is one of the features available in Microsoft Word. It mostly uses a statistical approach to identify the "important" words in a text (i. e., words that occur very frequently in the text in question but less frequently in general language use) and determine which sentences contain the most of these "important" words. These sentences are then extracted and put together to create the summary. In this very common commercial scenario, summarisation is simply a form of sentence extraction, and the text is reduced to a subset of its sentences. An alternative approach, for which some research has been carried out, is to generate brand new sentences that do not exist in the source text.

For the Danish language, research in most text technologies is much less developed than for the English language.

This requires a deeper understanding of the text, which means that so far this approach is far less robust. On the whole, a text generator is rarely used as a stand-alone application but is embedded into a larger software environment, such as a clinical information system that collects, stores and processes patient data. Creating reports is just one of many applications for text summarisation. For the Danish language, research in these text

technologies is much less developed than for the English language. Question answering, information extraction, and summarisation have been the focus of numerous open competitions in the USA since the 1990s, primarily organised by the government-sponsored organisations DARPA and NIST. These competitions have significantly improved the state of the art, but their focus has mostly been on the English language. As a result, there are hardly any annotated corpora or other special resources for these tasks in Danish.

Some Danish SMEs (such as RDFined and Ankiro) and media (Information, InfoMedia, DR) are engaged in small, task-driven initiatives on named entity recognition and knowledge extraction. Some of these draw on tools developed at the Centre for Language Technology at the University of Copenhagen: part of speech tagger, lemmatiser, keyword extractor. Some apply the computational dictionary STO, and/or the Danish wordnet (DanNet), which are resources that have been developed in collaboration with other institutions. Likewise a summarisation system, DanSum, is available online at the Centre's website. For text generation, reusable components have traditionally been limited to the surface realisation modules (the "generation grammar"); again, most available software is for English.

4.4 LANGUAGE TECHNOLOGY IN RESEARCH AND EDUCATION

Language technology research is performed at several research institutions in Denmark and through several collaborative infrastructure and research projects. In the following, only prime movers in the field as well as currently ongoing projects are listed, acknowledging the fact that related research takes place also at other institutions and in other projects than the ones mentioned here.

As previously mentioned, Centre for Language Technology at the University of Copenhagen is the national centre for language technology. The Centre's main research topics are language resources and tools such as LSP corpora and wordnets, multilinguality (machine translation, controlled language etc.), multimodality, information retrieval, use of ontologies and incorporation of language technology in other application areas, as for example e-learning.

The Department of International Language Studies and Computational Linguistics, Copenhagen Business School, performs research in text technology and computational linguistics including treebanks, statistical machine translation, terminology and speech technology. Taking their point of departure in language, text and translation studies, researchers at the department in general focus on how companies can deal with their language processes professionally in a globalised world and on issues related to language for special purposes (LSP). Copenhagen Business School includes the DANTERMcentre which is a centre for terminology and development of terminology tools.

At the Institute of Language and Communication, University of Southern Denmark, several researchers work within the field of language technology and computational linguistics, for instance within the VISL project (Visual Interactive Syntax Learning). Aalborg University, Department of Electronic Systems, is a prime mover in research regarding speech technology. The Danish Language Council as well as the Society for Danish Language and Literature are also engaged in language technology-related research, especially with regard to the development of dictionaries and corpora and the corpus-based identification of new words in Danish.

Several research infrastructure projects are ongoing in Denmark, such as DK-CLARIN, which has the aim to construct a Danish research infrastructure for the humanities integrating written, spoken, and visual records

into a coherent and systematic digital repository. Moreover, LARM has the aim of constructing a national, digital and user-driven research infrastructure, which will provide a solution for preserving and maturing cultural heritage radio source material.

There are also several collaborative research projects such as ESICT, NOMCO and DanTermBank. ESICT performs research in methods and development of technologies providing citizens with an innovative information system on health and disease, based on information technology, language technology and formalised medical knowledge. NOMCO is a Nordic collaborative project which deals with multimodal corpus analysis. DanTermBank is a recently initiated project concerned with the development of the technological foundations for national term bank for languages for special purposes. Furthermore, Danish research centers participate in several ongoing European projects related to language technology, LetsMT!, CLARIN, CLARA, META-NORD/META-NET to mention a few.

In contrast to the relatively high-level research activity in language technology in Denmark, the educational situation can only be labeled as critical. Currently, there exists no bachelor's or master's degree in language technology in Denmark. Language technology is only taught as a component of other bachelor's and master's educations in Denmark. Examples are the master's program in *IT and Cognition* at the University of Copenhagen, which encompasses an obligatory course in language technology, as well as the bachelor's program in *Business Communication and IT* at the University of Southern Denmark, which includes an optional branch of language technology. The lack of education in basic language technology is problematic and constitutes a rather new situation in the country, since computational linguistics was taught until recently both at the Copenhagen Business School and the University of Copenhagen. The current situation should encourage political actions since future research and development of Danish language technology may be in danger. Paradoxically, the lack of education in the field coincides with an increasing interest in and demand for language technology in Danish industry.

4.5 NATIONAL PROJECTS AND INITIATIVES

The Danish research councils have supported language technology projects over the years, and in the past also created specific research programmes supporting the field.

The European EUROTRA programme was probably the first large scale support for language technology in Denmark. The EUROTRA effort aimed at creating multilingual machine translation for all official European languages. It started late 1970s and ran till early 1990s. A Dane was chairing the EUROTRA Liaison Group 1986-1992, which brought some attention to the project from Danish players. Although EUROTRA did not produce a ready to use multilingual machine translation system, it did provide a prototype and thereby the necessary background for a patent machine translation system for English-Danish, running in production for more than 10 years.

One of the early programmes created by the Research Council for the Humanities was ETTO (*Edb for Tekst, Tale og Ordbøger* [Computing for Text, Speech and Dictionaries]) 1982-85. The next programme was *Eksperimentel Sprogvidenskab* [Experimental linguistics]) running from 1991. Since then, the Research Council for Humanities has not had specific programmes aiming at language technology.

The Research Council for the Technical Sciences had speech technology as one of the main themes in their strategy plan for 1998-2002, and they did support several language technology projects.

The inter-disciplinary IT research programmes running from mid 1990s to 2003 have been able to support a few language technology related projects (as was the case for OntoQuery and SIABO), but have not had a specific aim to do so.

The Research Agency of the Danish Ministry for Science, Technology and Development decided in 2001 to promote language technology by funding a collaborative effort to enlarge the Danish language technology lexicon developed in the European PAROLE project. The resulting dictionary (STO) is now available through ELRA, and is being used for research and commercial purposes. It was the specific aim of the Research Agency to support the Danish language in the digital age, by granting funds for the development of basic language technology building blocks.

Since this funding, to our knowledge there have been no programmes for the support of language technology specifically, but as always, language technology projects may be supported by research councils etc. in competition with other themes.

In parallel with the ESFRI initiative for research infrastructure, Denmark has started a roadmap process for research infrastructures. One of the research infrastructures to be supported will be a "Digital Humanities Laboratory" for the support of humanities research through the use of i.a. language technology.

As we have seen, previous programmes have led to the development of a number of LT tools and resources for the Danish language. The following section summarises the current state of LT support for Danish.

4.6 AVAILABILITY OF TOOLS AND RESOURCES

Figure 7 provides a rating for language technology support for the Danish language. This rating of existing tools and resources was generated by leading experts in the field who provided estimates based on a scale from 0 (very low) to 6 (very high) using seven criteria.

The key results for Danish language technology can be summed up as follows:

- In speech technology, there exist some tools, but almost exclusively available on a commercial basis. Regarding their quality however, commercial developers and researchers disagree. Where commercial developers argue that the recognition rate is at least 95%, the researchers point out that this high recognition rate can only be achieved if the speaker speaks very clearly and the system knows the user's voice in advance. Due to this disagreement, quality has not been rated in figure 7.
- With respect to basic text analysis the situation in Denmark is reasonably good. There exist a couple of tokenisers, part of speech taggers and morphological analysers of relatively high quality and broad coverage. However, not all of them are freely available. Broad coverage syntactic parsers on the other hand are sparse even if work has been done on constraint grammars and dependency parsing.
- Semantics is more difficult to process than syntax, and text semantics is more difficult to process than word and sentence semantics. Semantic tools and resources are scored very low.
- There is substantial research going on in the field of machine translation at several Danish research institutions, in particular on statistical machine translation. However, the quality is still low, and the number of language pairs is limited.
- With regard to resources such as reference corpora, lexicons, wordnets and terminologies, the situation is also reasonably good for Danish since substantial resources have been built in recent years; however, IPR issues prevent that all corpora become freely available. While some reference corpora of

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	4	2	–*	4	4	3	3
Speech Synthesis	5	2	–*	3	3	2	3
Grammatical analysis	3	2	4	4	3	2	3
Semantic analysis	1	0	1	1	1	1	1
Text generation	0	0	0	0	0	0	0
Machine translation	3	2	2	3	3	1	2
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	4	3	4	3	4	4	3
Speech corpora	3	3	3	1	2	2	2
Parallel corpora	3	1	3	2	2	2	2
Lexical resources	3	3	4	4	3	3	3
Grammars	2	1	4	1	2	2	2

7: State of language technology support for Danish (*quality has not been rated; see explanation in the text)

high quality exist, large syntactically and semantically annotated corpora are not available. For multimodal corpora, some annotated corpora are under development, but not yet available. There is a lack of parallel corpora needed for statistical and hybrid approaches to machine translation.

- Several of the resources lack standardisation, i. e., even if they exist, sustainability is not given.

In a number of specific areas of Danish language research, we have software with limited functionality available today. Obviously, further research efforts are required to meet the current deficit in processing texts on a deeper semantic level and to address the lack of resources such as parallel corpora for machine translation. Concerted programs and initiatives are needed to standardise data and interchange formats.

4.7 CROSS-LANGUAGE COMPARISON

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, this section will present an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The languages were categorised using a five-point scale:

1. Excellent support
2. Good support
3. Moderate support
4. Fragmentary support
5. Weak or no support

LT support was measured according to the following criteria:

Speech Processing: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Machine Translation: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Text Analysis: Quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars.

Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars.

Figures 8 to 11 show that Denmark is roughly estimated to be at the same level as its Scandinavian neighbors (Sweden, Norway and Finland). However, Denmark is in the second-worst category for speech processing (although, as mentioned above, there is some disagreement among researchers and commercial developers about the quality), text analysis and resources. The figures are even lower im machine translation. Here, Denmark is in the worst category, along with many other countries.

Thus, for building more sophisticated applications, such as machine translation, there is a clear need for resources and technologies that cover a wider range of linguistic aspects and enable a deep semantic analysis of the input text. By improving the quality and coverage of these basic resources and technologies, we shall be able to open up new opportunities for tackling a broader range of advanced application areas, including high-quality machine translation.

4.8 CONCLUSIONS

In this series of white papers, we have made an important effort by assessing the language technology support for 30 European languages, and by providing a high-level comparison across these languages. By identifying the gaps, needs and deficits, the European language technology community and its related stakeholders are now in a position to design a large scale research and development programme aimed at building a truly multilingual, technology-enabled communication across Europe.

The results of this white paper series show that there is a dramatic difference in language technology support between the various European languages. While there are good quality software and resources available for some languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources. Others have basic tools and resources but the implementation of, for example semantic methods, is still far away. Therefore a large-scale effort is needed to attain the ambitious goal of providing high-quality language technology support for all European languages, for example through high quality machine translation.

For Danish, the results indicate that only with respect to the most basic tools and resources the situation is reasonably good. Furthermore, there exist some systems for information extraction, machine translation and speech recognition and synthesis, as well as resources like parallel corpora, and speech corpora. However, these systems and resources are rather simple and have a limited functionality for some of the areas. For instance, parallel corpora only exist for very few language pairs.

With respect to more advanced fields like sentence and text semantics, information retrieval, language generation, and annotated multimodal data, Danish clearly

lacks systems, tools and resources even if some of these are currently under development. For advanced semantic and discourse processing and dialogue management resources are very scarce and systems have a quite limited scope.

Another point which is not fully reflected in the table above is that language technology resources and tools for Danish alone do not necessarily facilitate globalisation and international cooperation. "Cross- and multilingual" technologies that link Danish to other languages (for instance machine translation, multilingual retrieval, parallel corpora, linked wordnets, etc.) are a prerequisite for high-level, technology-aided interaction with our surroundings.

Given the present situation and the importance of language technology as the key for protecting and furthering the Danish language in the information-driven society, these results indicate that there is an indispensable need for new programmes specifically focusing on the development of language technology systems, tools and resources. In contrast to several other Nordic countries which have, for example, initiated projects concerning the development of a language bank encompassing a set of basic resources and tools (a BLARK), recent research and development in Denmark has been performed in full competition with other research themes. So even if Danish language technology has improved in recent years, and even if Danish industry is moving clearly forward in the field, there is an imminent danger that we are not moving fast enough. It is a political decision to change the speed of this development.

Finally there is a lack of continuity in research and development funding. Short-term coordinated programmes tend to alternate with periods of sparse or zero funding. In addition, there is an overall lack of coordination with programmes in other EU countries and at the European Commission level.

The long term goal of META-NET is to enable the creation of high-quality language technology for all languages. This requires all stakeholders – in politics, research, business, and society – to unite their efforts. The resulting technology will help tear down existing barriers and build bridges between Europe's languages, paving the way for political and economic unity through cultural diversity.

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French German Italian Portuguese Spanish	Basque Bulgarian Catalan **Danish** Estonian Galician Greek Hungarian Irish Norwegian Polish Serbian Slovak Slovene Swedish	Croatian Icelandic Latvian Lithuanian Maltese Romanian

8: Speech processing: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	Catalan Dutch German Hungarian Italian Polish Romanian	Basque Bulgarian Croatian Czech **Danish** Estonian Finnish Galician Greek Icelandic Irish Latvian Lithuanian Maltese Norwegian Portuguese Serbian Slovak Slovene Swedish

9: Machine translation: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch French German Italian Spanish	Basque Bulgarian Catalan Czech **Danish** Finnish Galician Greek Hungarian Norwegian Polish Portuguese Romanian Slovak Slovene Swedish	Croatian Estonian Icelandic Irish Latvian Lithuanian Maltese Serbian

10: Text analysis: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German Hungarian Italian Polish Spanish Swedish	Basque Bulgarian Catalan Croatian **Danish** Estonian Finnish Galician Greek Norwegian Portuguese Romanian Serbian Slovak Slovene	Icelandic Irish Latvian Lithuanian Maltese

11: Speech and text resources: State of support for 30 European languages

ABOUT META-NET

META-NET is a Network of Excellence partially funded by the European Commission. The network currently consists of 54 research centres in 33 European countries. META-NET forges **META**, the Multilingual Europe Technology Alliance, a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- grants all Europeans equal access to information and knowledge regardless of their language;
- builds upon and advances functionalities of networked information technology.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots.

Launched on 1 February 2010, META-NET has already conducted various activities in its three lines of action META-VISION, META-SHARE and META-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. The present White Paper was prepared together with volumes for 29 other languages. The shared technology vision was developed in three sectorial Vision Groups. The META Technology Council was established in order to discuss and to prepare the SRA based on the vision in close interaction with the entire LT community.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.

office@meta-net.eu – http://www.meta-net.eu

REFERENCER REFERENCES

[1] Aljoscha Burchardt, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk. *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.

[2] Directorate-General Information Society & Media of the European Commission. User Language Preferences Online, 2011. http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

[3] European Commission. Multilingualism: an Asset for Europe and a Shared Commitment, 2008. http://ec.europa.eu/languages/pdf/comm2008_en.pdf.

[4] Directorate-General of the UNESCO. Intersectoral Mid-term Strategy on Languages and Multilingualism, 2007. http://unesdoc.unesco.org/images/0015/001503/150335e.pdf.

[5] Directorate-General for Translation of the European Commission. Size of the Language Industry in the EU, 2009. http://ec.europa.eu/dgs/translation/publications/studies.

[6] Wikipedia: Den frie encyklopædi (Wikipedia: The Free Encyclopedia). Danmark (Denmark). http://da.wikipedia.org/wiki/Danmark.

[7] Frans Gregersen. Kom glad med happy ears (lit.: Come happy with happy ears). *Foredrag ved seminaret "Gang i Sproget" på Nationalmuseet (Talk at the seminar "Gang i Sproget" at The National Museum)*, 2010.

[8] Lars Hellan and Kirsti Koch Christensen (eds.). *Topics in Scandinavian Syntax*. D. Reidel, Dordrecht, 1986.

[9] Anna Braasch, Costanza Navarretta, Sanni Nimb, Sussi Olsen, Patrizia Paggio, Bolette Sandford Pedersen, and Jürgen Wedekind (eds.). *Sprogteknologi i et dansk perspektiv – En samling artikler om sprogforskning og automatisk sprogbehandling (Language Technology in a Danish Perspective – A Collection of Articles on Linguistics and Automatic Language Processing)*. C.A. Reitzels Forlag, København, 2006.

[10] Den Store Danske: Gyldendals åbne encyklopædi (The Great Danish: Gyldendal's Open Encyclopedia). Dansk (Danish). http://www.denstoredanske.dk/Samfund,_jura_og_politik/Sprog/Dansk/dansk.

[11] Dansk Oplagskontrol (The Danish Audit Bureau of Circulation). http://www.do.dk.

[12] Internet World Stats: Usage and Population Statistics. http://www.internetworldstats.com.

[13] DK Hostmaster A/S. https://www.dk-hostmaster.dk/presse/statistik/antal-registrerede-domaener.

[14] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 2nd edition, 2009.

[15] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[16] Language Technology World (LT World). http://www.lt-world.org.

[17] Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Battista Varile, Annie Zaenen, and Antonio Zampolli, editors. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, 1998.

[18] Jerrold Howard Zar. Candidate for a Pullet Surprise. *Journal of Irreproducible Results*, page 13, 1994.

[19] Spiegel Online. Google zieht weiter davon (Google is still leaving everybody behind), 2009. http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html.

[20] Juan Carlos Perez. Google Rolls out Semantic Search Capabilities, 2009. http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA, 2002.

[22] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 Machine Translation Systems for Europe. In *Proceedings of MT Summit XII*, 2009.

META-NET MEDLEMMER

META-NET MEMBERS

Belgien	Belgium	Computational Linguistics and Psycholinguistics Research Centre, University of Antwerp: Walter Daelemans
		Centre for Processing Speech and Images, University of Leuven: Dirk van Compernolle
Bulgarien	Bulgaria	Inst. for Bulgarian Language, Bulgarian Academy of Sciences: Svetla Koeva
Cypern	Cyprus	Language Centre, School of Humanities: Jack Burston
Danmark	Denmark	Centre for Language Technology, University of Copenhagen: Bolette Sandford Pedersen, Bente Maegaard
Estland	Estonia	Inst. of Computer Science, University of Tartu: Tiit Roosmaa, Kadri Vider
Finland	Finland	Computational Cognitive Systems Research Group, Aalto University: Timo Honkela
		Department of Modern Languages, University of Helsinki: Kimmo Koskenniemi, Krister Lindén
Frankrig	France	Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur and Institute for Multilingual and Multimedia Information: Joseph Mariani
		Evaluations and Language Resources Distribution Agency: Khalid Choukri
Grækenland	Greece	R.C. "Athena", Inst. for Language and Speech Processing: Stelios Piperidis
Holland	Netherlands	Utrecht Inst. of Linguistics, Utrecht University: Jan Odijk
		Computational Linguistics, University of Groningen: Gertjan van Noord
Irland	Ireland	School of Computing, Dublin City University: Josef van Genabith
Island	Iceland	School of Humanities, University of Iceland: Eiríkur Rögnvaldsson
Italien	Italy	Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli": Nicoletta Calzolari
		Human Language Technology Research Unit, Fondazione Bruno Kessler: Bernardo Magnini
Kroatien	Croatia	Inst. of Linguistics, Faculty of Humanities and Social Science, University of Zagreb: Marko Tadić
Letland	Latvia	Tilde: Andrejs Vasiļjevs
		Inst. of Mathematics and Computer Science, University of Latvia: Inguna Skadiņa

Litauen	Lithuania	Inst. of the Lithuanian Language: Jolanta Zabarskaitė
Luxemborg	Luxembourg	Arax Ltd.: Vartkes Goetcherian
Malta	Malta	Dept. Intelligent Computer Systems, University of Malta: Mike Rosner
Norge	Norway	Dept. of Linguistic, Literary and Aesthetic Studies, University of Bergen: Koenraad De Smedt
		Dept. of Informatics, Language Technology Group, University of Oslo: Stephan Oepen
Polen	Poland	Inst. of Computer Science, Polish Academy of Sciences: Adam Przepiórkowski, Maciej Ogrodniczuk
		University of Łódź: Barbara Lewandowska-Tomaszczyk, Piotr Pęzik
		Dept. of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz University: Zygmunt Vetulani
Portugal	Portugal	University of Lisbon: António Branco, Amália Mendes
		Spoken Language Systems Laboratory, Inst. for Systems Engineering and Computers: Isabel Trancoso
Rumænien	Romania	Research Inst. for Artificial Intelligence, Romanian Academy of Sciences: Dan Tufiș
		Faculty of Computer Science, University Alexandru Ioan Cuza of Iași: Dan Cristea
Schweiz	Switzerland	Idiap Research Inst.: Hervé Bourlard
Serbien	Serbia	University of Belgrade, Faculty of Mathematics: Duško Vitas, Cvetana Krstev, Ivan Obradović
		Pupin Institute: Sanja Vranes
Slovakiet	Slovakia	Ľudovít Štúr Inst. of Linguistics, Slovak Academy of Sciences: Radovan Garabík
Slovenien	Slovenia	Jožef Stefan Inst.: Marko Grobelnik
Spanien	Spain	Barcelona Media: Toni Badia, Maite Melero
		Inst. Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: Núria Bel
		Aholab Signal Processing Laboratory, University of the Basque Country: Inma Hernaez Rioja
		Center for Language and Speech Technologies and Applications, Universitat Politècnica de Catalunya: Asunción Moreno
		Dept. of Signal Processing and Communications, University of Vigo: Carmen García Mateo
Storbritannien	UK	School of Computer Science, University of Manchester: Sophia Ananiadou
		Inst. for Language, Cognition and Computation, Center for Speech Technology Research, University of Edinburgh: Steve Renals
		Research Inst. of Informatics and Language Processing, University of Wolverhampton: Ruslan Mitkov

Sverige	Sweden	Dept. of Swedish, University of Gothenburg: Lars Borin
Tjekkiet	Czech Republic	Inst. of Formal and Applied Linguistics, Charles University in Prague: Jan Hajič
Tyskland	Germany	Language Technology Lab, DFKI: Hans Uszkoreit, Georg Rehm
		Human Language Technology and Pattern Recognition, RWTH Aachen University: Hermann Ney
		Dept. of Computational Linguistics, Saarland University: Manfred Pinkal
Ungarn	Hungary	Research Inst. for Linguistics, Hungarian Academy of Sciences: Tamás Váradi
		Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics: Géza Németh, Gábor Olaszy
Østrig	Austria	Zentrum für Translationswissenschaft, Universität Wien: Gerhard Budin

Omkring 100 sprogteknologieksperter – repræsentanter for de lande og sprog, der er repræsenteret i META-NET – diskuterede og afsluttede de vigtigste resultater og budskaber af hvidbogsserien på et META-NET-møde i Berlin, Tyskland, oktober 21/22, 2011. – About 100 language technology experts – representatives of the countries and languages represented in META-NET – discussed and finalised the key results and messages of the White Paper Series at a META-NET meeting in Berlin, Germany, on October 21/22, 2011.

META-NET-HVIDBOGSSERIEN

THE META-NET WHITE PAPER SERIES

baskisk	Basque	euskara
bulgarsk	Bulgarian	български
dansk	Danish	dansk
engelsk	English	English
estisk	Estonian	eesti
finsk	Finnish	suomi
fransk	French	français
galicisk	Galician	galego
græsk	Greek	ελληνικά
hollandsk	Dutch	Nederlands
irsk	Irish	Gaeilge
islandsk	Icelandic	íslenska
italiensk	Italian	italiano
katalansk	Catalan	català
kroatisk	Croatian	hrvatski
lettisk	Latvian	latviešu valoda
litauisk	Lithuanian	lietuvių kalba
maltesisk	Maltese	Malti
norsk bokmål	Norwegian Bokmål	bokmål
norsk nynorsk	Norwegian Nynorsk	nynorsk
polsk	Polish	polski
portugisisk	Portuguese	português
rumænsk	Romanian	română
serbisk	Serbian	српски
slovakisk	Slovak	slovenčina
slovensk	Slovene	slovenščina
spansk	Spanish	español
svensk	Swedish	svenska
tjekkisk	Czech	čeština
tysk	German	Deutsch
ungarsk	Hungarian	magyar

In everyday communication, Europe's citizens, business partners and politicians are inevitably confronted with language barriers. Language technology has the potential to overcome these barriers and to provide innovative interfaces to technologies and knowledge. This white paper presents the state of language technology support for the Danish language. It is part of a series that analyses the available language resources and technologies for 30 European languages. The analysis was carried out by META-NET, a Network of Excellence funded by the European Commission. META-NET consists of 54 research centres in 33 countries, who cooperate with stakeholders from economy, government agencies, research organisations, non-governmental organisations, language communities and European universities. META-NET's vision is high-quality language technology for all European languages.

Når vi kommunikerer i hverdagen, bliver Europas borgere, forretningspartnere og politikere helt uundgåeligt konfronteret med sprogbarrierer. Sprogteknologien har potentiale til at nedbryde disse barrierer og levere helt nye grænseflader til teknologi og viden. Denne hvidbog præsenterer status for sprogteknologisk støtte til det danske sprog. Den er en del af en serie, som analyserer de sprogresurser og -teknologier der er tilgængelige for 30 EU-sprog. Analysen blev udført af META-NET, et Network of Excellence som er finansieret af EU-Kommissionen. META-NET består af 54 forskningscentre i 33 lande der samarbejder med interessenter fra økonomien, regeringer, forskningsinstitutioner, ikke-statslige organisationer, sprogfællesskaber og europæiske universiteter. META-NETs vision er sprogteknologi af høj kvalitet for alle europæiske sprog.

"The symbiosis of language and technology is in rapid growth today. Being able to use, understand, and communicate with the technology through our local languages imposes high demands on Danish research and development in language technology." – Kim Escherich (IBM Executive Innovation Architect, Sensor Solutions)

"Hvis vi har ambitioner om at bruge det danske sprog i fremtidens teknologiske univers, skal der gøres en indsats nu for at fastholde ekspertise og udbygge den viden vi har. Det viser META-NET rapporten med stor tydelighed. Ellers risikerer vi at kun folk der taler flydende engelsk, vil få glæde af de nye generationer af web-, tele- og robotteknologi der er på vej." – Sabine Kirchmeier-Andersen (Direktør for Dansk Sprognævn)