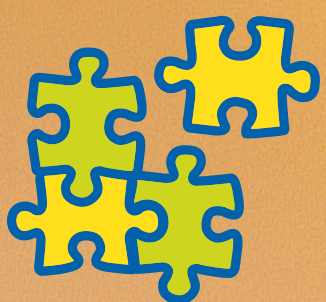


ISBN 87-90708-11-3

Resumé af

Benchmarkundersøgelse af den sprogteknologiske udvikling i Europa

EUROMAP-undersøgelsen
Rose Lockwood og Andrew Joscelyne

København 2003

Indholdsfortegnelse

Information Society Technologies

EUROMAP-projektet er støttet af Europa-Kommissionen
gennem HOPE-kontrakten under IST-programmet.

Dansk oversættelse: Lingtech A/S
ISBN 87-90708-11-3
DESIGN: Signatur TRYK: Trekroner Grafisk A/S

Indledning

Rapporten konkluderer, at de sprogteknologiske aktiviteter i Europa bør synliggøres

Sprogteknologi (HLT) sætter mennesker i stand til at kommunikere med computere og anvende dem på en mere enkel måde og på deres eget sprog, dvs. deltage i informationssamfundet på en helt naturlig måde. Sprogteknologi er særlig vigtig i Europa, da intet andet udviklet økonomisk område udviser en tilsvarende kulturel og sproglig mangfoldighed. Behovet for og evnen til at anvende mange forskellige sprog i hverdagen bliver stadig mere udtalt i erhvervslivet, i fritiden, hos offentlige myndigheder og borgerne i EU samt ansøgerlandene. Faktisk er evnen til at mestre flere forskellige sprog blevet en erhvervsmæssig nødvendighed.

EUROMAP-projektet har undersøgt den nyeste udvikling inden for sprogteknologisk forskning og anvendelse i Europa samt baggrunden for den nuværende situation i hvert enkelt land. På baggrund af dataindsamlinger fra forskellige forskningscentre, leverandører, nationale forskningstiltag og markedsanalyser er de europæiske lande blevet sammenlignet i en benchmarkundersøgelse. Undersøgelsen viser fx, at den omfattende og vedvarende investering, som de tyske, britiske og hollandske myndigheder har foretaget, har kunnet betale sig – disse lande er førende inden for sprogteknologi i Europa. Situationen i andre lande beskrives, ligesom der gives forslag til den fremtidige udvikling.

Rapporten konkluderer, at de sprogteknologiske aktiviteter i Europa bør **synliggøres**, og at denne synlighed bør knyttes tæt til Det Europæiske Forskningsrum (ERA). Målet bør være at tilvejebringe et antal robuste, stabile og flersprogede sprogteknologiske moduler, der kan indlejres i de nye IST-baserede anvendelser. Der bør oprettes et **Sprogteknologisk Agentur,** der kan overvåge og styre overgangen fra nationale sprogteknologiske tiltag til et egentligt europæisk teknologiniveau baseret på sproglig lighed. Der bør etableres **infrastrukturfonde** til anskaffelse af sprogressourcer og grundlæggende sprogteknologiske moduler til alle sprog, og disse bør overvåges af Det Sprogteknologiske Agentur.

Det har været utroligt interessant at arbejde med disse emner og at se, hvordan de enkelte europæiske lande har taklet informationssamfundets sproglige udfordringer, og hvad konsekvenserne er. Vi håber, at de fremlagte data samt analysen og anbefalingerne vil blive anvendt af beslutningstagere på nationalt og europæisk plan.

Endelig vil vi gerne takke for den støtte, som Europa-Kommissionen har ydet til projektet, men ønsker samtidig at understrege, at denne rapport er resultatet af EUROMAP-projektet, og at de holdninger, der fremgår af rapporten, ikke nødvendigvis er udtryk for Europa-Kommissionens holdning.

Bente Maegaard
Koordinator for EUROMAP
Language Technologies

Hvad er sprogteknologi?

Kombinationen af tale og natursprogsbehandling giver en stærk teknologi til menneske-maskine interaktion

Termen sprogteknologi (HLT) dækker den type softwarekomponenter, -værktøjer, -teknikker og anvendelser, der behandler naturligt, menneskeligt sprog. Sprogteknologi kan opdeles i to overordnede områder: talebehandling og natursprogsbehandling. Taleteknologi simulerer den menneskelige evne til at høre og anvende talt sprog. Teknologien bag natursprogsbehandling skaber en model af menneskets evne til at forstå og behandle indholdet i sproget – dvs. at forstå og omforme skrevet tekst. Automatisk oversættelse er et almindeligt eksempel på anvendelse af natursprogsbehandling. Kombinationen af tale og natursprogsbehandling giver et stærk teknologisk grundlag for forbedring af interaktionen mellem mennesker og maskiner samt mellem mennesker der *bruger* maskiner.

I løbet af de sidste 10 år har sprogteknologien udviklet sig fra at være et specialiseret og teoretisk forskningsemne til at blive en kerneteknologi i informationssamfundet. Målt ud fra samtidens normer – i omgivelser hvor de teknologiske innovationers cyklusser er blevet reduceret til måneder i stedet for år – kan den sprogteknologiske udvikling forekomme langsom. Dette indtryk er misvisende. Man har udført grundlæggende forskning på området i mere end 50 år. Betingelserne for anvendelse af sprogteknologi begyndte at vise sig i 1990'erne, og udbredelsen af sprogteknologi er siden vokset støt.

Sprogteknologien trives bedst under forhold, der støtter informationsrevolutionen – en høj grad af relativt prisbillig regnekraft og stort set universel konnektivitet. Igennem årtier var sprogteknologiens kompleksitet og behov for regnekraft en hindring for udviklingen; en hindring som også begrænsede forskningens omfang. Den teknologiske udvikling har imidlertid både skabt den infrastruktur, der understøtter sprogteknologien, og behovet for netop den type produkter og serviceydelser, som sprogteknologien kan understøtte.

Hvorfor er sprogteknologi vigtig for Europa?

Sprogteknologien spiller en afgørende rolle i EU på grund af den kulturelle og sproglige mangfoldighed

Sprogteknologi spiller en afgørende rolle i den Europæiske Union som følge af de usædvanlige kulturelle betingelser, der gælder i Europa. Der findes intet andet udviklet økonomisk område med tilsvarende kulturel og sproglig mangfoldighed som Europa. Antallet af officielle sprog i EU vil vokse fra de nuværende 11 til mere end 20, når den næste række ansøgerlande optages i EU. Der anvendes mange andre sprog i EU, herunder regionale sprog (såsom catalansk og baskisk i Spanien), uofficielle nationale sprog (såsom walisisk i Storbritannien) samt immigrantsprog (såsom urdu i Storbritannien, maghrebi/arabisk i Frankrig og tyrkisk i Tyskland).

Evnen og viljen til at anvende forskellige sprog i hverdagen bliver stadig mere udtalt i erhvervslivet, i fritiden, hos myndigheder og blandt borgere i EU. Dette afspejler europæernes stræben efter integration sideløbende med deres dybe respekt for det lokale. Europæerne er i stigende grad blevet bevidste om, at en aktiv indsats til fordel for sproglig mangfoldighed beskytter borgernes ret til at bevare deres lokale sprog – ikke på bekostning af andre sprog, men som del af de fælles værdier i EU.

Informationsrevolutionen skaber derfor specielle udfordringer for EU. I et stadigt mere kompakt informationsmiljø – for indbyggere og forbrugere, myndigheder og erhvervsliv – bliver sproglig gennemsigtighed afgørende. Hvis alle borgere i det udvidede EU skal kunne deltage fuldt ud i informationssamfundet, skal dette samfunds produkter og serviceydelser være tilgængelige på alle sprog. Hvis Europa

skal kunne fungere optimalt som et enhedsmarked, og hvis målene i eEurope-visionen skal nås, skal produkter og serviceydelser leveres tværsprogligt, så det bliver lige så let at krydse sprog som grænser.

Mange af produkterne og serviceydelserne i informationssamfundet vil bygge på grundlæggende sprogteknologiske komponenter. Sprogteknologiens betydning strækker sig ud over det umiddelbart indlysende; den trænger ind i de dybeste lag af internettet, hvor evnen til at behandle sprogets bestanddele – gennem indkodning af viden og intelligens i informationsinfrastrukturen – vil være grundlaget for den næste teknologigeneration.

EU har allerede forskningsmæssigt etableret en førende position på det sprogteknologiske område. Selve vanskeligheden ved at udvikle sprogteknologi for mange sprog giver faktisk europæiske forskere og teknologiudviklere en naturlig fordel inden for en af de vigtigste teknologier til den næste generation af informations- og kommunikationsteknologi (IKT). Som følge heraf er engagementet i den sprogteknologiske fremtid i Europa måske vigtigst i relation til den styrke, den vil give den europæiske IKT-sektor. En undersøgelse af Booz-Allen & Hamilton, *The Competitiveness of Europe's ICT Markets: The Crisis Amid the Growth* (fremlagt på ministerkonferencen i marts 2000), dokumenterer de konkurrencemæssige udfordringer, som Europa globalt står overfor i mange af IKT-sektorens nøglesegmenter, herunder software. En nylig undersøgelse fra The Conference Board, *Productivity, ICT and Service Industries: Europe and the United States,* vurderer IKT-sektorens betydning for produktiviteten. Europas produktivitetskløft inden for IKT-baserede brancher, herunder især servicesektoren, er markant. EU står derfor over for konkurrencemæssige udfordringer inden for IKT – både i relation til udbud og efterspørgsel.

Sprogteknologien er en 'lille' teknologi rent markedsmæssigt set, men dens potentielle betydning for tilgængelighed, innovation og integration er omfattende, og dens afgørende rolle i realiseringen af eEurope's potentiale er enestående. EUROMAP-undersøgelsen kortlægger, hvordan sprogteknologisk forskning passer ind i den større sammenhæng af avancerede teknologier i relation til eEurope's fremtid. Den fremhæver sprogteknologiens rolle i de nye paradigmer til næstegenerations-IKT, og foreslår, hvordan sprogteknologi kan integreres i Det Europæiske Forskningsrum.

Baggrunden for EUROMAP-undersøgelsen

Undersøgelsen foreslår en dagsorden for næste generation af IST-forskningen

EUROMAP Language Technologies er et initiativ, som er støttet af Europa-Kommissionen, og som har til formål at udbrede kendskabet til og fremme anvendelsen af sprogteknologi i Europa. Siden 1996 har projektet fungeret som en central enhed, der leverer oplysninger til alle aktører på det sprogteknologiske område, lige fra forskere og udviklere til leverandører og brugere. EUROMAP støtter relevante nationale sprogteknologiske miljøer gennem nationale knudepunkter, der yder direkte service til miljøerne. Gennem paneuropæiske initiativer som webstedet www.hltcentral.org og LangTech-konferencerne binder EUROMAP de forskellige interessenter i den europæiske sprogteknologi sammen.

EUROMAP-undersøgelsen samler de erfaringer, der er indsamlet gennem mere end fem år. Den gør udstrakt brug af ressourcer og viden fra mange eksperter og udøvere fra alle medlemslandene samt fra en række af de nye ansøgerlande. EUROMAP-netværket har dokumenteret det sprogteknologiske forskningsmiljø, og har kortlagt den stadigt stigende mængde af aktive virksomheder på det sprogteknologiske område. Gennem møder og feltarbejde har netværket dokumenteret det fremspirende marked for sprogteknologi. I samråd med førende sprogteknologiske forskere og i samarbejde med forskningsnetværket

ELSNET har EUROMAP tilvejebragt et bredt kendskab til områdets muligheder og udfordringer.

Undersøgelsen bygger på en rapport fra 1998, der gav det første paneuropæiske syn på denne nye teknologi i begyndelsen af det femte rammeprogram for forskning og teknologisk udvikling. EUROMAP har videreført undersøgelsen af den sprogteknologiske udvikling og har udviklet en foreløbig benchmarkingmetode til måling af fremskridtet på området.

Resultaterne af den aktuelle undersøgelse lægger op til den sprogteknologiske fremtid i Europa og identificerer politikker og foranstaltninger, der har givet positive resultater. Undersøgelsen giver forslag til, hvordan de positive resultater fra tidligere forskningstiltag kan udnyttes i næste generations IST-forskning i Europa.

De nyeste teknologier

Der er blevet identificeret omkring 300 europæiske virksomheder

EUROMAP-projektet har identificeret ca. 300 europæiske virksomheder, der tilbyder sprogteknologiske produkter og serviceydelser. De fleste af disse virksomheder ligger i de nuværende medlemslande, men der findes også en lille, men voksende gruppe af virksomheder i ansøgerlandene i Østeuropa. Mange af disse virksomheder tilbyder kombinationer af sprogteknologiske elementer og funktioner lige fra basiskomponenter til avancerede løsninger.

Komponenter og ressourcer

Al sprogteknologi hviler på kernekomponenter, der digitalt modellerer den menneskelige sprogbehandling. Disse komponenter kan være baseret på lingvistiske regler (såsom grammatik), på statistisk analyse (fx af sandsynligheden for, at en tekst eller ytring har en bestemt betydning) eller på en kombination af de to. Derudover har al sprogteknologi brug for en kilde af sproglige data som reference såsom et leksikon (en ordbog kodet med grammatiske oplysninger) eller et 'korpus', som fungerer som en omfattende database af sprogligt råmateriale. Eksistensen og tilgængeligheden af disse grundlæggende komponenter er udgangspunktet for udviklingen af sprogteknologi. EUROMAP har identificeret omkring 120 europæiske virksomheder, der tilbyder grundlæggende sprogteknologiske komponenter og sprogressourcer på rundt regnet 25 sprog.

Videnbehandling

Sprogteknologi kan integreres i det, man traditionelt kalder 'videnapplikationer' – dvs. produkter og serviceydelser, der anvender en eller anden form for sproglig intelligens. Søgemaskiner anvender sprogteknologi til at give en bedre matchning af søgeord, fx ved at finde forskellige morfologiske former af et ord eller endda synonymer. Mere avancerede applikationer såsom *knowledge mining* bruger kombinationer af sprogteknologiske værktøjer til at analysere data og rapportere om indholdet af tekst- eller dokumentlagre. Store virksomheder udvikler i stigende grad taksonomier (dvs. træstrukturer over sproglige begreber) i forbindelse med organiseringen af deres viden. EUROMAP har identificeret omkring 120 europæiske virksomheder, der tilbyder sprogteknologiske produkter og serviceydelser til videnbehandling på rundt regnet 25 sprog.

Interface og interaktion

Interface- og interaktionsteknologier er oftest talebaserede. Kendte eksempler på anvendelse af taleteknologi er telefonbaserede talegenkendelsessystemer, der fjerner behovet for et tastatur, og som normalt anvendes i call-centre og i telefonbaserede transaktionssystemer. Talegenkendelse anvendes også i dikteringssystemer, der overflødiggør tastaturet. Talesyntesesystemer (Text-To-Speech) bliver mere og mere almindelige i applikationer såsom 'aflytning af e-mail' efter at have været anvendt som støtte til blinde i lang tid. Blandt mere avancerede anvendelser kan nævnes stemmeautentificering. Talesystemer er kommet videre ud end til traditionelle platforme og er nu indbygget i almindelige forbrugsvarer og fx i biler. EUROMAP har identificeret omkring 130 europæiske virksomheder, der tilbyder sprogteknologiske interface- og interaktionsbaserede produkter og serviceydelser på ca. 25 sprog.

Tværsproglighed

Maskinoversættelse (MT) var det første eksempel på anvendelse af natursprogsbehandling og er teknisk set stadig et af de mest udfordrende. Ikke desto mindre har man udviklet produkter til mange forskellige sprogpar, og der tilbydes gratis rå-oversættelse mange steder på internettet. Foruden MT kan tværsproglige anvendelser også være både viden- og interfacebaserede applikationer. En tværsproglig søgemaskine kan oversætte en term med henblik på at søge i databaser på forskellige sprog, hente den 'fremmede' tekst og levere en rå-oversættelse eller endda et resumé på den oprindelige søgeterms sprog. Der findes mange prototyper på tværsprogli-ge taleapplikationer, fx systemer til telefonreservation, der gør det muligt for mennesker, der taler forskellige sprog, at kommunikere. EUROMAP har identificeret omkring 60 europæiske virksomheder, der tilbyder tværsproglige produkter og serviceydelser på 25 sprog.

Det flersprogede semantiske net

Den næste generation af internettet vil indeholde sproglige kernedata fra starten. Initiativet *Semantic Web* vil kode alle typer digitalt indhold med semantisk viden og anvende denne viden til at muliggøre en mere forudsigelig interaktion mellem forskellige systemer og serviceydelser. Agentteknologi forsynet med semantisk viden om en bruger kan interagere med praktisk talt alle elektroniske systemer, der har den samme viden om verden. Denne viden samles i 'ontologier' – strukturerede begreber med fastlagte relationer, der repræsenterer viden om verden. Den europæiske sprogteknologi bør være i stand til at bevare sin position som tankeleder inden for udviklingen af det flersprogede semantiske net og sikre, at alle europæiske sprogsamfund deltager i udviklingen af semantiske ressourcer, og at de serviceydelser, der bruger dem, er tilgængelige på alle de europæiske sprog.

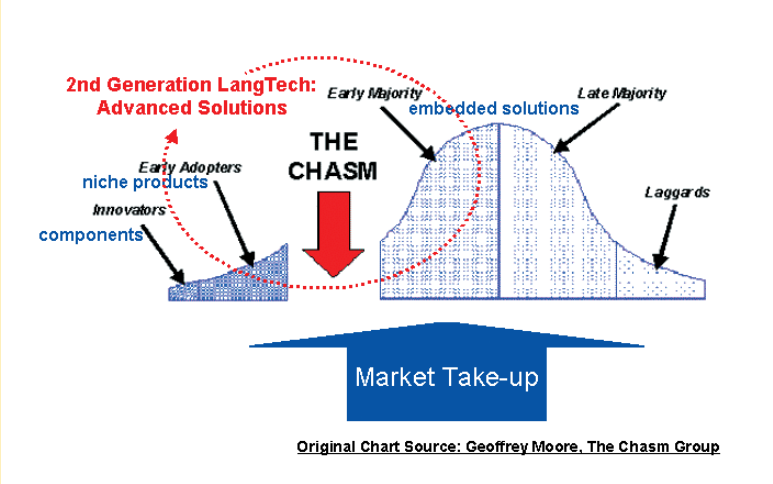
Visionære teknologier

Sprogteknologi vil være en integreret nøgleteknologi i takt med, at den næste generation af IKT-produkter bliver udviklet. Visionært arbejde inden for IT er fokuseret på 'intelligente omgivelser' i en 'allestedsnærværende IT-infrastruktur', hvor viden er indbygget i genstande i hele samfundet, der på naturlige måder kan reagere på mennesker. Forskningen inden for 'e-sense' vil skabe modeller af den måde, som alle menneskelige sanser berøres på under kommunikation. Grænsen mellem 'viden-håndtering' og 'interface' bliver derfor udvasket, og maskinerne vil ophøre med at dominere den elektroniske kommunikation. Maskiner vil interagere med mennesker på en mere menneskelig måde, og mennesker vil interagere med mennesker, der anvender maskiner, der er lettere at gennemskue. Nye IKT-paradigmer vil behandle oplysninger om menneskelige erfaringer gennem alle menneskets sanser i det mest naturlige kodningssystem, dvs. sprog, som skaber det, der er blevet kaldt 'det perceptuelt bevidste tværsproglige menneskelige interface'.

Det sprogteknologiske marked

På sin vej til markedet har sprogteknologien nu gennemgået én komplet 'kløftoverskridende' cyklus, der er gået fra 'innovators' (innovatorer) over 'early adopters' (tidlige brugere) til udbredt/'early majority' (tidligt flertal).

På dette marked for førstegenerations-sprogteknologi oplevede *innovatorer* enten meget vanskelige krav (fx Europa-Kommissionens og det amerikanske forsvars anvendelse af MT) eller eksperimenterede med komponenter på innovative måder (fx Reuters' tidlige anvendelse af sprogteknologi til fremfinding af information). *Tidlige brugere* udnyttede visse produkters stigende grad af modning til helt bestemte formål, fx Xerox' og Caterpillars anvendelse af MT til tekniske dokumenter samt indførelsen af talegenkendelse i medicinske transskriptionssystemer. Sprogteknologien er nu langt mere udbredt og nået til *tidligt flertal* med hensyn til indlejringsmuligheder. Telefonbaseret talegenkendelse anvendes i stort omfang; de fleste store søgemaskiner indeholder



Figur 1: Kløften

sprogteknologiske komponenter; millioner af internetsider oversættes dagligt automatisk ved hjælp af MT.

Denne udvikling afspejles i markedsudgifterne til sprogteknologi. Datamonitor vurderer det globale marked for taleteknologi til lige under 1 milliard € i 2003. IDC sætter det nuværende marked for naturprogsbehandling til omkring 400 millioner €. Det kombinerede marked for taleteknologi og naturprogsbehandling skønnes til mere end 2 milliarder € i 2005. Selv om disse muligheder i sig selv ser gode ud, afspejler de ikke den multiplikatoreffekt, der er forbundet med indlejret sprogteknologi. Den værdi, som sprogteknologien tilfører produkter og serviceydelser, skaber markedsværdier, der er mange gange større end kerneteknologien selv.

Det næste markedsniveau – udnyttelse af sproglig viden i mere komplekse og avancerede anvendelser – vil igangsætte en ny udviklingscyklus. I andengenerations-sprogteknologi vil innovatorer eksperimentere, mens det brede marked vil afvente gennemtestede løsninger. Det er ikke sandsynligt, at andengenerations-sprogteknologi vil afføde mange enkeltstående 'rendyrkede sprogteknologiske' produkter. I stedet vil sprogteknologien blive indlejret i andre anvendelser og skabe innovative elementer eller bedre resultater, der vil gøre en forskel. Den sandsynlige udformning af fremtidens sprogteknologiske marked bør styre den fremtidige forskning på området, der skal udføres i forbindelse med avanceret forskning inden for følge- eller værtsteknologi.

Den sprogteknologiske forskningsbase

Indtil nu har det sprogteknologiske miljø i Europa formået at bevare sin konkurrencedygtighed i forhold til stærke forskningsinitiativer i USA og Japan samt til det stigende niveau af F&U i andre dele af Asien (især oversættelsesteknologi i Kina, Korea og Indien). Sprogteknologien er nu et af de få områder inden for software-forskning, hvor europæisk forskning er i verdensklasse.

Vigtigheden af EU-støtte

Sprogteknologisk forskning har i mange år fået støtte gennem EU's rammeprogrammer, og tids- og strukturmæssigt har støtten passet godt til behovet på området. Indtil midten af 1990'erne havde forskningsprogrammerne et 'teknologisk drevet' fokus, der var meget effektivt for sprogteknologien, da markedsforholdene endnu ikke var gunstige. Støtten til sprogteknologi i 4. Rammeprogram og den sprogteknologiske indsats i 5. Rammeprogram har været mere markedsorienteret og har nøje fulgt udviklingen i de markedsrelaterede muligheder.

EU-støtten har spillet en afgørende rolle i etableringen af et sammenhængende forskningsmiljø i Europa. Erhvervslivets forskningsstøtte på det sprogteknologiske område har været begrænset og mest udtalt i relation til taleteknologi. Den offentlige støtte til den sprogteknologiske forskning har varieret meget på nationalt plan, og med enkelte nævneværdige undtagelser har den været noget svingende. EU-støtte til forskning i maskinoversættelse skabte et netværk af datalingvister i hele EU og fostrede samtidig en række forskningsinitiativer inden for forskellige sprog samt et etableret akademisk udgangspunkt for eksperter i MT. I tillæg hertil har tendensen til at støtte en lang række mindre projekter (sammenlignet med tendensen i USA og Japan) haft det resultat, at den tekniske basis er blevet udvidet i hele EU. Samtidig har opbygningen af rammeprogramsprojekter, der kræver samarbejde på tværs af grænser, skabt en ægte paneuropæisk forskningsbase.

EU-støtten har desuden haft stor betydning for teknologioverførslen på det sprogteknologiske område. Antallet af aktive leverandører på det sprogteknologiske marked er steget eksponentielt i løbet af de sidste 10 år, fra mindre end 30 virksomheder i 1993 til 10 gange så mange i 2003. Næsten alle disse europæiske leverandører har rødder i EU-støttede programmer enten via teknologi, der udspringer direkte af projekter, eller gennem den tekniske viden hos de forskere, der har deltaget i projekterne.

Benchmarkundersøgelse af sprogteknologien

Denne EUROMAP-undersøgelse er baseret på en benchmarkundersøgelse af resultaterne af den sprogteknologiske forskning i Europa og de muligheder den åbner. Undersøgelsen sammenlignede medlemslandene og oprettede indekser for to overordnede mål: mulighederne for at anvende sprogteknologi ('Mulighedsindekset') og succesen for den sprogteknologiske forskning og teknologioverførsel ('Det Sprogteknologiske Benchmark').

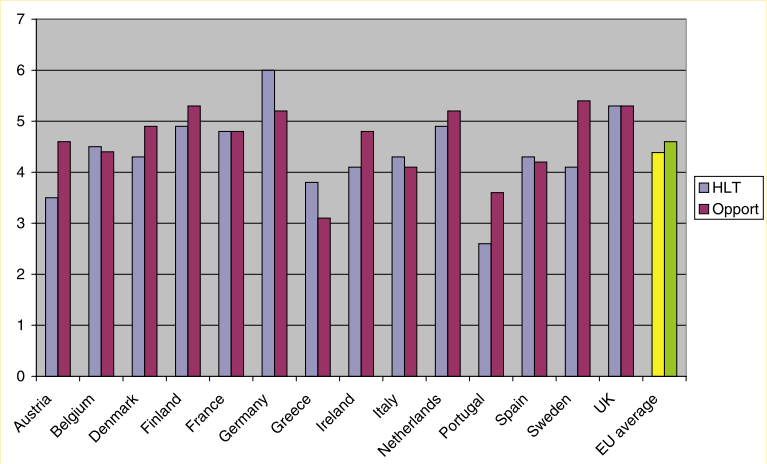
Mulighedsindekset blev baseret på uafhængig forskning, der måler forhold som det generelle miljø for forskningsinnovationen; udbudsfaktorer som hvor let det er at oprette virksomheder, adgang til vigtige kanaler (defineret i EUROMAP-undersøgelsen) for sprogteknologi og evnen til at implementere innovation; efterspørgselsfaktorer som konkurrencedygtighed, IKT-infrastruktur og evnen til at integrere innovation. Faktorerne blev dernæst vægtet for at afspejle en vurdering af deres relative vigtighed som potentielle succesfaktorer for sprogteknologi.

Det Sprogteknologiske Benchmark var baseret på EUROMAP-relateret skrivebordsforskning og feltarbejde. Faktorerne omfattede dybde og bredde i sprogteknologisk forskning (i både tale- og natursprogsbehandling), støtte fra både den offentlige sektor og erhvervslivet samt den sproglige dækning inden for forskning og produkter (der både tager hensyn til antallet og valget af sprog samt dækningen af mindre udbredte sprog og minoritetssprog). Målingen af den forskningsmæssige dybde tog hensyn til, om de grundlæggende sprogteknologiske komponenter er fuldt udviklede, og undersøgte, i hvilket omfang mere avancerede anvendelser er genstand for F&U-projekter.

Mulighedsindekset blev sammenholdt med Det Sprogteknologiske Benchmark for at nå frem til det såkaldte 'sprogteknologiske scorekort' – en sum af målingerne der viser sammenhængen mellem de to indekser. Der var en tydelig sammenhæng mellem Mulighedsindekset og Det Sprogteknologiske Benchmark. Generelt har landene med det mest gunstige forretningsmiljø og den bedst udviklede infrastruktur også mest succes med sprogteknologisk forskning.

Medlemslandenes scorekort

De **'førende'** lande er Tyskland, Holland og Storbritannien. Disse lande har oplevet et stærkt nationalt engagement i sprogteknologisk forskning. I Tyskland, der står øverst på Det Sprogteknologiske Benchmark, har både den offentlige og den private sektor siden SPICOS-projektet i 1985 gennemført en konsekvent, effektiv investering i sprogteknologi. De førende lande anses for at være 'markedsklar' til avanceret sprogteknologisk forskning.



Figur 2: Sammenligning - sprogteknologi/mulighed

Gruppen med **'stort potentiale'**, der står på gennemsnittet eller lidt under på Mulighedsindekset, men på gennemsnittet på Det Sprogteknologiske Benchmark, omfatter Frankrig, Belgien og Spanien. Frankrig ligger som de førende lande på Det Sprogteknologiske Benchmark, men ligger markant lavere på Mulighedsindekset. Disse lande har veludviklede forskningsmiljøer og en markant dybde i den sprogteknologiske forskning, og de står derfor stærkt i forhold til at udnytte sprogteknologien, i takt med at mulighedsfaktorerne forbedres.

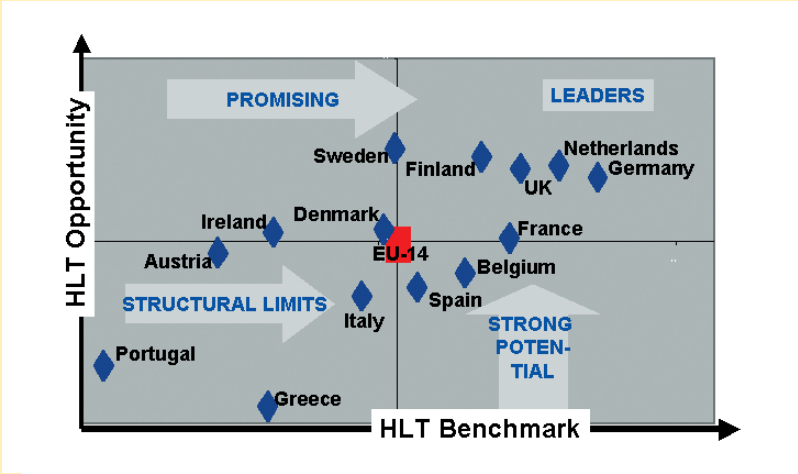
En tredje gruppe består af **'lovende'** lande, hvor Irland og Danmark ligger omkring gennemsnittet i henhold til begge indekser lige efter Sverige, der står højest på Mulighedsindekset, og Finland der ligger over gennemsnittet på Det Sprogteknologiske Benchmark. Mens disse lande mere eller mindre befinder sig på EU-gennemsnittet med sammenlignelige resultater, når det gælder førstegenerations-F&U og teknologioverførsel, skal de både øge investeringen i den sprogteknologiske forskning og forbedre teknologioverførslen, hvis de ønsker næstegenerationsstandarder.

Endelig er der en gruppe på fire lande (Grækenland, Italien, Portugal og Østrig), der har nået den **'strukturelle grænse'** for deres sprogteknologiske markedssituation, og som har behov for nytænkning for at indhente de førende lande. De ligger under gennemsnittet på begge indekser, men har dog forskellige profiler. Både Grækenland og Portugal står lavt på Mulighedsindekset, men Grækenland står højt på Det Sprogteknologiske Benchmark på grund af sin solide F&U-basis. Begge lande har muligvis behov for at kigge ud over deres landegrænser for at finde anvendelsesmuligheder for deres sprogteknologi, og de vil have gavn af et udvidet EU-samarbejde. Især Portugal kunne forbedre sine forskningsmæssige muligheder gennem et mere omfattende samarbejde på tværs af landegrænser. Italien har en lidt stærkere F&U-basis end de fleste i gruppen, men står sammen med Østrig på gennemsnittet på Mulighedsindekset. Østrig har fordel af at have samme sprog som det førende land, Tyskland, men netop denne kendsgerning kan virke motivationshæmmende for udvidelsen af Østrigs egne sprogteknologiske aktiviteter.

Generelle konklusioner

Sprogteknologisk forskning og udvikling

Af indlysende grunde har den sprogteknologiske forskning historisk set udviklet sig med en national hældning i retning af F&U inden for de nationale forskningsmiljøers sprog. Mens dette var afgørende for den tidligste sprogteknologiske forskning, ser man i stadig stigende grad et flersproget fokus, især i de mere succesrige forskningsafdelinger og -centre.



Figur 3: Det sprogteknologiske scorekort

Dette er en sund udvikling som bør kunne afhjælpe uhensigtsmæssige tendenser til 'ejerrettighederne' over sprogteknologien for et bestemt sprog. I takt med at det sprogteknologiske forskningsmiljø i Europa bliver mere integreret, spreder sprogekspertisen sig i hele EU, mens den naturligvis bevarer sine rødder i de nationale sprogsamfund. Det er vigtigt, at den sprogteknologiske og lingvistiske ekspertise frit kan spredes og integreres i hele EU's forskningsmiljø.

Sprogteknologisk F&U er en lang og kompleks proces, der kræver betydelig offentlig støtte. Den nødvendige udvikling af uddannelse, ressourcer, værktøjer og teknologi kan ikke sikres af markedskræfterne alene. Europas succes på det sprogteknologiske område har været baseret på offentlig støtte til universiteter, nationale forskningsinstitutter og særlige projekter. Det er usandsynligt, at dette område kan udvikles effektivt – især med henblik på at bringe alle sprog op på det samme avancerede niveau og inddrage de nye EU-sprog – uden fortsat omfattende offentlig støtte.

Konsekvent og langsigtet national støtte til den sprogteknologiske forskning har givet rigt afkast og har i stor stil bidraget til den stærke nationale forskningsbase i Tyskland, Frankrig og Storbritannien. Det er imidlertid usandsynligt, at alle medlemslande, især i det udvidede EU, vil være i stand til at støtte projekter på samme niveau som de mere teknologisk avancerede medlemslande (herunder Holland samt Finland og andre nordiske lande). Som følge heraf må EU-støtten struktureres på en sådan måde, at den tager højde for variationerne i omfanget af den nationale støtte.

Mens de nationale programmer i EU's nøglestater har været afgørende for opbygningen af sprogteknologiske kernekompetencer (som et supplement til EU-programmer), så har de langt fra været ensartede. Den nationale tilgang til sprogteknologisk forskning har afspejlet lokale prioriteter og strukturer. I Tyskland, for eksempel, har store omfattende programmer med et samlet fokus (fx Verbmobil) forbundet industri og forskningsmiljøer på en meget struktureret måde. I Frankrig var forskningen tæt knyttet til de (daværende) nationale centre (fx France Telecom). I Storbritannien skabte et relativt tidligt tale- og sprogteknologiprogram et stærkt netværk af nationale forskere, der muliggjorde teknologioverførsel på det tidspunkt, hvor teleindustrien blev dereguleret. Dette tyder på, at den ægte 'europæiske' tilgang til den fremtidige sprogteknologiske forskning skal være smidig og justerbar til medlemslandenes forskellige forhold.

Mens de forskningsaktiviteter, der modtager støtte via de sprogteknologiske tiltag i EU's rammeprogrammer for forskning, er relativt synlige både inden

og uden for forskningsmiljøet, er billedet alligevel ufuldstændigt. For eksempel er der indtil videre ikke noget sammenhængende, klart overblik over den omfattende sprogrelaterede F&U på andre IST-områder (for eksempel inden for digitale biblioteker på ERCIM eller bedrageriforebyggelse på JRC) eller over de vigtige, om end strukturelt ret forskellige, nationale programmer (for eksempel i Frankrig, Italien, Litauen og Estland) eller den kommercielt støttede forskning (især på området for taleteknologi til biler i Telematics Valley i Sverige eller systemer til kontrolleret sprog i den aeronautiske og køretøjsrelaterede dokumentationssektor). Denne mangel på sammenhængende overblik vil sandsynligvis forværres, når det 6. Rammeprogram gradvist implementerer en politik der indlejrer tidligere selvstændige sprogteknologiske aktiviteter i den generelle IST-baserede forskning. Uden en klar plan til at identificere mønstrene i den løbende F&U, er der ikke blot risiko for, at der opstår et unødvendigt overlap mellem indsatserne, men også for, at investeringsselskaberne får vanskeligere ved at bidrage effektivt til teknologioverførslen.

Status for forskningen varierer meget inden for Europas sprog. Enkelte sprog (engelsk, tysk og fransk) er godt understøttet, hvilket bidrager til at der arbejdes med mere avancerede forskningsemner og anvendelser. Nogle af de mindre 'udbredte' sprog (målt i antallet af talere) har endnu ikke det nødvendige til førstegenerations-anvendelser. Dette betyder, at der er behov for yderligere offentlige investeringer for at bringe alle sprog op på et relativt ensartet niveau – et baseline-niveau – eftersom dette er en altafgørende forudsætning for den fremtidige udvikling af en avanceret informationssamfundsteknologi, der kan tilbyde alle europæiske borgere samme støtte.

Samtidig har den sprogteknologiske forskning igennem flere år bevæget sig i retning af 'teknikker' og væk fra teoretisk forskning. Selv åbenlyse teoretiske skift (fx statistisk og datadrevet natursprogsbehandling, i modsætning til regelbaseret) minder mere om naturlige hybrider end ægte paradigmeskift. Selv om dette er en naturlig cyklus, vil området sandsynligvis have gavn af omfattende nytænkning af de grundlæggende forudsætninger. At gøre plads til grundlæggende teoretisk arbejde på dette niveau (i modsætning til opnåelse af den grundlæggende baseline-F&U for de 'nye' EU-sprog) bør derfor være på dagsordenen for den næste generation af sprogteknologisk forskning. Det forekommer meget sandsynligt, at der vil opstå nye teoretiske tilgange gennem krydsfertilisation med andre tekniske IT-discipliner, hvilket fremhæver fordelene ved at medtage avancerede sprogteknologiske komponenter i den primære forskningsdagsorden for 6. Rammeprogram.

For at sikre at Europa ikke udvikler en sprogteknologisk kultur med to hastigheder, hvor den ene velfunderede halvdel af den sprogteknologiske F&U-dagsorden har fokus på avancerede systemer inden for et eller flere 'strategiske' sprog og den anden halvdel har fokus på at sikre grundlæggende dækning for de mindre udbredte eller mindre 'strategiske' sprog, er det imidlertid altafgørende, at der oprettes en eller anden form for selvstyrende 'Sprogteknologisk Agentur'. Dets opgave bør være at opretholde en passende grad af sprogteknologiske selvstændighed (især på det kritiske område for grundlæggende sprogressourcer) uafhængigt af sprogteknologiens ultimative teknologiske skæbne som indlejret komponent i informationssamfundets infrastruktur.

Teknologioverførsel til markedet

Indtil videre har der ikke været en direkte forbindelse imellem omfanget af forskningsindsatsen i et bestemt sprogsamfund og hvor effektiv teknologioverførslen har været. Der er naturligvis en klar forskel imellem eksemplerne på vellykket teknologioverførsel for de udbredte sprog (især engelsk, tysk og fransk) og eksemplerne for de mindre udbredte sprog, hvilket tydeligvis skyldes det kommercielle potentiale på de store markeder, hvor man taler de udbredte sprog. Der findes imidlertid en markant undtagelse til dette nemlig europæisk-spansk, hvor forskningsindsatsen stadig er forholdsvis diffus, hvilket til dels skyldes national støtte til en række 'regionale' sprog, hvoraf alle har officiel status.

Et andet særligt tilfælde er italiensk, der globalt set tales mindre end spansk, men som er meget udbredt i Europa, og alligevel har en relativt svag teknologioverførsel, hvilket uden tvivl skyldes specifikke vilkår i forretningsmiljøet og den tekniske infrastruktur. Italien har en lang og stærk tradition for sprogteknologisk forskning, der går helt tilbage til datalingvistikkens begyndelse i 1950'erne, og det er tydeligt, at landets aktuelle position mere skyldes kommerciel 'timing' end en indre teknologisk svaghed.

I modsætning hertil har det relativt stærke forskningsmiljø i både Finland og Holland, hvor forretningsmiljøet og infrastrukturen er blandt de stærkeste i Europa, ikke desto mindre overført mindre teknologi end man havde kunnet forvente. Konklusionen er, at overførslen af sprogteknologi til markedet er påvirket af tre vægtige faktorer: sprogsamfundets størrelse, forretningsmiljøet og infrastrukturen samt forskningsfokus. Eftersom 'udgifterne' til teknologisering af et sprog i sidste ende er de samme, uanset om sproget tales af 2 millioner eller 200 millioner, bliver det nu tydeligere, at der ikke nødvendigvis er nogen entydig forbindelse imellem den sprogspecifikke forskning, den teknologiske udvikling, teknologioverførslen og geografien.

Et af resultaterne af et ægte tværeuropæisk, i modsætning til nationalt, marked for F&U kunne for eksempel være, at det tilskynder til oprettelsen af centre for bedste praksis inden for sprogteknologisk udvikling, således at det, der viser sig at være de 'bedste' sprogteknologiske arkitekturer, vælges som de optimale udviklingsmiljøer for alle 'nationale' sprog, uanset hvor de pågældende sprog tales eller skrives.

I et ægte interaktivt europæisk informationssamfund er der helt klart behov for sproglig lighed på alle niveauer. Indtil nu har der været en naturlig, og begrænsende, tendens til at udvikle sprogteknologi med henblik på at tilvejebringe informationer *på* det nationale sprog. I praksis har jeg naturligvis behov for adgang på *mit* sprog til indhold og interaktion på *dit* sprog, ligesom du har behov for adgang til *mit* indhold og interaktion på *dit* sprog. Opnåelsen af en sådan sproglig lighed på teknologisk niveau kan udelukkende opnås gennem etablering af en omfattende flersproget infrastruktur, der i langt større udstrækning end i dag kunne gøre sprogbehandlingen uafhængig af F&U-geografier. Opnåelsen af en sådan lighed ville være et vigtigt punkt på dagsordenen for et eventuelt europæisk 'Sprogteknologisk Agentur'.

Politisk prioritering

Det er almindeligt anerkendt, at indsatsen inden for innovation og F&U skal styrkes, hvis Europa skal blive en førende viden- og teknologibaseret økonomi, der opfylder eEurope's formål. Sprogteknologien bør fortsat fremmes som en væsentlig teknologisk fordel for Europa.

Sprogteknologiens markedsmuligheder korrelerer i vid udstrækning med de 'omgivende' muligheder såsom den gældende IKT-infrastruktur, de eksisterende lokale og nationale IKT-markeders styrke, forbrugernes og erhvervslivets villighed til at acceptere nye produkter og serviceydelser samt tilgængeligheden af markedskanaler for de sprogteknologiske komponenter. Den tætte forbindelse imellem markedsmulighederne og den omfattende sprogteknologiske forskning tyder på, at man på dette område, ligesom på andre områder inden for den IST-baserede forskning, skal medregne forretningsmiljøet og den tekniske infrastruktur, hvis Europa fuldt ud skal udnytte sit potentiale på dette vigtige teknologiske område.

Analogien med 'informationsmotorvejen' er meget sigende i relation til sprogteknologien, eftersom den praktisk talt vil eliminere alle hindringer for kommunikation på tværs af EU's netværk og muliggøre den fri informationsstrøm og de serviceydelser, der er baseret på information. Ud fra denne synsvinkel burde den sprogteknologiske infrastruktur have den samme status og prioritet som den fysiske infrastruktur, der tillader fri bevægelighed for varer og mennesker i EU.

Anbefalinger

Sprogteknologi i det europæiske forskningsrum
Der bør etableres en konkret synliggørelse af de sprogteknologiske aktiviteter i det europæiske forskningsrum: Målet bør være et antal robuste, bredt dækkende, stabile, flersprogede sprogteknologiske moduler. Dette mål opnås sandsynligvis bedst, hvis den sprogteknologiske forskning både er en prioritet inden for IST i det europæiske forskningsrum (grænseflader, kognitive processer, interaktion, videnteknologier og semantik) og behandles som en følgeteknologi på den innovative forskningsdagsorden. Et grundlæggende mål bør være at tilvejebringe ovennævnte moduler for alle nuværende og fremtidige EU-sprog og at sikre, at komponenter og ressourcer for mindre udbredte sprog også kommer med.

Visionære strukturer
Der bør etableres et 'Sprogteknologisk Agentur', der kan forestå overvågningen af den gradvise overgang fra de nationale indsatser til et egentligt europæisk teknologiniveau baseret på sproglig lighed, og som kan samle og udbrede en bedste praksis på alle niveauer inden for den sprogteknologiske F&U. Et sandsynligt første skridt kunne være oprettelsen af et LangTech-observatorium, hvilket ville give det europæiske forskningsmiljø adskillige fordele:

- Kortlægning af forskningen for at undgå overlap, en mere åben adgang til udnyttelsen af resultaterne samt rådgivning i forbindelse med fastlæggelsen af den sprogteknologiske forskningsdagsorden.
- Støtte til at indtænke sprogteknologi i alle relevante europæiske forskningsområder ved at gøre området mere synligt.
- Ændring af fokus fra rent geografisk definerede nationale data til en mere 'sprogligt orienteret' observationsfunktion ved at overføre den bedste europæiske praksis til nationalt plan.
- Tilvejebringelse af pålidelige data til kortlægning af innovationen i EU.

Sprogteknologiske infrastrukturfonde
Noget i retning af 'støtte til lingvistisk infrastruktur' ville være en passende investering for at støtte sprog, der mangler en stærk kerne af komponenter og ressourcer, og som halter bagefter i overgangen til næstegenerations-sprogteknologi.

Forskningsplanlæggere bør overveje at adskille 'sprog' fra 'geografi' og at støtte F&U i den lingvistiske infrastruktur, hvor der er størst sandsynlighed for, at det vil bære frugt. Dette bør omfatte samarbejde på tværs af grænser mellem steder med stærke forskningsmiljøer og geografiske områder med teknologisk mindre udviklede sprog (især de nye ansøgerlande). Dette ville også gavne sprogsamfund med en relativt stærk forskning, men med dårligere lokale muligheder for at udnytte den.

Den digitale sproginfrastruktur
Selv om 'ejerforholdet' til sprog i sidste ende bør være adskilt fra geografien ud fra et finansieringsperspektiv, vil dette tydeligvis tage tid. Nationale sprogteknologiske 'agenturer' eller sponsorer vil stadig spille en stor rolle, som for eksempel i den 'digitale sproginfrastruktur', der udvikles af Nederlandse Taalunie for hollandsk-flamsk, samt i lignende initiativer under det franske TechnoLangue-program. Dette kunne skabe én mekanisme til sikring af, at alle europæiske sprog får samme kerneressourcer og -komponenter – eller i det mindste at manglende elementer identificeres.

EU's rolle (via det foreslåede 'Sprogteknologiske Agentur') bør være at støtte den fælles definition af, hvad der forstås ved en grundlæggende 'sprogværktøjskasse', uden hvilken den sprogteknologiske udvikling ikke kan fortsætte, at fremme udviklingen af åbne platforme for sådanne værktøjskasser samt at igangsætte processen med at opstille standarder for samspil med og mellem sprogkomponenterne. Dette vil indebære, at man definerer og aftaler krav til form og indhold, tilgængelighed, multifunktionalitet og genanvendelighed.

I en startfase bør man overveje Taalunie's erfaringer som en model, der kan udvides til alle europæiske sprog, hvorved der iværksættes en proces, der kunne føre til et paneuropæisk netværk til sponsering af sprogteknologi for bestemte sprog, med konkrete fordele for teknologioverførslen. Taalunie har vurderet, at udgifterne til agenturet vil blive omkring 500.000 € om året for hollandsk-flamsk. Det betyder, at EU løbende kunne støtte grundlæggende 'sprogværktøjskasser' for 20 sprog til en pris af 10 millioner €/år – en relativt beskeden sum i forhold til de aktuelle udgifter til sprogservice i Europa. Målet bør være, at udviklingen af en digital sproginfrastruktur for Europa baseres på 'open source' – dette ville konvergere med de andre 'open source'-initiativer fra fx digital forvaltning.

Alt dette skulle overvåges af det foreslåede 'Sprogteknologiske Agentur', hvis berettigelse, status og sammensætning skulle være genstand for udstrakt samråd. Agenturet kan sikre, at Europas fundamentale sprogteknologiske dagsorden vil opnå en uafhængighed af rammeprogrammerne og en kontinuitet i indsatsen, der er større end ved den nuværende projektorienterede støtteform.

Forkortelser

ERA	(The European Research Area) Det Europæiske Forskningsrum
ERCIM	(European Consortium for Informatics and Mathematics) Det europæiske forskningskonsortium for informatik og matematik
HLT	(Human Language Technologies) Sprogteknologi
IKT	Informations- og kommunikationsteknologier
IST	(Information Society Technologies) Informationssamfundets teknologier (del af 5. og 6. Rammeprogram)
JRC	(Joint Research Center) Det fælles forskningscenter
MT	(Machine Translation) Maskinoversættelse
F&U	Forskning og udvikling (R&D Research and Development)