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.

http://www.springer.com/978-3-642-30626-6

The Danish Language in the Digital Age
Rehm, G.; Uszkoreit, H. (Eds.)
2012, VI, 73 p. 24 illus. in color., Softcover
ISBN: 978-3-642-30626-6