

HOTET MOT VÅRA SPRÅK: EN UTMANING FÖR SPRÅKTEKNOLOGIN

Vi bevittnar för närvarande en digital revolution med enorma effekter på kommunikation och samhälle. Den senaste utvecklingen inom den digitala informations- och kommunikationsteknologin jämförs ibland med Gutenbergs uppfinning av boktryckarkonsten. Vad säger oss den liknelsen om framtiden för det europeiska informationssamhället och särskilt för våra språk?

Den digitala revolutionen kan jämföras med Gutenbergs uppfinning av boktryckarkonsten.

Gutenbergs uppfinning ledde till såna stora genombrott i informations- och kunskapsutbyte som t. ex. Luthers översättning av bibeln till folkspråket. Senare århundraden bevittnade framväxten av kulturella teknologier för mer effektiv språkanvändning och kunskapsutbyte:

- Ortografisk, lexikalisk och grammatisk standardisering av språken möjliggjorde snabb spridning av nya vetenskapliga och intellektuella idéer.
- Skapandet av standardspråk gjorde det möjligt för medborgare att kommunicera fritt inom vissa – ofta politiska – gränser.
- Språkundervisning och översättning underlättade meningsutbyte mellan språken.
- Utvecklingen av redaktionell och bibliografisk praxis garanterade kvaliteten i tryckt text.

- Uppkomsten av olika medier som böcker, tidningar, radio, television uppfyllde olika och varierade kommunikationsbehov.

Under de senaste två årtiondena har informationsteknologin möjliggjort automatisering och förenkling av en rad aktiviteter:

- Skrivmaskiner och textsättning har ersatts av ordbehandling och desktopprogram.
- Presentationsprogramvara har ersatt overheadbilder.
- Meddelanden och dokument kan skickas mycket snabbare och enklare med epost än med fax eller telex.
- Skype erbjuder telefoni och telekonferenser över internet till ingen eller låg kostnad.
- Digitala audio- och videoformat underlättar utbyte av multimediainnehåll.
- Sökmotorer ger tillgång till webbsidor med enkla sökord.
- Onlinetjänster som Google Translate levererar snabba grovöversättningar.
- Sociala medier (Facebook, Twitter) underlättar kommunikation och informationsutbyte.

Alla dessa verktyg och tillämpningar är helt klart praktiska, men långt ifrån tillräckliga för att säkerställa ett obehindrat flöde av information och varor i ett europeiskt samhälle som ska förbli varaktigt flerspråkigt.

2.1 SPRÅKGRÄNSER HÅLLER TILLBAKA DET EUROPEISKA INFORMATIONSSAMHÄLLET

Vi kan inte förutsäga exakt hur det framtida informationssamhället kommer att se ut. Det är ändå mycket troligt att kommunikationsteknologirevolutionen kommer att föra samman talare av olika språk på nya sätt. Därmed ökar kraven på individen, som behöver lära sig nya språk, men i synnerhet på teknikutvecklare, som behöver ta fram nya lösningar för ömsesidig förståelse och kunskapsutbyte. I dagens globala ekonomi och informationssamhälle leder nya typer av media till ökad interaktion mellan olika språk, språkbrukare och informationsinnehåll. Den popularitet som vi ser hos sociala medier (Wikipedia, Facebook, Twitter, YouTube och Google+) är bara toppen på isberget.

I det globala informationssamhället konfronteras vi med olika språk, språkbrukare och informationsinnehåll.

Att skicka text i gigabytemängder runt världen är idag gjort på några få sekunder, så snabbt att vi inte ens hinner uppfatta att texten är på ett språk som vi inte förstår. Enligt en färsk EU-rapport köper 57 % av internetanvändarna i Europa varor och tjänster på ett språk som inte är deras modersmål. Engelska är det vanligaste främmande språket, följt av franska, tyska och spanska. Av användarna läser 55 % innehåll på ett främmande språk och 35 % använder ett annat språk för att skriva epost eller kommentarer på webben [2]. Så sent som för några år sen kunde man kalla engelska webbens lingua franca – den överväldigande merparten av innehållet på webben var då på engelska – men situationen har nu förändrats drastiskt. Andelen webbinnehåll på andra europeiska språk (och andra språk överhuvudtaget) har vuxit explosionsartat.

Överraskande nog har denna globala språkliga klyfta inte fått särskilt mycket uppmärksamhet i det offentliga samtalet, trots att den väcker en stor och akut fråga: Vilka av Europas språk kommer att frodas i framtidens sammanlänkade informations- och kunskapssamhälle och vilka är dömda till undergång?

2.2 HOTET MOT VÅRA SPRÅK

Boktryckarkonsten ökade informationsutbytet i Europa, men samtidigt ledde den till många europeiska språks undergång. Regional- och minoritetsspråk upphöjdes sällan till rangen av skrivna standardspråk. Språk som korniska (nästan utdött på 1700-talet men nu återupplivat) och dalmatiska (utdött på 1800-talet) förblev därför enbart talade språkformer, vilket i sin tur begränsade deras användbarhet i Europas nya språkliga ekologi. Har turen nu kommit till våra nutida skriftspråk på grund av internet?

Europas språkliga mångfald är en av våra rikaste och viktigaste kulturskatter.

De ungefär 80 språk som talas i Europa är en av våra rikaste och viktigaste kulturskatter och en central del av den unika europeiska samhällsmodellen [3]. Även om språk som engelska och spanska troligen kommer att överleva på den framväxande digitala marknaden, kan många andra av våra språk sannolikt bli överflödiga i ett sammanlänkat informationssamhälle. En sådan utveckling skulle försvaga Europas globala position och den skulle stå i motsats till den strategiska principen om varje europeisk medborgares samhällsdeltagande på lika villkor oavsett språk.

I en UNESCO-rapport om flerspråkighet understryks språkets nyckelroll för utövandet av grundläggande rättigheter såsom uttryckande av politiska åsikter, utbildning och samhällsdeltagande [4].

2.3 SPRÅKTEKNOLOGI ÄR EN NYCKELTEKNOLOGI

Ekonomiska satsningar på språkbevarande handlar traditionellt framför allt om språkundervisning och översättning. Enligt en uppskattning uppgick marknaden för översättning, tolkning, mjukvarulokalisering och webbplatsglobalisering i Europa till 8,4 miljarder euro år 2008 och beräknades stiga med 10 % årligen [5]. Ändå motsvarar detta bara en liten del av dagens och morgondagens behov av informationsutbyte mellan språk. Den enda realistiska lösningen för att säkerställa att morgondagens europeiska språkliga ekologi uppvisar samma mångfald och djup är att använda oss av teknologi, precis som vi använder teknologi för att uppfylla våra energi- och transportbehov, m.m.

Europa behöver robust språkteknologi till låg kostnad för alla europeiska språk.

Språkteknologi för alla former av skriven text och talat språk kan hjälpa människor att samarbeta, göra affärer, utbyta kunskap och delta i den samhälleliga och politiska debatten oavsett språkskillnader och datormognad. Språkteknologi finns ofta dold under ytan som en komponent i komplexa mjukvarusystem. Redan idag möjliggör den:

- informationssökning med sökmotorer
- stavnings- och grammatikkontroll
- produktrekommendationer i webbutiker
- GPS:er som talar till användaren
- översättning av webbsidor online

Språkteknologi består av en rad basteknologier, som kan användas i olika typer av tillämpningar. Syftet med META-NET-vitböckerna är att belysa i vilken grad dessa basteknologier är tillgängliga för Europas språk.

För att behålla sin ledande position inom global innovation, behöver Europa robust språkteknologi till låg kostnad för alla sina språk, för integrering i nyckelapplikationer. Utan språkteknologi kommer vi inte i framtiden att kunna åstadkomma en genuint effektiv användarupplevelse präglad av interaktivitet, multimedialitet och flerspråkighet.

2.4 SPRÅKTEKNOLOGINS MÖJLIGHETER

Boktryckarkonsten innebar ett teknologiskt genombrott som ledde till att en text snabbt kunde mångfaldigas med en mekanisk tryckpress. Människor behövde utföra det mödosamma arbetet med att lokalisera, bedöma, översätta och sammanfatta kunskap. Det dröjde till Edison innan det gick att bevara talat språk för eftervärlden, och då med en teknik för enbart analog lagring och kopiering.

Med hjälp av språkteknologi kan vi idag förenkla och automatisera översättning, innehållsproduktion och informationshantering för alla Europas språk. Teknologi möjliggör också lättanvända talbaserade gränssnitt för hemelektronik, maskineri, fordon, datorer och robotar. Fullskaliga kommersiella och industriella tillämpningar är fortfarande i sin linda, men forskning och utveckling inom språkteknologi uppvisar redan resultat som antyder en stor potential. Exempelvis finns nu maskinöversättning av godtagbar kvalitet inom specifika fackområden och prototypsystem har tagits fram för flerspråkig informationshantering och innehållsproduktion på flera europeiska språk.

Precis som har varit fallet med många andra teknologier, utvecklades de första språkteknologitillämpningarna – som t. ex. talbaserade användargränssnitt och dialogsystem – för smala domäner, och hade ofta begränsad funktionalitet. Marknadspotentialen är dock enorm inom utbildnings- och nöjesindustrin för integrering

av språkteknologi i spel, edutainmentpaket, bibliotek, simulerings- och utbildningsprogramvara.

Mobila informationstjänster, datorstödd språkinläring, e-utbildningsplattformar, programvara för självtest och plagiatdetektering är några tillämpningsområden där språkteknologi kan spela en viktig roll.

Den popularitet som sociala media som Twitter och Facebook åtnjuter pekar på ett behov av sofistikerade språkteknologifunktioner som kan följa inlägg, sammanfatta diskussioner, påvisa opinionstrender, identifiera känsloreaktioner, upptäcka upphovsrättsintrång eller spåra missbruk.

Språkteknologi bidrar till att motverka
att språklig mångfald uppfattas
som ett "handikapp".

Språkteknologi innebär en oerhörd chans för EU, genom att den erbjuder ett sätt att hantera den komplexa frågan om mångspråkighet i Europa, det faktum att olika språk används naturligt sida vid sida i Europa i näringsliv, organisationer och skolor. Medborgarna behöver därmed ständigt kunna kommunicera över språkgränser, och språkteknologi kan bidra till att övervinna denna sista barriär och samtidigt främja fri och allmän användning av de enskilda språken.

På längre sikt kommer innovativ europeisk språkteknologi att visa vägen för våra globala partners när de börjar stödja sina egna mångspråkiga samhällen. Språkteknologi kan ses som ett slags tekniskt hjälpmedel för att kompensera för det "handikapp" som språklig mångfald kan uppfattas som, genom att det ger språkgemenskaperna större tillgång till varandra.

Slutligen är ett aktivt forskningsområde användning av språkteknologi vid räddningsinsatser i katastrofområden, där systemfunktionen kan betyda skillnaden mellan liv och död. I framtiden kan vi få se livräddare i form av intelligenta flerspråkiga robotar.

2.5 SPRÅKTEKNOLOGINS UTMANINGAR

Även om vi har sett stora framsteg inom språkteknologi under de senaste åren, är takten i tekniska framsteg och produktinnovation fortfarande för låg. Allmänt använda funktioner som stavnings- och grammatikkontroll i ordbehandlingsprogram är typiskt enspråkiga och finns bara för en handfull språk.

Teknikutvecklingen behöver skyndas på.

Även om man nu med de översättningstjänster som är tillgängliga online snabbt kan få en grovöversättning av ett dokument, kommer de till korta om man kräver en exakt och komplett översättning. På grund av det mänskliga språkets komplexitet, är det ett tids- och resurskrävande företag att bygga modeller av våra språk i mjukvara och testa modellerna i verkliga livet, något som kräver ett stabilt långsiktigt finansieringsåtagande. Europa måste därför behålla sin roll som pionjär när det gäller att ta sig an de teknologiska utmaningar som ett mångspråkigt samhälle innebär genom att utveckla ny metodologi för att accelerera utvecklingen på bred front. Här kan det handla såväl om nya komputationella paradigmer som om tekniker för storskaligt decentraliserat kollektivt samarbete av den typ som Wikipedia har stått modell för ("crowdsourcing").

2.6 HUR MÄNNISKOR OCH MASKINER LÄR SIG SPRÅK

För att illustrera hur datorer hanterar språk och varför det är ett så svårt problem att programmera dem så att de förstår och producerar språk på mänsklig nivå, ska vi ta en översiktlig titt på hur människor lär sig sitt eller sina modersmål och andra språk för att sedan se hur språkteknologisystem fungerar.

Människor lär sig språk på två sätt. Spädbarn lär sig språk genom att höra och ta del i interaktionen bland sina föräldrar, syskon och andra personer i deras omgivning. Vid ungefär två års ålder börjar barnen själva yttra sina första ord och korta fraser. Detta är möjligt enbart därför att människor har en genetiskt betingad förmåga att upprepa och så småningom lära sig att förstå språk (talat språk eller teckenspråk) som riktas till dem.

Att lära sig ett andraspråk efter de tidiga barndomsåren kräver betydligt större medveten ansträngning, framför allt därför att barnet då inte är omgivet av en språkgemenskap av modersmålsstalare. I skolan lär man sig ofta främmande språk genom att grammatisk struktur, ordförråd och stavning övas med hjälp av explicita lingvistiska regler, tabeller och exempel.

Om vi nu istället ser på hur språkteknologisystem ”lär sig” språk, finner vi samma två huvudtyper av inläring. Statistiska (eller ”datadrivna”) metoder får sin språkkunskap ur enorma mängder konkreta textexempel genom en process som kallas ”maskininläring”. För att ta fram exempelvis ett stavningskontrollprogram räcker det med text på ett språk, medan parallella texter på två eller flera språk behövs för att träna ett maskinöversättningssystem. Maskininlärningsalgoritmen ”lär sig” då mönster för hur ord, korta fraser och hela meningar översätts.

De statistiska metoderna kräver normalt miljontals meningar för att uppnå godtagbar kvalitet. Detta är en viktig anledning till att sökmotorföretag vill samla in så mycket text som möjligt. Stavningsrättning i ordbehandlare och tjänster som Googles sökmotor och översättningstjänst bygger alla på statistiska metoder. Deras stora fördel är att datorn lär sig snabbt i en serie successiva träningsomgångar, även om kvaliteten kan variera godtyckligt.

Den andra typen av språkteknologisystem använder explicit formulerade regler. Ett regelbaserat maskinöversättningssystem bygger t. ex. på att språkvetare, dataling-

vister och datavetare tillsammans explicit kodar grammatiska analyser (översättningsregler) och sammanställer lexikal information (ordlistor), något som kräver mycket tid och arbete. Utvecklingen av några av de ledande regelbaserade maskinöversättningssystemen har bedrivits kontinuerligt under mer än två decennier. Den stora fördelen med regelbaserade system är att experterna har noggrannare kontroll över språkbearbetningen, vilket gör det möjligt att systematiskt korrigera fel i bearbetningen. Det är också lätt att ge användaren detaljerad återkoppling, vilket är en fördel särskilt när regelbaserade system används i datorstödd språkinläring. Då utvecklingen av regelbaserade språkteknologisystem är förknippad med så höga kostnader, har sådana system med få undantag utvecklats enbart för några få stora språk.

Människor lär sig språk på två sätt: genom exempel och genom att lära sig språkliga regler.

Eftersom de statistiska och regelbaserade systemen tenderar att uppvisa komplementära styrkor och svagheter, fokuserar forskningen nu på att utveckla hybridssystem med kombinationer av de två metoderna. Dessa har dock hittills inte rönt samma framgång i kommersiella tillämpningar som i forskningslaboratorierna.

Som vi har sett i detta avsnitt, är många av de mest använda tillämpningarna och tjänsterna i dagens informationssamhälle starkt beroende av språkteknologi. Detta gäller inte minst den europeiska ekonomin och informationssamhället. Även om denna teknologi har utvecklats starkt under senare år, har språkteknologin fortfarande en enorm förbättringspotential när det gäller systemens kvalitet. I de två följande avsnitten beskriver vi vilken roll svenska språket spelar i det europeiska informationssamhället samt presenterar en översikt över befintlig språkteknologi för svenska.

The Swedish Language in the Digital Age

Rehm, G.; Uszkoreit, H. (Eds.)

2012, VI, 79 p. 25 illus., 2 illus. in color., Softcover

ISBN: 978-3-642-30831-4