

Språkteknologi för ökad tillgänglighet – vilka möjligheter finns?

Rickard Domeij

Språkrådet vid Institutet för språk och folkminnen

Rickard.Domeij@sprakradet.se

Sammanfattning

Kraven på myndigheterna att använda andra språk än svenska i kommunikationen med medborgarna växer. Kan automatisk översättning och andra språkteknologiska hjälpmedel göra svenska myndigheter mer tillgängliga i en flerspråkig och digital värld? Frågan diskuterades på det nordiska seminariet *Språkteknologi för ökad tillgänglighet* i anslutning till den svenska språkteknologikonferensen SLTC 2011. Denna artikel beskriver översiktligt språkteknologins möjligheter att bidra till ökad tillgänglighet utifrån innehållet på seminariet.

Vad kan språkteknologin bidra med?

Efterfrågan på flerspråkig information och kommunikation ökar kraftigt i hela världen med tilltagande globalisering. Storleken på EU:s översättningsbudget sägs vara uppe i en miljard euro om året. Också inom Sverige ökar behoven av kommunikation på olika språk, inte bara med andra länder utan även mellan svenska myndigheter och en flerspråkig befolkning.

Samtidigt utvecklas tekniken i rasande takt och skapar nya sätt att hantera flerspråkighet på, vilket myndigheterna inte är sena att uppmärksamma. Som exempel erbjuds en besökare på Stockholms kommuns webbplats idag översättning av allt innehåll till över 50 olika språk helt automatiskt. Men är resultatet verkligen tillräckligt bra? Vad tycker nätmedborgarna själva?

På den svenska språkteknologikonferensen SLTC 2010 i Linköping i slutet av oktober arrangerades ett seminarium om vad språkteknologin kan bidra med för att göra myndigheterna i de nordiska länderna mer tillgängliga för alla. Bakom seminariet stod Astin – de nordiska språkrådens arbetsgrupp för språkteknologi. Seminariet gav en intressant inblick i vad som sker.

Statistisk översättning med Google Translate

Keith B. Hall från Googles forskningsavdelning kom till språkteknologikonferensen i Linköping för att tala om Googles arbete med språkteknologi. Han berättar att Google tidigare erbjöd översättning av eftersökta webbsidor med Systrans maskinöversättningssystem BabelFish. Det var ett traditionellt maskinöversättningssystem med handskrivna regler som beskriver hur en mening i ett språk ska översättas till ett annat.

Sedan 2006 har Google ett egenutvecklat system där översättningen i stället baseras på statistik över tidigare översättningar. Googles tillgång till enorma mängder mänskligt översatta texter gör att de statistiska modellerna ger ett bättre resultat än de tidigare regelbaserade metoderna.

–Idag finns Google Translate på 57 olika språk och kan användas både för att söka information på ett annat språk och få de eftersökta webbsidorna översatta till det egna språket, säger Keith B. Hall. Systemet kan också användas för att översätta e-brev och chatt i Gmail och textremsor i filmer på Youtube.

Och faktum är att de automatiska översättningarna idag är mycket bättre än tidigare med BabelFish. Men är de tillräckligt bra för att användas på myndigheternas webbplatser?

Kvaliteten räcker inte

Stefan Johansson från tillgänglighetskonsulten Funka Nu berättade på seminariet i Linköping att de har börjat undersöka hur användbara Googles automatiska översättningar är för personer med svenska som andraspråk. Än så länge har de bara hunnit göra inledande tester för att jämföra begripligheten hos automatiskt översatta webbtexter med mänskliga översättningar av samma texter.

–De preliminära testerna visar att användarna inte ens förstår de mänskliga översättningarna. Det pekar på ett grundläggande problem med myndighetstexter: att de ofta är för krångligt skrivna redan från början. Och det blir förstås inte bättre om översättningarna inte är att lita på.

Och det är de inte. Stefan Johansson visar exempel på både godtagbara och mindre godtagbara översättningar. Han använder svenska och norska som exempel på att resultatet ibland är svårt att förstå, som när meningen ”Hur upplever läsaren texten?” översätts till ”Hvad gjør leseren texten?”.

I det här fallet kan man gissa att det är engelska do-omskrivningar som ställer till det. Jo, du läste rätt. Google använder nämligen engelska som övergångsspråk när det inte finns tillräckligt med text som underlag för att göra tillförlitliga statistiska beräkningar för andra språk.

Maskinöversättning att lita på?

Frågan är om det överhuvudtaget är möjligt att göra pålitliga översättningar på automatisk väg. Svaret kan komma att stavas Molto. EU-projektet som koordineras från Göteborgs universitet har fått 25 miljoner kronor för att skapa ett pålitligt verktyg för automatisk översättning som ska fungera mellan EU-språken.

Projektet vara klart 2013. På längre sikt är förhoppningen att Molto ska bidra till att alla EU-medborgare, oavsett modersmål, ska få samma tillgång till information och service på nätet.

Molto skiljer sig från Google på flera sätt. En avgörande skillnad är att Googles översättningssystem riktar sig till informationskonsumenterna, medan Molto riktar sig till producenterna.

–Eftersom producenterna har ett ansvar för informationens kvalitet, måste översättningarna gå att lita på, säger Aarne Ranta från Göteborgs universitet som leder projektet. Google Translate fungerar oftast mycket bra för den som snabbt vill skaffa sig en uppfattning om innehållet på en webbsida, men det räcker inte för den som har ett ansvar för det som står där, fortsätter han.

Han påpekar att statistiska metoder fungerar hyfsat bra på engelska, som har fast ordföljd och mycket enkel morfologi, men mindre bra på andra språk. Därför använder Molto regelbaserade metoder som klarar både tysk ordföljd och finska kasus lika bra i översättningar till och från alla andra EU-språk.

– Den som ska sälja något kan inte ha ett system som översätter ”prix 99 euros” till ”pris 99 kronor”, ett vanligt fel i statistiska översättningar, säger Aarne Ranta.

Bara inom begränsade domäner

Men Molto kommer inte att kunna översätta en godtycklig text som Google Translate gör. För att få en mer pålitlig översättning måste man begränsa sig till vissa domäner, enligt Aarne Ranta.

–I stället för att översätta texter av alla möjliga slag håller vi oss till en viss typ, som medicinska patent. Det gör att ordförrådet kan begränsas och grammatiken beskrivas med regler på ett mer förutsägbart och kontrollerat sätt.

Grammatiska beskrivningar för många EU-språk finns sedan tidigare, liksom det system som beskrivningarna görs i, Grammatical Framework. Skillnaden är att arbetet nu ska göras i större skala och på ett mera samordnat sätt. Systemet ska också kompletteras med statistik när det är motiverat.

Tre domäner som ska provas till att börja med är medicinska patent, matematiska uppgifter och beskrivningar av objekt på museer. För varje domän skapas en semantisk beskrivning av ingående begrepp och relationer, en så kallad ontologi. Den fungerar som ett mellanspråk, ett slags interlingua. Sedan görs grammatiska beskrivningar för varje språk så att allt som finns beskrivet inom domänen kan översättas mellan språken.

Terminologi viktigt

Det låter svårt, men i grunden handlar det om ett gediget terminologiskt arbete, något som Henrik Nilsson från Terminologicentrum TNC underströk betydelsen av.

–Många tror att terminologi gör språket svårare, men ofta är det tvärt om. När man klargjort olika begrepps innebörd och relationer, och bestämt sig för vilka termer som ska användas för de olika begreppen blir myndighetsspråket tillgängligare för alla.

Ibland behöver myndigheterna använda facktermer och uttryck som kan vara svåra att förstå, som ”bostadsbidrag” och ”rehabiliteringsersättning”. Då är det viktigt att förklara ordet första gången det används i texten eller göra en ordlista med förklaringar på webbplatsen.

I översatta texter är det viktigt att man förstår att en översatt term som ”housing allowance” är detsamma som ”bostadsbidrag” genom att till exempel skriva den svenska termen inom parentes i den översatta texten så att termerna kopplas till varandra.

En konsekvent användning av termer gör det lättare att översätta både för människor och maskiner, säger Henrik Nilsson.

TNC har tillsammans med språkteknologiföretaget Fodina tagit fram Term-O-Stat, ett fyrstegsprogram för att hjälpa myndigheter att arbeta med den egna terminologin. I språklagen står det uttryckligen att myndigheterna har ett ansvar för att terminologin inom det egna verksamhetsområdet ”finns tillgänglig, används och utvecklas”.

Magnus Merkel på Fodina menar att arbetet med terminologi är en viktig förutsättning när man utvecklar och använder sig av språkteknologi.

–Har man en enhetlig terminologi för flera språk parallellt kan man till exempel utveckla teknik för att söka efter ett samiskt ord och få träff på ett svenskt med samma betydelse. Då skulle det gå att använda nationella minoritetsspråk på myndigheternas webbplatser på helt andra sätt än idag, säger han.

Och han vet vad han talar om. Fodina har för Patent- och registreringsverket PRV tagit fram en svensk-engelsk databas med patenttermer, som ska användas som grund för att automatiskt översätta patent inom den europeiska patentorganisationen EPO.

En infrastruktur för språkanvändningen på internet

För att se till att också den samiska minoritetsbefolkningen får ta del av språkteknologins möjligheter utvecklar Sametinget i Norge elektroniska ordböcker, termdatabaser, korrekturprogram, språkinlärningsprogram, talsyntes, översättningsminnen och maskinöversättning för samiska.

–Om samhället ska ta ansvar för de nationella minoritetsspråken, som det står i språklagen, måste vi också se till att minoritetsspråken får liknande möjligheter att användas på internet som majoritetsspråken, säger Sjur Moshagen från Sametinget i Norge.

Som grund för teknikutvecklingen behövs uppmärksatta språkdatabaser och tillhörande analysverktyg. Man kan se det som ett slags infrastruktur i den digitala världen, motsvarande vägar och tågräls, för att möjliggöra kommunikation på olika språk på internet så att inte alla tvingas använda majoritetsspråket.

Svenskan har samma problem i förhållande till engelskan, majoritetsspråket på hela internet, som samiskan har i förhållande till svenskan, majoritetsspråket inom den svenska delen av internet. Men än har vi i Sverige inte gjort som i Nederländerna, Storbritannien och nu senast i Norge, där stortinget avsatt 100 miljoner norska kronor för att skapa en nationell språklig infrastruktur som grund för språkanvändningen i det digitala samhället.

Johan Hartman, tillgänglighetsansvarig på SVT, pekar på nödvändigheten av en liknande satsning i Sverige:

–Vi har som mål att texta alla våra program, även direktsändningar, som man numera gör på BBC. För att nå målet behöver vi utveckla teknik för automatisk igenkänning av talad svenska, men det går inte utan statsfinansierade språkdatabaser eftersom det kommersiella intresset inte är tillräckligt stort.

Samma språkdata-baser kan användas för att driva på teknikutvecklingen och öka tillgängligheten till information också i andra sammanhang, vilket gör det till en nationell angelägenhet.

Möjligheter och begränsningar

På seminariet i Linköping blir det uppenbart att den språkteknologiska utvecklingen har skapat helt nya möjligheter att göra information mer tillgänglig för människor med olika språkliga och kommunikativa behov.

Björn Granström från Kungliga tekniska högskolan avslutade seminariet med att visa att det i framtiden också kan bli möjligt att automatiskt översätta texter till teckenspråk och få dem tecknade med konstgjorda gester från så kallade teckenspråksavatarer. Sådana dataanimerade människofigurer används redan för att komplettera konstgjort tal med munrörelser och gester.

Men samtidigt finns svårfråkomliga begränsningar i tekniken som det gäller att vara uppmärksam på, om inte de negativa effekterna ska överskugga de positiva. Talsyntes fås till exempel till priset av sämre kvalitet på uppläsningen i jämförelse med om en människa skulle läsa upp samma text. Fördelen med automatisk uppläsning är att den kan användas när mänsklig textuppläsning saknas eller inte är möjlig. Som i uppläsning av texter på webben.

På liknande sätt är det med maskinöversättning, men där är svårigheterna större och resultatet mer opålitligt. Visst kan man använda det för att översätta en godtycklig text på webben för att få en grov uppfattning om vad som står där. Men ska man verkligen kunna lita på det behövs en kvalitetsöversättning som än så länge bara en mänsklig översättare kan åstadkomma, utom möjligen inom begränsade domäner.

– Från majoritetsspråk till minoritetsspråk måste kvaliteten på översättningen vara bra. Då får man översätta manuellt med hjälp av elektroniska ordböcker, termdatabaser och översättningsminnen. Men från minoritetsspråket till majoritetsspråket kan det räcka med maskinöversatta texter som ger en grov förståelse av innehållet, säger Sjur Moshagen på Sametinget i Norge.

Man får inte heller glömma den potential som finns hos användarna själva att bidra till och förädla innehållet på internet när möjlighet ges. Facebook har översatts till 75 språk, de allra flesta av användarna själva. På Wikipedia finns 197 språk med fler än 10 000 uppslag. Allt skapat av användarna själva.

Så visst finns det möjligheter. Men det gäller att kombinera och använda dem på bästa sätt.

Delaktighet på ett flerspråkigt internet

Frågan är om de myndigheter som erbjuder översättning via Google Translate på sina webbplatser har tänkt igenom vad det kan få för konsekvenser. Troligen inte. De försöker bara möta de ökade kraven på flerspråkig information genom att gripa tag i det som finns närmast till hands.

Men hur uppfattar användarna det? Vems ansvar är det om någon användare förlitar sig på felaktigt översatt information? Myndighetens? Googles? Eller är det kanske nätmedborgarens eget fel?

Språklagen gör klart att språket i offentlig verksamhet ”ska vara vårdat, enkelt och begripligt”. Det gäller oavsett vilket språk som används. Det duger alltså inte med krångligt skrivna grundtexter och slarviga översättningar. Om myndigheterna vill kommunicera på ett tillgängligt sätt med en flerspråkig befolkning måste de anstränga sig mer.

Flerspråkig kommunikation handlar i grunden inte om att mekaniskt följa föreskrifter, utan om att lära sig att analysera kommunikationssituationen och ta reda på hur behoven ser ut i det enskilda fallet. Hur ser situationen ut i den egna kommunen? Vilka vill man nå? Vilka

behov av information och service har olika mottagargrupper? Med vilka medel når man dem bäst inom ramen för befintliga resurser?

Ju bättre man känner till de språkteknologiska hjälpmedel som står till buds, desto bättre kan behoven mötas på olika sätt: med klarspråk och lättläst, i tal och skrift, på det svenska teckenspråket, på något av de nationella minoritetsspråken, på engelska, spanska, arabiska, somaliska eller något annat språk. Men det måste vara människors behov som står i centrum, inte tekniken.

Språkteknologin kan erbjuda helt nya möjligheter att öka tillgängligheten till information och service på internet. Problemet är att den lika gärna kan användas som ett sätt att smita ifrån ansvaret.

Slutnot

En kortare version av denna text publicerades i Språktidningen, april 2011, s. 44–47, med titeln *Datorn god översättare i snäv mening*.