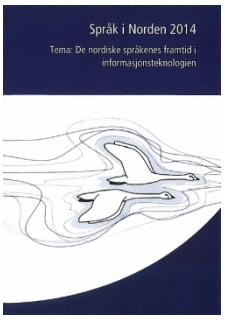
Sprog i Norden

Titel:	Språkteknologi och språkresurser för språken i Sverige: En statusrapport
Forfatter:	Lars Borin & Rickard Domeij
Kilde:	Sprog i Norden, 2014, s. 33-47
URL:	http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive



Betingelser for brug af denne artikel
Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Søgbarhed
Artiklerne i de ældre numre af Sprog i Norden (1970-2004) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Språkteknologi och språkresurser för språken i Sverige: En statusrapport

Lars Borin och Rickard Domeij

Under senare år har behovet av språkteknologi och språkresurser för svenska och andra språk i Sverige kommit att uppmärksammas allt mer. Detta handlar dels om att på bästa sätt förvalta Sveriges under lång tid upparbetade spetskompetens inom akademisk forskning i språkteknologi, dels även om att se till att denna forsknings resultat omsätts i produkter och tjänster som kommer det svenska samhället och dess medborgare till gagn. I denna artikel lämnar vi en rapport om den akademiska språkteknologiforskningen i Sverige (avsnitt 1) och om ett färskt politiskt initiativ för att få till stånd en nationell språkresursinfrastruktur som ska kunna användas för produkt- och tjänsteutveckling (avsnitt 2).

Mot en svensk språkteknologi för forskning och utveckling – en rapport från fronten

När det gäller språkteknologi för svenska och dess användning i forskning och utveckling, har denna å ena sidan en lång historia i Sverige men å den andra ser man att utvecklingen i viss mån har stagnerat.

Den svenska forskningen inom området går tillbaka till 1950-talet för talteknologi och 1960-talet för textkorpusar och språkverktyg för det skrivna språket och har lett till att det idag finns ett antal aktiva forskargrupper i landet, som representerar de flesta av språkteknologins områden:[1]

- svenska korpusar (Språkbanken vid Göteborgs universitet, Stockholm, Uppsala)

1 Dessutom finns en rad företag som utvecklar språkteknologiska produkter och tjänster och som bedriver egen forskningsverksamhet, ofta i samarbete med akademiska forskningsmiljöer. Fokus i det här avsnittet ligger dock på den akademiska forskningen i snävare bemärkelse.

- flerspråkiga korpusar (Uppsala, Linköping, Språkbanken)
- taldatabaser (KTH)
- resurser för informationsåtkomst (SICS, KTH, Stockholm, Språkbanken)
- lexikondatabaser (Språkbanken, KTH, Språkrådet)
- många olika språkverktyg för text och tal (samtliga grupper)

Huvudsakligen bedrivs denna forskning i form av avgränsade korta forskningsprojekt och den är mer fragmenterad än vad som skulle vara önskvärt. De huvudsakliga finansieringsformerna täcker inte långsiktigt underhåll och tillgängliggörande av språkresurser. Språkbanken i Göteborg är den enda grupp som systematiskt arbetar med att säkerställa detta, men strikt på eget initiativ och med lokal finansiering från Göteborgs universitets humanistiska fakultet.

Det finns således ett stort behov av samordning av resursuppbyggandet samt av harmonisering av språkresurser och språkverktyg med avseende på informationsstruktur, dataformat, programgränssnitt, licensvillkor, etc., så att en positiv spiraleffekt kan uppstå, där både ny forskning och kommersiell produktutveckling kan använda befintliga resurser fullt ut, utan att ständigt behöva återuppfinna hjulet eller hålla tillgodo med undermåliga lösningar.

I någon mån går utvecklingen i denna riktning, genom svensk medverkan i två breda europeiska initiativ, META-NORD och CLARIN, som beskrivs i de närmast följande avsnitten nedan.

META-NORD

META-NORD-projektet (2011–2013) var ett EU-finansierat samarbete mellan de fem nordiska och de tre baltiska staterna, med syftet att stärka och utveckla potentialen för språkteknologibaserad forskning och utveckling i det nordisk-baltiska området. META-NORD ingick tillsammans med två andra liknande projektsamarbeten under det vidare paraplyet META-NET,[2] ett europeiskt spetsforskningsnätverk med huvudsyftet att

2 <http://www.meta-net.eu/>

främja den teknologiska grunden för ett flerspråkigt europeiskt informationssamhälle.

Den svenska delen av META-NORD åstadkom tre typer av resultat:

1. En vitbok om tillståndet för svenska språket med avseende på språkresurser och språkteknologi (Borin m.fl. 2012a), samt medvetandegörande om detta tillstånd hos forskare, politiker, forskningsfinansiärer, språkteknologiföretag, m.fl.

2. En kartläggning av språkresurser och språkverktyg i Sverige samt katalogisering av dessa i den gemensamma databasen META-SHARE[3] (Skadina m.fl. 2011)

3. Anpassning av utvalda befintliga resurser till gemensamma standarder samt vidareutveckling av vissa resurser, i det svenska fallet länkning av ett svenskt ordnät till de danska, estniska och finska ordnäten via det engelska Princeton WordNet (Pedersen m.fl. 2012, 2013)

På grundval av den information som samlats in för de sammanlagt 32 META-NET-vitböckerna om 31 språk (bokmål och nynorska har varsin vitbok) har resursläget för dessa språk kunnat sammanfattas som i figur 1.

Sammanfattningsvis har META-NORD inom de relativt begränsade ramar som projektet förfogade över ändå kunnat lägga en god grund för vidare arbete, nationellt, nordiskt och europeiskt, och bidragit till att höja medvetenheten om behovet av fokuserade satsningar på svensk språkteknologi.

SWE-CLARIN

Termen *e-vetenskap* används ofta för att beteckna vetenskap som med hjälp av modern informationsteknik kan angripa problem av en omfattning och komplexitet som inte skulle kunna hanteras utan elektroniska hjälpmedel och möjligheten att dra fördel av geografiskt spridda resurser. Traditionellt har detta angreppssätt främst förknippats med extremt beräkningsintensiv forskning inom teknik- och naturvetenskap men i själva verket skulle man kunna säga att t.ex. korpuslingvistiken (som med hjälp

3 <http://spraakbanken.gu.se/metashare/>

av språkteknologi bedriver forskning på stora insamlade textmassor) är ett mycket tidigt exempel på e-vetenskap. Detta är också ett område som utvecklats starkt över femtio år, och den teknologi och teoribildning som utvecklats i anslutning till detta forskningsfält har idag fått en rik flora av applikationer utanför den rent lingvistiska intressesfären.

Figur 1. Resursläget för 31 av Europas språk enligt META-NETs vitbok (svenska inringat; Borin m.fl. 2012a: 33f).

Det stora europeiska infrastrukturprojektet CLARIN[4] (Common Language Resources and Technology Infrastructure) syftar till att göra digitala språkresurser (textsamlingar, inspelningar av ljud och bild, lexikon och så vidare), tillsammans med de språkteknologiska verktyg som behövs för att hantera dem, tillgängliga för forskare inom alla discipliner. Särskilt tonvikt läggs på forskare inom humaniora och samhällsvetenskap. Med andra ord: CLARIN erbjuder en möjlighet för alla forskare inom humaniora och samhällsvetenskap (och även andra discipliner som t.ex. medi-

4 <http://www.clarin.eu/>

cin och vårdvetenskap) att använda sig av e-vetenskapliga metoder i sin verksamhet.

Efter en fyraårig förberedelsefas fick CLARIN under 2012 status av ERIC (European Research Infrastructure Consortium), en sammanslutning av stater och mellanstatliga organisationer. Vid inrättandet den 29/2 2012 hade CLARIN ERIC 9 medlemmar (8 EU-medlemsländer och en mellanstatlig organisation). Sverige var inte en av de initiala medlemmarna. En rad svenska forskargrupper hade dock deltagit i CLARINs förberedelsefas, och nio organisationer bildade därför ett konsortium för att förbereda en ansökan till Vetenskapsrådet om svenskt medlemskap i CLARIN ERIC samt fullt svenskt deltagande i CLARINs arbete, genom att:

1. definiera en nationell organisation för CLARIN-arbetet,

2. utforma en struktur och organisation för CLARIN-centra i Sverige,

3. utforma en teknisk infrastruktur och definiera de tjänster som ska vara tillgängliga via CLARIN,

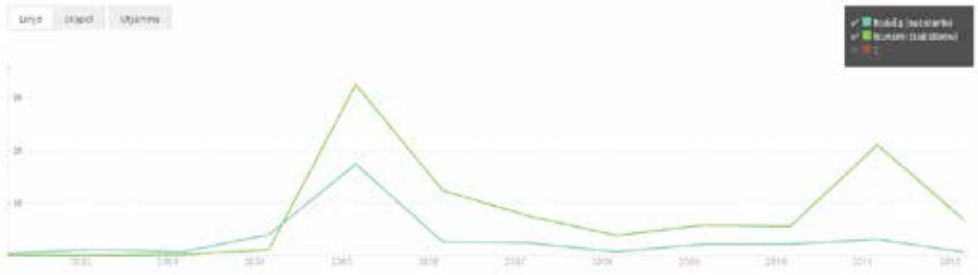
4. utreda behovet av språkresurser och språkverktyg för Sveriges språk (svenska, minoritetsspråken och andra viktiga språk) i ljuset av CLARINs behov och målsättningar,

5. identifiera och mobilisera viktiga aktörer i forskningsvärlden, i kulturarvsinstitutionerna och i relevanta digitaliseringsprojekt.

Ansökan beviljades av Vetenskapsrådet i november 2013, men exakt belopp och villkor är i skrivande stund inte kända utan ska bli föremål för förhandling under första kvartalet 2014. Den planerade nationella infrastrukturen inom området – SWE-CLARIN – kommer organisatoriskt att bestå av ett antal centra, baserade på väletablerade forsknings- och utvecklingsmiljöer. Dessa centra ska dels skapa, uppgradera och underhålla de digitala resurserna och verktygen, och dels fungera som utbildnings-, rådgivnings- och stödenheter för forskare utanför språkteknologigemenskapen.

Mot digital spetsforskningspotential med svensk språkteknologi: Kunskapsbaserad kulturomik

Som ett konkret exempel på den sorts aktiviteter som hör hemma inom CLARIN kan vi här anföra det nystartade VR-ramprogrammet *Mot kunskapsbaserad storskalig kunskapsutvinning ur svensk text* ('Towards a knowledge-based culturomics'), ett samarbete mellan Språkbanken i Göteborg, datavetare vid Chalmers tekniska högskola i Göteborg och språkteknologer vid Lunds universitet.



Figur 2. Förekomster av orden *flodvåg* (den undre blå kurvan) och *tsunami* (den övre gröna kurvan) i dagstidningar 2001–2012 (normaliserade lemmafrekvenser)

Nyligen har några forskare börjat utnyttja de enorma textmängder som resulterat ur Googles massiva bokdigitaliseringsprojekt för att i dessa textmassor försöka följa språklig och kulturell utveckling över de två senaste seklerna (Michel m.fl. 2011). Forskningsområdet har med buller och bång lanserats under namnet *culturomics*[5] (analogt med *genomics, proteomics,* etc.), men de första studierna har med rätta kritiserats för att helt ignorera relevanta tidigare arbeten i språkteknologi och lingvistik, och t.ex. inte diskutera det inte alldeles enkla begreppet "ord" i den här kontexten. Samtidigt är detta forskning som verkligen ligger i tiden. Det finns nu enorma mängder digital text att tillgå på svenska. Bara de svenska bloggarna uppgår till miljarder ord. Dessutom pågår ett antal kulturarvsdigitaliseringsprojekt, t.ex. har Kungliga biblioteket och Riksarkivet digitaliserat stora mängder svensk dagspress från de senaste 300 åren, sammanlagt miljarder ord.

Syftet med detta projekt är att lyfta *kulturomik* till kunskapsbaserad storskalig kunskapsutvinning ur stora mängder digitaliserad svensk text,

5 Se <http://www.culturomics.org/>

såväl modern som äldre (Borin och Johansson 2014). De inledande studierna i projektet behandlar politiska opinionsyttringar i sociala medier, utveckling av system som kan svara på frågor om textinnehåll, samt ordutveckling över tid. Sålunda visar figur 2 hur den tidigare huvudsakligen tekniska termen *tsunami* blev ett allmänord i samband med den tragiska naturkatastrofen i Sydostasien i slutet av 2004 och hur ordet sedan helt hade ersatt det tidigare *flodvåg* vid rapporteringen från Fukushimaolyckan 2011. Visualiseringen i figur 2 har gjorts med trenddiagramfunktionen i Språkbankens korpushanteringssystem Korp[6] (Borin m.fl. 2012b) på basis av ett stort dagstidningsmaterial.

Förslag till nationell infrastruktur för det digitala samhället

Språkrådet i Sverige har på uppdrag av den svenska regeringen (Ku2011/860/KA) tagit fram ett förslag på hur en nationell infrastruktur för språken i Sverige kan skapas med utgångspunkt i behovet av språkresurser för utveckling av taligenkänning för tv-textning (*Infrastruktur för språken i Sverige* 2012). Som bakgrund till uppdraget skriver regeringen:

> För att främja utvecklingen av teknik som ökar tillgängligheten till information för alla har flera berörda aktörer uttryckt behovet av att etablera en nationell språkdatabank med öppet tillgängliga språkdatabaser och tillhörande analysverktyg. En nationell språkdatabank avseende det svenska språket, de nationella minoritetsspråken och det svenska teckenspråket som omfattas av språklagen (2009:600), kan också utgöra grund för olika tillgänglighetstjänster, t.ex. på tv-området.

I det följande presenteras bakgrunden till förslaget, huvuddragen i det och hur regeringen tänker sig att gå vidare med det.

Ökade krav på tillgänglig information och kommunikation

Kraven på informationstillgänglighet har ökat i takt med att den digitala tekniken skapat nya möjligheter att göra information och service mer tillgängliga för alla. Enligt FN:s konvention om rättigheter för personer med funktionsnedsättning ska medlemsstaterna ”säkerställa att personer med

6 Se <http://spraakbanken.gu.se/korp/>

funktionsnedsättning får tillgång på samma villkor som andra till information och kommunikation" (artikel 9), även "till tv-program, film, teater och annan kulturell verksamhet i tillgänglig form" (artikel 30).

Detta har lett till ökade krav i Radio- och tv-lagen (2010:696) och i sändningstillståndet för Sveriges television (SVT) att tillgängliggöra tv-utbudet via undertextning, uppläst textremsa och teckenspråkstolkning. Liknande krav ställs på public service-kanalerna i andra nordiska och europeiska länder (Nordens välfärdscenter 2011, ITU 2011). Det uppsatta målet i SVT:s sändningstillstånd är att samtliga program på svenska ska textas med hög kvalitet, även de direktsända. För att uppnå det målet behöver SVT utveckla ny teknik för tv-textning baserad på taligenkänning (SOU 2012:59), vilket uppmärksammades i den svenska Public service-utredningen som presenterades våren 2013 (SOU 2012:59).

Behovet av nationell språkinfrastruktur

För att utveckla sådan teknik krävs tillgång till språkresurser i form av taldatabaser, lexikon och andra språkdata som kräver stora arbetsinsatser och kostnader för att ta fram. Hittills har skapandet av sådana språkresurser främst inriktats på ett begränsat antal kommersiellt intressanta språk, i synnerhet engelska. Taligenkänningstekniken är därför i dag enbart förbehållen tv-företag som verkar inom stora språkområden, som exempelvis BBC. I Sverige och andra nordiska länder har marknadskrafterna inte varit starka nog att driva på utvecklingen i tillräcklig utsträckning (Domeij 2013). För de flesta europeiska språk saknas enligt den undersökning som redovisas i META-NETs vitböcker (Borin m.fl. 2012a; se avsnittet om META-NORD ovan) många av de grundläggande språkresurser som ses som nödvändiga för att stimulera forskningen och teknikutvecklingen på området. På sikt utgör det ett hot mot många av de europeiska språken, i synnerhet minoritetsspråken. I en rapport om tillgängliga medier från Nordens välfärdscenter (2011) nämns avsaknad av språkteknologisk utveckling som en av tre huvudorsaker till att alla inte har tillgång till tv i de nordiska länderna.

Tillgången till språkresurser är alltså en central språkpolitisk fråga med avgörande betydelse för språkutvecklingen och tillgängligheten till information och service i det digitala samhället (Domeij m.fl. 2011). När det kommersiella intresset inte är tillräckligt stort behöver samhället ta ansvar för att skapa den språkliga infrastruktur som krävs för att tillgodose behovet av språkresurser och driva på teknikutvecklingen, så att vi i Sverige

och andra nordiska länder kan ta del av de nya tekniska möjligheterna på någorlunda lika villkor som andra länder med mer kommersiellt gångbara språk. I Norge har man redan etablerat en nationell språkbank med uppgift att tillgodose behovet av språklig infrastruktur. I Sverige har frågan under senare år uppmärksammats av regeringen.

I *It i människans tjänst – en digital agenda för Sverige* (2011) presenterade den svenska regeringen en strategi med målet att göra Sverige bäst i världen på att utnyttja digitaliseringens möjligheter. Särskilt intressant ur ett språkligt perspektiv är betoningen på e-tillgänglighet och infrastruktur för att stimulera teknikutvecklingen. Där uppmärksammas också behovet av en nationell språkinfrastruktur för att stimulera utvecklingen av digitala produkter och tjänster. Regeringen skriver att ”det långsiktiga etablerandet av en nationell språkbank med språkdatabaser främjar utvecklingen av teknik vilket gagnar språken i Sverige och ökar tillgängligheten till information för alla” (s. 37). Därmed finns frågan om en nationell språkinfrastruktur på den politiska agendan.

Formuleringen rimmar väl med regeringens satsning på öppna data för att stimulera utvecklingen av digitala tjänster genom att uppmuntra myndigheter och andra organisationer att fritt dela med sig av befintliga data för utveckling av digitala tjänster och produkter. Öppna data kan förenklat sägas vara information som är tillgänglig utan inskränkningar i form av kostnader eller immaterialrättsliga hinder, vilket är av stor betydelse för utvecklingen av språkteknologi, inte minst för mindre resurssvaga språk (Nørstebø Moshagen och Langgård 2011).

Förslag till nationell infrastruktur för språken i Sverige

I enlighet med den digitala agendan gav regeringen i uppdrag åt Institutet för språk och folkminnen, som Språkrådet är en del av, att ta fram ett beredningsunderlag för att utveckla formerna för drift och samordning för en nationell språkdatabank (ku2011/860/KA). Det har ingått i uppdraget att ta fram en behovsanalys och kostnadsberäkning för projektet, undersöka vilka tillgängliga tjänster som kan användas samt lämna förslag till långsiktig finansiering av språkdatabanken. En viktig utgångspunkt är att databanken bör vara öppen för aktörer som vill använda den som bas för produkt- och tjänsteutveckling.

I februari 2012 lämnades underlaget till kulturdepartementet i form av rapporten *Infrastruktur för språken i Sverige – Förslag till nationell språkinfrastruktur för det digitala samhället* (2012). Det förslag som läggs

fram i rapporten innebär i korthet att en nationell språkinfrastruktur för talbaserade tjänster stegvis etableras med utgångspunkt i behovet av direkttextning för tv. Rapporten hänvisar till att utvecklingen av taligenkänning nu står inför ett genombrott, vilket ger stora möjligheter att stimulera teknikutvecklingen och öka tillgängligheten till information och tjänster för alla, förutsatt att erforderlig infrastruktur snabbt kan utvecklas och göras tillgänglig för tjänsteutveckling. En första uppgift bör vara att ta fram taldatabaser för direkttextning av tv med siktet inställt på den långsiktiga uppgiften att se till att en bred och representativ taldatabas för svenska görs tillgänglig för tjänsteutveckling och hålls uppdaterad över tid.

Att utveckla en taldatabas för direkttextning av tv är ett konkret exempel på språkresursutveckling i samarbete med tjänsteutvecklare som öppnar möjligheter till ett långsiktigt samarbete, där SVT kan anpassa sin verksamhet till höjda tillgänglighetskrav samtidigt som SVT:s omfattande språkmaterial kan användas för utveckling av språkresurser och göras fritt tillgängliga för andra tjänsteutvecklare. Dörren bör öppnas även för andra intressenter som TV4. Ett sådant samarbete skapar också möjligheter till långsiktig finansiering av verksamheten.

Rapporten föreslog vidare att Språkrådet vid Institutet för språk och folkminnen får i uppdrag att samordna och följa upp arbetet med språklig infrastruktur på nationell nivå. I uppdraget ingår att ta fram en nationell strategi för arbetet med språklig infrastruktur, fortlöpande undersöka språkresursbehovet, inventera språkresurstillgången och förmedla befintliga språkresurser till teknikutvecklare.

Språkresurssituationen i Sverige

Som nämnts i avsnitt 1 är läget i Sverige nu det att språkresurssituationen är fragmenterad. Flera nödvändiga basresurser saknas, bl.a. representativa tal- och textdatabaser för svenska. En del språkresurser som tagits fram i olika forskningsprojekt finns tillgängliga för forskningsändamål på de forskningsinstitutioner som ansvarat för projektet. Dessa språkresurser underhålls vanligen inte efter att projektfinansieringen upphört.

Språkbanken i Göteborg bedriver som sagt en mer systematisk språkbanksverksamhet med inriktning på textresurser. Däremot saknas motsvarande verksamhet med inriktning på talresurser. Terminologicentrum TNC tillgängliggör terminologi via Rikstermbanken. Det finns också flera myndigheter som sitter på viktiga språkresurser, t.ex. Myndigheten för

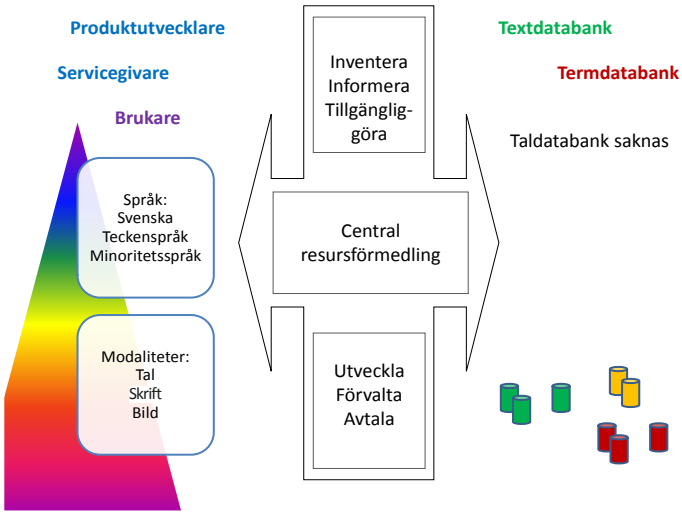
tillgängliga medier (MTM) och Institutet för språk och folkminnen dit Språkrådet hör.

Figur 3 ger en schematisk bild som utgångspunkt för överväganden om vad som skulle behöva göras för att förbättra språkresurssituationen i Sverige och stimulera produkt- och tjänsteutvecklingen. Högersidan i bilden illustrerar på ett förenklat sätt det faktum att det redan finns en del språkresurser i form av textdatabaser (grönt), termdatabaser (rött) och i taldatabaser (gult). För att stimulera utvecklingen av de tillgängliga tjänster (pyramiden på vänster sida) som ska komma brukare och servicegivare till del behöver efterfrågade språkresurser utvecklas och tillgängliggöras för produkt- och tjänsteutveckling.

Tre verksamheter behöver realiseras

Rapporten till regeringen föreslår att de tre verksamheter som skisseras i rektanglarna i mitten av figur 3 bör realiseras i syfte att utveckla och tillgängliggöra efterfrågade resurser. För det första behöver man göra fortlöpande inventeringar av resursutbudet. Man behöver ta reda på vilka resurser som finns på olika organisationer, vilken kvalitet de har (med hänsyn till standarder m.m.), var man hittar dem och hur de är tillgängliga. Detta gäller inte bara för svenska utan också för de nationella minoritetsspråken och det svenska teckenspråket.

För det andra behövs en central resursförmedling som hjälper tjänsteutvecklarna att få tag på befintliga resurser på ett smidigt sätt. Detta kan göras genom ett samordnat arbete mellan olika språkresursinnehavare och -förvaltare som ställer sina språkresurser till förfogande via den centrala resursförmedlingen. I anslutning till förmedlingsverksamheten behöver det finnas information och service av olika slag, som att ge svar på frågor om språkresurserna från tjänsteutvecklare som vill använda dem och att ge stöd till personer eller organisationer som vill tillgängliggöra sina resurser via resursförmedlingen. Det kan handla om hjälp med att hantera juridiska frågor, med att göra metabeskrivningar, med att hitta lämplig förvaltare och liknande.



Figur 3. Schematisk bild med förslag till funktioner för bl.a. inventering, förmedling och utveckling av språkresurser.

För det tredje måste det finnas ekonomiska möjligheter att se till att efterfrågade resurser utvecklas och sedan också underhålls så att de kan hållas uppdaterade och standardanpassade. De kan också behöva anpassas på andra sätt med hänsyn till tjänsteutvecklarnas behov. I samband med detta bör man också kunna hantera licenser. Som alternativ till att utveckla nya resurser kan redan befintliga men inlåsta resurser frigöras så att de fritt kan användas för tjänsteutveckling, vilket innebär att frågor om upphovsrätt och personuppgifter måste kunna hanteras.

Utöver ovan beskrivna verksamheter behövs språkbanksverksamhet för lagring, underhåll och distribution av olika typer av språkresurser. Som framgår av figurens övre högra del finns i dag en språkbank för textresurser, dvs. en textbank, i form av Språkbanken vid Göteborgs universitet. Det finns dessutom en termbank i form av Rikstermbanken på Terminologicentrum TNC (Rikstermbanken). Däremot saknas en talbank för underhåll och tillgängliggörande av talresurser, vilket är ett stort hinder för utveckling av talbaserade tjänster. För att direkttextning av tv och andra talbaserade tjänster ska kunna utvecklas behöver därför en språkbanksverksamhet med inriktning på tal etableras, en taldatabank. Detta bör rimligen göras i nära samarbete med KTH som har internationellt gångbar kompetens på talteknologiområdet.

Slutord och framåtblick

Som framgått av ovanstående beskrivning behöver det på ett övergripande plan finnas en funktion för samordning av olika verksamheter och organisationer som deltar i ett nationellt samarbete för utveckling och tillgängliggörande av språkresurser. I anslutning till denna behöver en nationell strategi utarbetas för hur arbetet kring språkresurser ska bedrivas i samverkan mellan olika parter.

En nödvändig grundförutsättning för en god utveckling på språkteknologiområdet är att ett nödvändigt basutbud av språkresurser (en s.k. BLARK – *Basic LAnguage Resource Kit*) tas fram och tillgängliggörs för forskning och utveckling av språkteknologi genom finansiering från Vetenskapsrådet och andra statliga finansiärer av grundläggande infrastruktur. Utöver det behövs riktade satsningar på utveckling av efterfrågade språkresurser för tjänsteutveckling. Eftersom behovet av språklig infrastruktur för forskningen till stora delar överlappar med den infrastruktur som behövs för produkt- och tjänsteutveckling är det viktigt att arbetet med språkinfrastruktur för forskning respektive produktutveckling samordnas.

Efter att under en tid ha behandlat det ovan beskrivna förslaget i kulturdepartementet aviserade den svenska regeringen i juni 2013 att den avser att starta ett pilotprojekt för att arbeta vidare på basis av förslaget. I december 2013 beviljade också Vetenskapsrådet en ansökan om svenskt deltagande i CLARIN vilket gör att förutsättningarna nu ser goda ut för att en mera organiserad och samordnad verksamhet kring en nationell språkinfrastruktur ska komma till stånd.

Nyckelord: språkteknologi, språkbank, språkinfrastruktur, språkresurser, META-NORD, SWE-CLARIN.

Lars Borin är professor i språkvetenskaplig databehandling, föreståndare för Språkbanken och Centre for Language Technology, Göteborgs universitet. Nationell samordnare för SWE-CLARIN, den svenska CLARIN-noden. Språkteknolog med bakgrund i jämförande språkvetenskap (slaviska och finsk-ugriska språk) och med ett aktivt forskningsintresse inom e-vetenskap, i skärningspunkten mellan språkteknologi och en rad humanistiska och samhällsvetenskapliga discipliner.

Rickard Domeij är doktor i datorlingvistik och arbetar som språkvårdare på Språkrådet i Sverige med språk och it som specialområde.

Summary

Over the last few years there has been an increased appreciation of the need for high-quality language resources (LR) and language technology (LT) for Swedish and the other languages of Sweden. This issue has two facets: (1) Safeguarding the achievements of an old and strong Swedish LT research tradition; and (2) ensuring that the results of this research are realized in products and services that will benefit Swedish society and its denizens. This paper describes some connected efforts in this direction, which manifest themselves both in academic initiatives aiming at surveying, standardizing, upgrading and interlinking Swedish LRs and LT (the META-NORD and SWE-CLARIN R&D collaborations), and in a recent political initiative by the Swedish Ministry of Culture with the goal of establishing a national LR/LT infrastructure supporting the development of language-aware products and services.

Litteratur

Borin, Lars, M.D., Brandt, J. Edlund, J. Lindh, M. Parkvall, 2012a: *Svenska språket i den digitala tidsåldern. The Swedish Language in the Digital Age.* META-NETs vitboksserie. Berlin: Springer. <http://www.meta-net.eu/whitepapers/volumes/swedish-sv?set_language=sv>.

Borin, Lars, M. Forsberg, J. Roxendal, 2012b: Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*, s. 474–478. Istanbul: ELRA.

Borin, Lars och Richard Johansson, 2014: Kulturomik: Att spana efter språkliga och kulturella förändringar i digitala textarkiv. Populärvetenskaplig framställning för bloggen *Historia i en digital värld*. <http://digihist.se/5-metoder-inom-digital-historia/fordjupning-kulturomik-att-spana-efter-sprakliga-och-kulturella-forandringar-i-digitala-textarkiv/>

Moshagen, Sjur N. and Per Langgård (eds.), 2011: *Visibility and Availability of LT Resources. Proceedings of the NODALIDA 2011 workshop*. NEALT Proceedings Series Vol 13. Riga: NEALT.

Domeij, Rickard, T. Breivik, J. Halskov, S. Kirchmeier-Andersen, P. Langgård, S. Nørstebø Moshagen (red.), 2011: *Språkteknologi för ökad tillgänglighet – vilka möjligheter finns?* Utarbetad av Astin för Nätverket

för de nordiska språknämnderna. Linköping: Linköping university electronic press. ISBN 978-91-7393-094-9.

Domeij, Rickard, 2013: Behovet av språkinfrastruktur för direkttextad tv. I: Breivik m.fl. (red.), *Språk i Norden 2013*.

Infrastruktur för språken i Sverige. Förslag till nationell språkinfrastruktur för det digitala samhället., 2012: Beredningsunderlag för att utveckla formerna för en nationell språkdatabank enligt regeringsuppdrag Ku2011/860/KA.

It i människans tjänst – en digital agenda för Sverige, 2011: Rapport från *Näringsdepartementet* N2011.12.

ITU, *2011*: *Making television accessible.* International Telecommunication Union och G3ICT.

Michel, Jean-Baptiste, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, The Google Books Team, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, E. Lieberman Aiden, 2011: Quantitative analysis of culture using millions of digitized books. *Science*, 331.

Nordens välfärdscenter, 2011: *Medier för alla? Tillgänglighet till tv i Danmark, Finland, Norge och Sverige.* Rapport från ett möte mellan tre nordiska samarbetsorgan för funktionshindersorganisationer.

Pedersen, Bolette Sandford, L. Borin, M. Forsberg, K. Lindén, H. Orav, E. Rögnvaldsson, 2012: Linking and validating Nordic and Baltic wordnets. *Proceedings of 6th International Global Wordnet Conference*. s. 254-260.

Pedersen, Bolette Sandford, L. Borin, M. Forsberg, N. Kahusk, K. Lindén, J. Niemi, N. Nisbeth, L. Nygaard, H. Orav, E. Rögnvaldsson, M. Seaton, K. Vider, K. Voionmaa, 2013: Nordic and Baltic wordnets aligned and compared through "WordTies". *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013). NEALT Proceedings Series* 16, s. 147–162.

Proposition *2012*/13:164. *Bildning och tillgänglighet – radio och tv i allmänhetens tjänst 2014–2019*.

Skadina, Inguna, A. Vasiljevs, L. Borin, K. De Smedt, K. Lindén, E. Rögnvaldsson, 2011: META-NORD: Towards sharing of language resources in Nordic and Baltic countries. *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, s. 107–114. Chiang Mai, Thailand: Asian Federation of Natural Language Processing.

SOU (*2012:59*) *Nya villkor för public service*. Kulturdepartementet, Public service-kommittén (Ku 2011:05).