Vetenskapsrådet

Database Infrastructure Committee
DISC

Svensk språkteknologi - existerande forskningsinfrastruktur och framtida behov

Februari 2007

Förord

Styrelsen för Database Infrastructure Committee (DISC) beslutade vid sammanträde 2006-05-10 att uppdra åt undertecknad att initiera arbetet med en kartläggning av förhållandena kring databaser för forskning inom humaniora. Vid följande möte 2006-09-13 preciserades uppdraget till att gälla

- En inventering av befintliga stora databaser inom humaniora.
- En kartläggning av framtida behov av stora databaser inom humanistiska ämnesområden.
- En rimlig avgränsning av datamaterial vid myndigheter och organisationer utanför universiteten som bör ligga inom DISC:s ansvarsområde.

Denna kartläggning av forskningsinfrastruktur inom hela det humanistiska fältet delades upp i två delar, en med inriktning på humaniora i hela dess bredd och en med inriktning på språkteknologi. Denna rapport beskriver situationen i fråga om forskningsinfrastruktur inom språkteknologiområdet.

Uppdraget redovisades den xx februari inför DISC:s styrelse.

Umeå xx februari 2007

Eva Strangert

INNEHÅLL

INLEDNING

Uppdraget från Database Infrastructure Committee (DISC) var att inom ramen för en bredare kartläggning av forskningsinfrastruktur inom humaniora beskriva läget – befintliga resurser såväl som behovet i ett femårsperspektiv – inom språkteknologiområdet. Med forskningsinfrastruktur avses här databaser, i första hand digitala datasamlingar, men också befintligt icke-digitalt material, av stort värde for forskningen nationellt och internationellt.

En utgångspunkt var en tidigare rapport från april 2005[1] som sammanfattar resultatet av ett regeringsuppdrag till Vetenskapsrådet att ”… som underlag för sitt arbete med vetenskaplig infrastruktur … kartlägga den nationella infrastrukturen inom humaniora och samhällsvetenskap…” Denna kartläggning begränsades till befintlig infrastruktur, främst digitala forskningsdatabaser, men även icke-digitaliserade materialsamlingar ingick. För humanioras del framgick att cirka hälften av de redovisade databaserna fanns inom områdena språkvetenskap, inklusive språkteknologi. Den nu aktuella kartläggningen är att se som en komplettering och utvidgning av denna tidigare inventering. Fokus ligger dock på språkteknologi och språkvetenskap med anknytning till språkteknologi.

Ytterligare motiv för en kartläggning av språkteknologiområdet står att finna i Vetenskapsrådets långsiktiga plan för forskningsinfrastruktur från juni 2006[2] utarbetad inom Kommittén för forskningens infrastrukturer (KFI). Där (sid 20) pekas språkteknologi ut som ett område ”i en unik situation i och med att ett välfungerande samarbete har vuxit fram mellan svenska universitet och tekniska högskolor”. KFI ser ”ett stort behov att se över den nationella infrastrukturen för språkteknologi och verka för samordning av databaser och analysverktyg. I första hand avvaktar KFI resultat från DISC:s utredning om digitaliserade databaser inom humaniora.”

Underlag för kartläggningen samlades in via ett rundfrågebrev till institutioner och andra enheter med språkteknologisk forskning samt genom annan tillgänglig dokumentation och genom samtal med företrädare för området. Deltagande i ett seminarium i språkteknologi i Göteborg i oktober 2006 med ett trettiotal deltagare från samtliga nordiska länder bidrog vidare till värdefull belysning av det nordiska samarbetet kring språkteknologisk infrastruktur.

Merle Horne, professor i allmän språkvetenskap vid Lunds universitet, har i inledningsfasen bidragit med synpunkter på rundskrivelsen och i slutfasen med värdefulla synpunkter på och kompletteringar av manuskriptet.

Därefter har rapporten gått ut på remiss till de institutioner/enheter som besvarat den och inkomna synpunkter har inarbetats.

I det följande presenteras inledningsvis en sammanfattning av arbetet och de viktigaste slutsatserna. Därefter ges en bakgrund – en beskrivning av språkteknologin och dess frågeställningar och metoder samt relationen till andra vetenskaper. En summarisk beskrivning av språkteknologins roll i det språkpolitiska arbetet ingår också. Därefter redovisas resultatet av kartläggningen med beskrivningar av forskningsmiljöerna och den

[1] ”Om forskningens infrastruktur inom humaniora och samhällsvetenskap i Sverige”, http://www.vr.se/download/18.4e0e826e108d316a03080004002/Om%20forskningens%20inf...rastrukturer_KFI-HS_slutrapport.pdf

[2] ”Vetenskapsrådets guide till infrastrukturen”, http://www.vr.se/download/18.7bea596910e36c19cbc80001735/Rapport+14.2006.pdf

forskning som bedrivs, de tillgängliga resurserna och de framtida behoven av infrastrukturella resurser. Avslutningsvis framläggs slutsatserna av arbetet, där möjligheten till nya strategiska satsningar ingår som ett naturligt nästa steg för att säkra svensk språkteknologi för framtiden.

SAMMANFATTNING OCH SLUTSATSER

Kartläggningen visar bilden av ett antal starka forskningsmiljöer förenade i ett nätverk med bred samverkan inom ramen för olika forskningsprogram och projekt sedan lång tid tillbaka. Utöver sådan nationell samverkan förekommer nära knytningar till andra nordiska länder, till EU-området och till övriga världen. Svenska språkteknologer är, och har sedan lång tid tillbaka varit, involverade i nordiska och europeiska samarbetsprojekt och relationerna till omvärlden är i övrigt mångfasetterade och starka.

Dagens läge, som det framgår av kartläggningen är att det finns relativt stora kvantiteter forskningsdata samlade i Sverige. Databaserna är dels sådana som utvecklats i Sverige med svenskt material, dels databaser med annat språkligt material utvecklat på andra håll i världen. Att annat språkligt material än svenska utnyttjas beror inte bara på att sådant behövs för översättningsteknologier och mer allmänt för jämförelser mellan svenska och andra språk. Det beror också på att det svenska materialet är högst otillräckligt för att utföra många olika typer av undersökningar på ett vetenskapligt tillfredsställande sätt.

För att svensk språkteknologisk forskning ska kunna hålla jämna steg med och fortsatt bidra till den internationella utvecklingen på området krävs nya och omfattande text-och taldatabaser omfattande många olika genrer. Dessutom behövs nödvändiga redskap för att hantera databaserna. Verktyg måste utvecklas för sökning, för uppmärkning av tal och text – i största möjliga utsträckning med automatiska metoder – för utveckling av lexikon och för olika typer av analyser av tal och text. Här visar kartläggningen på stora och varierande behov reflekterande de olika inriktningarna som finns inom språkteknologin.

Vid sidan om de behov som framkommer i svaren från de olika forskargrupperna har emellertid svenska språkteknologer också gemensamt utarbetat en plan för att ta fram en nationell infrastrukturell resurs. Den ska innehålla omfattande tal- och textkorpusar samt verktyg för uppbyggnad, inmatning, uppmärkning, sökning, uppdatering och underhåll av databaserna. Konceptet är en bred, väl dokumenterad och allmänt tillgänglig basresurs som efter behov ska kunna vidareutvecklas och kompletteras med mer material och nya redskap i samklang med utvecklingen på området.

Tillgängligheten är central. Många problem existerar idag på grund av bristande tillgänglighet i fråga om befintliga databaser. Värdefulla material kan t ex inte användas till följd av olika slag av restriktioner. Det gäller inte bara svenska utan också utländska databaser. Men också när nya databaser ska byggas finns hinder. Hit hör upphovsrättsliga regler, copyright-bestämmelser, som är ett mycket stort problem. I den mån de kan lösas medför de oftast stora kostnader. Här finns paralleller till Personuppgiftslagen (PUL) och dess konsekvenser för den samhällsvetenskapliga och medicinska forskningen.

Det är således en uppenbar framtidsinriktning och samsyn som kommer till uttryck i den nämnda planen för uppbyggnad av en gemensam språkteknologisk basresurs och syftet att göra den allmänt tillgänglig och utvecklingsbar. Den språkteknologiska forskningen har som mål att agera i en internationell kontext och att föra utvecklingen framåt. Förutsättningarna för en positiv och kraftfull utveckling finns också. Det svenska försörjningsläget är gott genom den nationella forskarskola (Graduate School of Language Technology) som för närvarande hyser ett 40-tal doktorander. I denna utveckling har de infrastrukturella resurserna en nyckelroll.

Den språkteknologiska forskningen och betydelsen av god språkteknologisk infrastruktur har också relevans i ett samhälleligt perspektiv och är en förutsättning för många olika tillämpningar inom såväl myndigheter som industri. Sådan mer tillämpat utvecklingsinriktad forskning pågår också. Tillämpningsaspekten är likaså den centrala när språkteknologins betydelse i ett språkpolitiskt perspektiv framhävs, bl a i utredningen *Mål i mun.* Språkteknologin ses där som ett medel att slå vakt om det svenska språket och ett medel att göra medborgarna delaktiga i det svenska samhället.

Beroendet och vikten av goda infrastrukturella resurser framkommer dessutom tydligt i olika strategiska dokument under senare år – bl a IT-kommissionens rapport *Svensk språkteknologi -vadan och varthän?*, den tioårsplan för att utveckla språkteknologin i Norden, *Språkvis*, som tagits fram på uppdrag av Nordiska Ministerrådet liksom EU-dokumentet *Human Language Technology for Europe* som har till målsättning att genom språkteknologi överbrygga språkgränser inom Europa. CLARIN (Common Language Resources and Technologies Infrastructure), ett EU-projekt med mål att tillgängliggöra språkteknologiska resurser via webben, är ytterligare ett uttryck för nödvändigheten av allmänt tillgängliga och lätt åtkomliga databaser och språkteknologiska redskap. Kartläggningen visar att i den utvecklingen har svensk språkteknologi en roll att spela. Det förutsätter fortsatta strategiska satsningar, där denna kartläggning kan ses som ett första steg i att säkra den svenska språkteknologiska forskningen för framtiden.

BAKGRUND

Språkteknologiområdet och relationerna till lingvistik, datavetenskap och teknik

Som utgångspunkt för kartläggningen ges en kortfattad beskrivning av forskningsområdet . Den är hämtad ur dokumentet ”Språkteknologi för Sverige” (Bilaga 1), som ger en något mer utförlig bild och som rekommenderas som läsning i sin helhet.

”Språkteknologi är informationsteknologi som utvecklas för att hantera mänskligt språk i dess olika former. Till *talteknologin* räknas tekniker som konverterar mellan tal och text eller som känner igen en individuell röst. Tekniker som hanterar språket i dess skrivna former, t.ex. för stavnings- och grammatikkontroll, kan kallas *textteknologi*, men en betydande del av språkteknologin handlar om *tekniker som är gemensamma för tal och text*. Detta gäller exempelvis tekniker för att komma åt information som är lagrad i databaser eller tekniker för översättning mellan olika språk.

Ett språkteknologiskt system är i bästa fall uppbyggt med generell programvara och språkspecifika data. Ofta kan en och samma grundteknik användas för många olika språk och de tekniker som utvecklats för de stora språken kan med viss framgång tillämpas på de mindre. Förutsättningen är då givetvis att nödvändiga data för språket i fråga finns till hands i den form som tekniken kräver. Att få fram dessa data kan dock vara svårare än det i förstone kan verka. God täckning av ett språk kräver databaser med tiotals miljoner ord. Det handlar oftast inte heller om rådata i form av text eller inspelat tal utan om analyserade data som märkts upp med språklig information, eller bearbetade data i form av lexikon. Sådana analyser och bearbetningar kräver utvecklade verktyg och språkteknologisk expertis. Därtill kommer att mycket av de data man vill använda är upphovsrättsligt skyddat.

Tekniken är inte heller helt oberoende av språk. Även mellan så närbesläktade språk som engelska och svenska finns stora skillnader. Svenskans sammansättningar är hopskrivna till ett ord medan engelskans i regel är uppdelade på flera. Denna enkla skillnad innebär t.ex. att om man söker med en ordbaserad sökmotor som Google på ordet ’stuga’ så får man ingen träff i dokument som bara innehåller ordet ’sommarstuga’ medan en motsvarande sökning med det engelska ’cottage’ ger träffar vare sig det står ’summer’ framför eller något annat. Svenskan har ordtoner, vilket innebär att ett talsyntessystem ska kunna uttala ordet ’tomten’ olika beroende på om det syftar på ett markområde eller en jultomte. Omvänt bör ett taligenkänningssystem kunna höra skillnaden, men sådana prosodiska fenomen är svåra att hantera på grund av dialektvariationer och brytning. Större språk som engelska och franska, vilka saknar ordtoner, kan delvis bortse från sådana problem.

En viktig trend i dagens språkteknologi är utveckling av metoder som utifrån stora datamängder automatiskt skapar språkmodeller, s.k. maskininlärning. Maskininlärning fick sitt genombrott inom språkteknologin som en kraftfull metod inom taligenkänning. En vägledande princip för den typen av system har varit att ”mer data är bättre data”. Detta innebär att man kunnat se förbättringar av systemen för varje gång man utvidgat sin databas. Maskininlärning tillämpas i dag på många språkteknologiska problem inklusive översättning. Men varken taligenkänning eller s.k. statistisk maskinöversättning är perfekta teknologier. Det finns därför ett behov av metodutveckling till vilken svensk forskning har kapacitet att bidra.”

Språkteknologin är ett tvärvetenskapligt forskningsområde som alltså inkluderar såväl tal som text. (En alternativ beteckning för textteknologi är datalingvistik.) Dessutom ingår i viss utsträckning också bild, dels inom ramen för teckenspråksforskning, dels som gester och mimik inom ramen för multimodal interaktion.

Språkteknologin vilar på tre ben, lingvistik (allmän språkvetenskap och fonetik), datavetenskap och teknik i förening. Det innebär att det finns mycket nära relationer mellan språkteknologin och respektive lingvistik, datavetenskap och teknik. Forskare i lingvistikämnena arbetar ofta i nära samarbete med språkteknologer och språkteknologer finns vid såväl lingvistik- som datavetenskapliga institutioner samt vid de tekniska högskolorna. Centrala forskningsfrågor kring det mänskliga språket förenar språkteknologer och lingvister och utvecklingen av språk- och kommunikationsmodeller utgör ett gemensamt

fält där också datavetare med intresse för språk kommer in. Här finns också kopplingar till kognitionsvetenskap och AI-området.

Den här beskrivna tvärvetenskapligheten har varit avgörande för att i den aktuella kartläggningen inkludera också databaser inom lingvistik. Den nära relationen mellan de olika områdena framgår också av att den forskning som skett inom ramen för olika språkteknologisatsningar från 1990 och framåt inkluderat såväl språkteknologer/datalingvister som språkvetare och datavetare. (En kort redogörelse för dessa satsningar i form av speciella språkteknologiprogram ingår under rubriken ”Forskningsmiljöer”.)

Utöver sådana mer grundforskningsinriktade frågeställningar rymmer språkteknologin en stor potential för tillämpningar. Det kan gälla effektiv informationsåtkomst av olika slag i sökmotorer eller genom utveckling av dialogsystem. Det kan också gälla kommunikation över språkgränserna som i t ex maskinöversättning, eller flerspråkig informationssökning och det kan gälla integrerade system av tal, text och visuella signaler för kommunikation, däribland hjälpmedelsutveckling (se Bilaga 1).

Språkteknologins roll i det språkpolitiska arbetet

Språkteknologi intar en nyckelroll i utredningen *Mål i mun* (2002)[3] som hade till syfte att överblicka språksituationen i dagens och morgondagens Sverige. Den proposition som den resulterade (*Bästa språket – en samlad svensk språkpolitik*, 2005)[4] i formulerade de språkpolitiska målen: att svenska är huvudspråk i Sverige, att det ska vara ett komplett och samhällsbärande språk och att alla ska ha rätt till ett väl fungerande språk för att inte hamna utanför språkliga gemenskaper. Det innebär att svenska ska vara användbart i alla sammanhang och måste kunna erbjuda sina användare ett rikt utbud av kommunikationsmöjligheter. De som lever i Sverige ska vidare inte bara ha rätt till svenska utan också till sitt modersmål och det ska vara möjligt att kommunicera på minoritetsspråk och på främmande språk.

I allt detta ses språkteknologi som ett verksamt medel att uppnå målen. För att vi inte ska halka efter språk med bättre tillgång på språkteknologiska resurser måste medborgarna ha tillgång till välfungerande språkteknologiska tillämpningar, vilket kräver forskning och utveckling för att möta behoven. Utvecklingen av språkteknologin har också betydelse för människor med olika slag av funktionshinder, där olika teknologiska redskap kan utnyttjas som kommunikationshjälpmedel. Dessutom spelar språkteknologin en roll i utvecklingen av den sk 24-timmarsmyndigheten, dvs myndigheternas nätverksamhet för information till medborgarna.

Språkteknologin är således viktig i det språkpolitiska och språkvårdande arbetet. Institutet för språk och folkminnen, som bildades 2006 och där den tidigare Svenska språknämnden, nu Språkrådet[5], ingår som en del, har i sin instruktion att den särskilt ska främja språkteknologiskt arbete.

[3] www.regeringen.se/sb/d/108/a/1443
[4] www.regeringen.se/sb/d/5359/a/50761
[5] www.sprakradet.se. Språkrådet har en hemsida med samlad information om språkteknolgi, www.sprakteknologi.se

På likartat sätt ses språkteknologins betydelse för övriga nordiska språk. Riktlinjerna för en gemensam nordisk språkpolitik antogs nyligen av Nordiska Rådet (*Deklaration om nordisk språkpolitik*, 2006)[6] med förhoppningen att Norden ska vara föregångsregion i det språkpolitiska arbetet. De nordiska länderna, bl a Norge och Danmark, har också satt igång ett arbete med att ta fram nationella, språk- och forskningspolitiska handlingsplaner, där språkteknologin ingår som en del.

De nordiska språknämnderna samarbetar nära om språkteknologiska frågor. Det sker i Arbetsgruppen för språkteknologi och språkvård i Norden, som finansieras av Nordens Språkråd (som ingår i Nordiska Ministerrådet). På uppdrag av Nordiska Ministerrådet har också en tioårsplan för att utveckla språkteknologin i Norden tagits fram. Den föreligger i form av en sk vismansrapport (*Språkvis*, 2006)[7].

Men också utvecklingen av den europeiska gemenskapen förutsätter en fungerande språkpolitik. Det måste gå att upprätthålla en god och effektiv kommunikation över språkgränserna och i det har språkteknologin med metoder för maskinöversättning och flerspråkig teknik av andra slag en roll att spela. Det framgår bl a av Europeiska Ministerrådets rapport *En ny ramstrategi för flerspråkighet*[8] från 2005, liksom av EU-dokumentet *Human Language Technologies for Europe*[9] från 2006.

En mer utförlig beskrivning av språkteknologins roll inom ramen för språkpolitiska ställningstaganden i Sverige, Norden och EU ges i Bilaga 2 från det svenska Språkrådet.

[6] http://www.norden.org/sagsarkiv/sk/sag_vis.asp?vis=2&id=335
[7] Rapporten kan laddas ner från www.sprakteknologi.se under ”dokument”.
[8] eurlex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!DocNumber&lg=sv&type_doc=COMfinal&an_doc=2005&nu_doc=596
[9] www.tc-star.org/pubblicazioni/D17_HLT_ENG.pdf

KARTLÄGGNINGEN

Den infrastruktur som står i fokus för kartläggningen är i första hand digitala databaser, text- och talkorpusar, som används inom språkteknologisk och språkvetenskaplig forskning. Här ingår också databasverktyg, dvs de redskap som ingår som en integrerad del i strukturering en och den övriga hanteringen av forskningsmaterialet, program för uppmärkning och sökning i materialet samt dessutom nödvändiga lexikon- och grammatikresurser. I andra hand ingår också icke-digitala samlingar av text och tal. Hit hör bl a bandupptagningar (audio eller video) med uppläst tal, dialoger och andra typer av samtal, uppmärkt helt eller delvis, som efter digitalisering utgör en rik potential för språkvetenskaplig/språkteknologisk forskning. Dessutom ingår eventuell annan utrustning som behövs som komplement i verksamheten som rör databaser.

Resultaten av kartläggningen föreligger som svar på den rundskrivelse som gick ut till forskargrupper och institutioner vid svenska lärosäten som bedriver språkteknologisk/ språkvetenskaplig forskning.

I rundskrivelsen ombads institutionerna beskriva situationen så att den gav en så god bild som möjligt av verksamheten, inklusive olika former av samarbeten nationellt och internationellt. För de institutioner där språkvetenskaplig forskning med egna databaser finns parallellt med specifikt språkteknologisk forskning fanns alternativet att antingen lämna svaren separat eller tillsammans för de två verksamhetsgrenarna. Det har i de flesta fall lett till att det kommit in två eller flera svar från samma institution, ofta med viss överlappning dem emellan.

Den bild som framträder genom svaren på rundskrivelsen måste betraktas som en väl täckande beskrivning av situationen inom området. Svar har med ett par undantag inkommit från samtliga de tillskrivna forskargrupperna/institutionerna och återfinns i Bilaga 3 ordnade efter universitets-/högskoleort. Denna ordning har valts med hänsyn till den tidigare nämnda, och ofta förekommande överlappningen mellan språkteknologin och lingvistiken, även om den långt ifrån alltid är geografiskt bunden. Bilaga 4 innehåller rundskrivelsen med sändlista.

Frågorna i rundskrivelsen var indelade i tre avsnitt med uppgifter centrerade kring

- Den språkteknologiska forskning som sker/planeras inom den egna institutionen/enheten och som kräver tillgång till stora databaser och databasverktyg
- De forskningsdatabaser (inklusive verktyg) som finns vid institutionen/enheten
- Det framtida behovet av resurser för planering, utveckling, uppbyggnad, drift och avveckling av databaser samt ev annan utrustning som behövs som komplement i verksamheten som rör databaser.

Utöver de infordrade svaren har också annan dokumentation utnyttjats för att ge en så god beskrivning av området som möjligt. Det svenska Språkrådet ombads ge sin syn på språkteknologin utifrån det språkpolitiska perspektiv, som är dess huvuduppgift (se ovan och Bilaga 2). Konsultationer och samtal med olika nyckelpersoner för att få fram ett helhetsperspektiv på språkteknologin har också givit tillgång till dokument som utgjort underlag för denna rapport. Deltagande i ett nordiskt seminarium under hösten om språkteknologisk forskning gav också värdefull information.

Fortsättningsvis beskrivs nu de olika forskningsmiljöerna och deras inbördes relationer och samverkan i ett nordiskt och internationellt perspektiv. Därefter ges en sammanställning över

tillgängliga databaser och verktyg samt de förutsedda behoven som de framgår ur underlagen från institutionerna/enheterna och övrig tillgänglig dokumentation.

Forskningsmiljöerna

De avgivna svaren ger en bild av en inom området väl samverkande grupp av sinsemellan starka forskningsmiljöer.

Den talteknologiska forskningen bedrivs huvudsakligen vid Centrum för talteknologi (CTT) vid KTH, även om viss talteknologiskt inriktad forskning också förekommer inom språkteknologigruppen i Göteborg. Forskare vid KTH har varit pionjärer i utvecklingen av talsyntes och automatisk taligenkänning och i verksamheten vid CTT utnyttjas sådana system i utvecklingen på områden som dialogsystem, modellering av naturlig talad dialog, vidareutveckling av automatisk taligenkänning och talarkaraktärisering, bl a för talarverifiering. CTT har under en tioårsperiod finansierats genom stöd från VINNOVA, men kommer då stödet upphör att leva vidare som centrumbildning inom KTH. Institutionen har en stark ställning internationellt och lång erfarenhet av samarbete inom såväl Europa som övriga världen. Härom vittnar bl a den listning av pågående samarbeten, varav flera EU-projekt, såväl som avslutade samverkansprojet under en lång tidsperiod som ingår i underlaget. Gruppen har också lång erfarenhet av industriellt samarbete (och avknoppning av företag. Den talteknologiskt inriktade forskningen omfattar för närvarande ett 25-tal personer.

Inom den textbaserade språkteknologiforskningen pågår en rikt förgrenad verksamhet med såväl grundläggande forskning som tillämpningar inom områden som bl a informationssökning och informationsåtkomst, maskinöversättning, läs-och-skrivprocessen, datorstödd undervisning och språkstöd i olika former. Forskningen bedrivs framförallt vid universiteten i Göteborg, Linköping, Lund, Stockholm, Uppsala och Växjö, men också vid högskolorna i Skövde och Borås, samt dessutom vid KTH (Språkteknologigruppen vid NADA) och Chalmers samt inom Swedish Institute of Computer Science (SICS), ett fristående institut med nära knytning till universitetsforskningen. Göteborg har en väl förgrenad verksamhet på flera institutioner förenad i en samverkansgrupp benämnd Centre for Language Technology (CLT) som man vill utveckla till en mer formaliserad organisation. I den ingår bl a Språkbanken, en verksamhet som går tillbaka till den uppbyggnad av språkkorpusar som initierades på 1960-talet. Gruppens ansökan om Linnéstöd 2005 beviljades trots utmärkta omdömen inte av Vetenskapsrådet, men ny ansökan planeras inför nästa ansökningsomgång. Ett nytt initiativ i Uppsala är Engelska parkens informationsteknologiska centrum för forskning och utbildning i humaniora, språk och teologi (EPARIT), där den gemensamma nämnaren är användning av stora språkdatabaser liksom humanistiska databaser av andra slag. Den textlingvistiska projektverksamheten är, liksom den talteknologiska, omfattande, såväl på det nationella planet som internationellt (främst inom EU). Industriellt samarbete (bl a SCANIA, Microsoft och IBM) förekommer också och eftersträvas i ökande grad. De olika grupperna varierar i storlek mellan 4-5 och 30 personer och det totala antalet involverade forskare på textområdet (exklusive lingvistik) uppgår till ca 80.

Mellan de olika grupperna – och det gäller hela språkteknologiområdet, tal såväl som text – sker samverkan på flera plan. En del av grunden för denna samverkan kan utan tvivel spåras tillbaka till de satsningar som under 1990-talet gjorts på svensk språkteknologi i ett antal program finansierade av HSFR i samverkan med STU respektive NUTEK och från 2000 i ett

nyligen avslutat program med finansiering från VINNOVA. Bakgrunden och en sammanställning av de olika programmen ges i Bilaga 5.

En annan bidragande faktor är den gemensamma nationella forskarskolan i språkteknologi, Graduate School of Language Technology (GSLT). Förutom Göteborgs universitet (Humanistiska fakulteten) som är värdhögskola medverkar nio andra svenska universitet/högskolor[10] (se också Bilaga 6). Forskarskolan, tillsammans med andra satsningar på språkteknologi har sedan starten 2001 varit en betydelsefull faktor i utvecklingen av språkteknologin. Den möjliggör ett stort utbud av kurser och högt kvalificerad handledning, därtill viss tillgång på databaser och andra resurser för hantering av databaser. Forskarskolan bidrar också till att säkra tillgången på kompetens fortsättningsvis. För närvarande finns över 40 doktorander vid skolan. Den samverkan som forskarskolan inneburit har således bidragit till koordinering av och samarbete kring forskningen i övrigt liksom samordnade forskningsansökningar för gemensamma nationella resurser.

Nordiskt och internationellt samarbete

Samverkan mellan Sverige och övriga nordiska länder sker, delvis informellt och delvis genom speciella satsningar. Från 2000 till 2004 pågick ett samfinansierat nordiskt forskningsuppdrag inom ramen för Nordiska Ministerrådets verksamhet. Det ledde bl a till initieringen av en nordisk forskarskola, NGSLT, och webb-baserade språkteknologiska dokumentationscentrum i de olika nordiska länderna med Språkteknologi.se som svensk representant. På terminologiområdet finns också ett nordiskt nätverk, Nordtermnet. Sedan ett antal år anordnas den nordiska språkteknologikonferensen Nodalida vartannat år mellan de nordiska länderna. För att ytterligare stärka och bredda samarbetet inom norden och kringliggande länder bildades hösten 2006 organisationen Northern European Association for Language Technology (NEALT).

Mycket av samarbetet sker, som tidigare nämnts inom ramen för det språkpolitiska arbetet. Det sker bl a genom Nordiska Språkrådet som består av språkvårdsmyndigheterna inom Norden och inom ramen för det nämnda femårsuppdraget från Nordiska Ministerrådet. Syftet är att samordna och öka kommunikationen mellan språkteknologin länderna emellan samt att säkerställa tillgänglighet och återanvändning av forskningsresultat, textsamlingar och verktyg. Språkteknologin ses som en nyckelfaktor för att språken ska överleva och utvecklas. För detta har också Ministerrådet bidragit med viss projektfinansiering.

Att svensk språkteknologi utmärks av väl utvecklade samarbeten också inom Europa och med världen i övrigt har redan nämnts. Svenska språkteknologer har bred erfarenhet av EU-samarbete och deltar i flera nu pågående projekt. Det finns också en lång tradition av samarbete, främst inom talteknologi, med institutioner i bl a USA (t ex MIT) och Japan (ATR). Beskrivningen av nu existerande databaser i ett senare avsnitt (”Tillgängliga resurser i form av databaser och databasverktyg”) ger ytterligare bidrag till beskrivningen av svensk språkteknologi i internationellt perspektiv.

[10] Se forskarskolans hemsida, http://www.gslt.hum.gu.se/(sv)/nslp.html

Pågående och planerad forskning som kräver tillgång till stora databaser och databasverktyg

Här följer fortsättningsvis en översiktlig beskrivning av den språkteknologiska forskningen i Sverige. Den inkluderar såväl tal- som textteknologi, som dock inte är oberoende av varandra. I båda fallen behövs tillgång till grammatik, lexikon mm, dvs språklig information som till stora delar är gemensam för text och tal och som kan betraktas som basen för båda dessa modaliteter. På samma sätt som texten speglar en bakomliggande språklig struktur gäller det också för talet. Tal-, text- och språkstruktur är med andra ord integrerade system i mänsklig kommunikation. I ett dialogsystem baserat på talsyntes och taligenkänning behövs t ex tillgång till grammatik och lexikon. Och samma typ av kunskap behövs också för forskning baserad på text, t ex maskinöversättning. Tal- och textteknologi existerar därför inte bara jämsides utan korsbefruktar också varandra genom det gemensamma intresset för språkstrukturen. Det innebär också ett nära samarbete mellan de två områdena.

Här finns som nämnts också i många fall nära relationer till kognitionsvetenskapliga frågeställningar och till artificiell intelligens. Språkteknologin kan där utnyttjas som ett redskap att modellera mänsklig språkförmåga och språkligt processande parallellt med att samma metoder driver den teknologiska utvecklingen. Det huvudsakliga redskapet här, maskininlärning, kräver storskaliga resurser i form av data och vare sig forskningen gäller tal eller text är beroendet av tillgång till stora databaser detsamma.

Uppbyggnad och utveckling av databaserna är en väsentlig del av forskningen, som kräver sina egna metoder och där metodutvecklingen är en integrerad del av databasuppbyggnaden. Det gäller för både tal- och textdata och mycket av forskningsresurserna idag går åt till att utveckla metoderna för att göra data tillgängliga i en sådan form att användbarheten ökar. Mycket forskningsarbete läggs t ex ned på att driva det krävande arbetet med annotering – uppmärkning – av materialet mot en ökande grad av automatisering. Även i övrigt, vare sig det handlar om grundforskningen eller den mer tillämpade språkteknologin har utvecklingen av metodologi och grundläggande algoritmer central betydelse inom området.

Databaser inom språkteknologi, liksom databaser inom medicin och samhällsvetenskap utgör det grundläggande materialet på basis av vilket man kan komma fram till svaren på de stora vetenskapliga frågor man ställer. I detta fall gäller frågorna det unikt mänskliga och komplexa system som gör det möjligt att kommunicera i form av text, tal och tecken, ett system som vi ännu bara delvis förstår oss på. Databaserna behövs för modellering och simulering av tal och språk i olika situationer. Till exempel, för att kunna förstå hur en optimal multimodal människa-maskininteraktion skulle kunna designas, måste man utveckla metoder för simulering av många typer av tänkbara kommunikationssituationer. Utgångspunkten måste då vara data från naturliga interaktioner människa-människa i jämförbara kontexter. En vison är att tillämpa språkteknologiska modeller i den mänskliga interaktionen mellan människor och med maskiner utan att dessa upplevs som hinder utan som naturliga hjälpmedel i det framtida samhället.

Databaserna utgör grunden i korpuslingvistik och överhuvudtaget i empiriskt baserad forskning som rör språklig struktur, vare sig det gäller tal eller text. Frågeställningar kring det talade eller skrivna språket bearbetas och analyseras med utgångspunkt i olika slags materialsamlingar, databaser med olika typer av texter och talsituationer. Här sker forskningen ofta i samverkan mellan språkteknologin och den teoretiskt och empiriskt inriktade lingvistiken. Lingvistiken kan dra nytta av språkteknologiskt utvecklingsarbete kring

automatiska metoder för syntaktisk[11] och morfologisk[12] analys och många av de databaser som utvecklats av allmänlingvister och fonetiker, primärt för egna syften, utgör resurser som utnyttjas också inom språkteknologin.

Forskningen kring tillämpningen av språkteknologin sker på flera olika fronter men kan grupperas inom två mer övergripande områden, informationsåtkomst samt kommunikation över språkgränser. Härutöver förekommer forskning inom en rad andra områden.

Informationsåtkomst via sökmotorer utgör ett forskningsområde med stark svensk representation. Här bidrar Sverige till utvecklingen av mer avancerade system, sk frågebesvarande system, än de idag förekommande som bygger på fritextsökning. Sökningen i frågebaserade system utgår ifrån en direkt fråga från användaren, vilket kräver såväl grammatisk som semantisk analys för att kunna förstå frågan samt för att kunna generera relevanta svar. Angränsande områden är sökning och extraktion av information i strukturerade databaser, liksom klustring och sammanfattning av information. Huvuddelen av svensk språkteknologisk forskning om informationsåtkomst är inriktad på text, speciellt mer avancerad frågebaserad sökning, som kräver sk ”data mining”[13], men även system för sökning av talad information är under utveckling, t ex information i TV - och radioprogram. Ytterligare ett stort delområde utgörs av forskning kring dialogsystem. Sådana system kan ses som en vidareutveckling av textbaserade frågebesvarande system som utvecklats så att de idag också kan fungera med tal som input såväl som output. Dialogsystem baserade på tal utvecklas och förbättras i takt med att kunskap om metoder för automatisk taligenkänning samt talsyntes ökar.

Det andra större området, kommunikation över språkgränser, utgör också en rikt förgrenad verksamhet i Sverige, främst i Stockholm (Stockholms universitet), Uppsala och Linköping. Hit hör automatiska system för översättning mellan språk, sk maskinöversättning. Hittills har forskningen huvudsakligen gällt översättning mellan svenska och närbesläktade språk som engelska eller tyska, men idag finns en stark inriktning mot flerspråkig översättning som kan involvera också sinsemellan mer olika språk. Här bidrar inte minst den utmaning som uppkommit genom EU med dokument som kräver översättning mellan många olika språk. Samtidigt gynnas översättningsforskningen av existensen av EU-dokument med samma innehåll på många olika språk, eftersom de kan utnyttjas för parallella analyser av strukturen i de olika språken. På detta område finns planer på storskaliga analyser av många olika språkliga aspekter. Andra planerade satsningar gäller maskinöversättning till orientaliska och slaviska språk som liksom all maskinöversättning förutsätter parallella korpusar, dvs korpusar från två eller flera språk som länkas till varandra på olika språkliga nivåer.

Övrig språkteknologisk forskning gäller bl a språkstöd av olika slag och utbildning. Forskningen här är inriktad på system för språkgranskning, stavning såväl som grammatik. Satsningar görs på teknologi för automatisk upptäckt av problem hos vissa målgrupper och forskning inriktad på barns och ungdomars skrivande har möjliggjort utveckling av system som också är anpassade till barn och ungdomar. Här spelar sk inlärarkorpusar en viktig roll. Men området utvecklas för närvarande också med ett vidare perspektiv på språkutbildning i form av t ex datorstödd språkinlärning, som kan gälla såväl text som tal. Svensk talteknologi

[11] Syntaktisk analys innebär att strukturen/relationerna mellan orden i en sats eller mening beskrivs. Det sker ofta i form av en hierarkiskt ordnad trädstruktur.
[12] Morfologisk analys innebär att strukturen i orden får en analys i termer av stam och olika slag av ändelser eller, i fråga om sammansatta ord, en analys av de olika ordleden i sammansättningen.
[13] Automatiska tekniker för analys av stora mängder data för att hitta betydelsekopplingar i text/tal.

ligger vidare väl framme när det gäller integrerade system av tal och bild (ansiktsrörelser), sk talande huvuden som redan förekommer i kommersiella system som utnyttjas av bl a hörselhandikappade.

Tillgängliga resurser

Databaser och databasverktyg

Genomgående gäller att institutionerna/enheterna, utöver databaser utvecklade av de egna forskarna eller i samverkan med andra (huvudsakligen svenska) kolleger, förfogar över databaser som skapats av andra forskare. En stor del av dessa databaser som köpts eller på annat sätt tillförts verksamheten är utländska och representerar en stor mångfald främmande språk. Sådana databaser kan ha ett speciellt värde i studier där svenska behöver kontrasteras mot andra språk med annan struktur, men i många fall är det bristen på motsvarande svenskt material som ligger bakom. Kostnaden för att införskaffa materialet är ibland ringa, men kan i andra fall vara avsevärd och belopp stora som 2500 USD förekommer. Medlemskap i organisationer som LDC (Linguistic Data Consortium) i USA bidrar till att hålla nere kostnaderna, men en stor del av de inköpta resurserna (också från ELDA (Evaluations and Language resources Distribution Agency i Paris) är ändå ett surrogat för svenska databaser som ännu inte existerar.

Likväl förfogar svensk språkteknologi (och språkvetenskap) över relativt stora egna svenska material. Det gäller tal såväl som text, och i viss utsträckning också bild (bl a teckenkommunikationsdata). En del av materialet är som nämnts utvecklat i samverkan mellan olika universitet och inom ramen för gemensamma forskningsprojekt. Till denna kategori hör t ex Swediadatabasen med insamlat och uppmärkt tal från sammanlagt 1400 talare från alla delar av Sverige och svensktalande områden i Finland. Hit hör också den översättningskorpus som huvudsakligen byggts upp i Linköping men med bidrag från såväl gruppen i Uppsala som Språkbanken. Ytterligare ett exempel är SUC (Stockholm Umeå Corpus), utvecklad inom HSFR-NUTEK-språkteknologiramprogrammet under 1990-talet, som bygger på ett omfattande arbete med texter (1,1 milj ord totalt) representerande olika genrer. Det speciellt värdefulla med SUC är dess storlek och framförallt att den i sin helhet är uppmärkt morfosyntaktiskt[14]och detta gör den ytterst användbar och därför oerhört väl utnyttjad inom såväl lingvistisk som språkteknologi.

I andra fall har databaser kommit till genom större EU-projekt med svenska deltagare. Här finns exempel som svenska SpeechDat (inspelningar av 6000 personer från olika delar av Sverige) och SpeCon (550 personer), som innehåller taldata insamlade av forskare vid CTT och där det svenska materialet utgör delmängder av de samlade databaserna med många olika språk representerade. PAROLE-databasen, där den svenska delen samlats in vid Språkdata i Göteborg, utgör en liknande mångspråkig resurs när det gäller textdata. Lexikala databaser som Svenskt OrdNät och Svenskt RamNät (uppbyggda inom lingvistik i Uppsala) har kopplingar till liknande arbeten inom organisationer som Global WordNet respektive Global Fram/Net.

Härutöver finns en stor mängd databasmaterial, större och mindre. Generellt kan sägas att medan de databaser som utvecklats och utnyttjas inom språkteknologi (naturligen) är digitala, så förekommer inom lingvistik såväl digitalt som icke-digitaliserat material. Den digitala

[14] Det innebär att strukturen hos varje enskilt ord såväl som ordens struktur inom satser och meningar framgår.

taldatabasen IRIS (Invandrarröster i Sverige) innehåller inspelningar av ca 150 språk (främst invandrar- och andra minoritetsspråk i Sverige), med i vissa fall upp till 12 talare per språk. Många av talarna har dessutom producerat motsvarande material på bruten svenska. Till större icke-digitaliserade material – vissa delar har dock digitaliserats – hör Göteborg Spoken Language Corpus (GSLC), insamlad under ca 30 år och med för närvarande totalt 1,5 miljoner ord. Vartefter den digitaliseras i ökande grad, något som planeras, kan den väntas få stort värde inom den språkteknologiska forskningen. Många exempel på sådana icke-digitaliserade material beskrivs i rundfrågesvaret från språkteknologigruppen i Uppsala (Bilaga 2k). Sammanställningen där, som täcker både digitala och icke-digitala korpusar och lexikon vid hela den språkvetenskapliga fakulteten i Uppsala ger sannolikt en bild som stämmer även för andra universitet. Alla svenska universitet har dock inte på långt när den rika mångfald av språk som finns representerade i Uppsala.

Rundfrågesvaren redovisar således en stor mängd digitala databaser, text så väl som tal inom en mängd genrer. Till dessa hör även inlärardatabaser, en växande grupp av databaser med svenska (eller andra språk som målspråk) som är en värdefull resurs vad gäller bl a utveckling av olika språkstöd och datorstödd undervisning. Ett par exempel är de databaser som utvecklats inom två projekt om svenska som andraspråk vid Stockholms universitet och liknande databaser (vad gäller såväl skrift som tal) byggs för närvarande upp vid flera av universiteten och högskolorna.

Här ska också Språkbanken i Göteborg, och den i Språkbanken inkluderade Litteraturbanken, nämnas. Båda är att betrakta som nationella resurser av stort värde, även om de huvudsakligen finansieras lokalt eller i fråga om Litteraturbanken med medel från bl a Svenska Akademien. Språkbanken förfogar över mycket stora materialsamlingar – svenska och parallella textkorpusar och svenska elektroniska lexikon liksom språkverktyg av olika slag – och tillhandahåller dem för forskare.

Parallellt med databaserna har svenska språkteknologer utvecklat verktyg som behövs för uppbyggnad och hantering av databaserna. Det gäller bl a verktyg för sökning och annotering av såväl text som tal. Hit hör t ex visualiseringsverktyg för den akustiska informationen i taldatabaser och utveckling av automatiska metoder som hjälp vid annotering av tal (som är en mycket tidskrävande process). Dessutom finns olika slag av översättnings- och lexikonverktyg och utveckling av automatiska metoder för taggning (syntaktisk uppmärkning av text) i arbetet med att åstadkomma en trädbank för svenska. Andra resurser under utveckling är ett svenskt ordnät (för en beskrivning av betydelserelationer såsom synonymi, antonymi liksom andra betydelserelationer mellan ord) som behövs för att utveckla mer avancerade sökmetoder (t ex text-mining) för svenska.

Nyttjande, tillgänglighet och standardisering

Nyttjandefrekvensen varierar stort mellan olika databaser. I svaren anges allt ifrån att nyttjandet endast gäller enstaka forskare eller forskargrupp till i vissa fall mycket frekvent utnyttjande både inom den egna gruppen och utanför den. Variationen här beror till stor del på syftet när databasen tillkom. Smärre databaser insamlade för specifika projektändamål har givetvis oftast mindre nyttjandegrad än databaser skapade just för att vara resurser med bredare användning. En databas som i detta senare sammanhang framstår värd att nämna särskilt är SUC som genom sin rika uppbyggnad har en väl dokumenterad hög nyttjandegrad och vid användning. Till samma kategori kan också den tidigare nämnda PAROLE-korpusen räknas, liksom när det gäller talat språk, Swedia-databasen. Europeiska infrastrukturprojekt, t

ex SpeechDat och SpeeCon, där svenska ingår som ett av många språk och som bidrar till en användning som också går över nationsgränserna, har också hög nyttjandegrad.

Emellertid är ofta tillgängligheten ett problem. Databaserna är ofta insamlade under restriktioner av olika slag och därmed inte öppna för användning för många av de syften där de skulle göra stor nytta. Det gäller också t ex data som är lagrade i Språkbanken. Materialet där tillgängliggörs i stor utsträckning via Språkbankens hemsida, men det får i de flesta fall inte lämnas ut för vidare bearbetning eller användning, något som är en absolut nödvändighet för den språkteknologiska forskningen. Denna restriktion har bl a med upphovsrättsliga regler att göra och svensk språkteknologi delar denna problematik med språkteknologin i andra länder. Tillgång till och utnyttjande av värdefulla material hindras av copyright-bestämmelser, som i den mån de kan lösas oftast medför mycket stora kostnader.

Svenska språkteknologer är väl medvetna om vikten av nationella och internationella standarder. Till det bidrar bl a gemensamma projekt inom EU samt den internationalisering i övrigt som kännetecknar svensk språkteknologi, liksom språkvetenskapen generellt. Standardiseringsfrågor står också högt på dagordningen inom den nordiska språkteknologin och den nybildade organisationen North European Association for Language Technology (NEALT). Strävan att följa internationellt fastställda standarder är således stark och manifesteras bl a i de samarbeten som förekommer. Databasarbetet vid CTT inom ramen för olika EU-samverkansprojekt följer t ex av helt naturliga skäl internationell standard. I detta arbete spelar också Språkbanken en central roll. Bland annat har medel sökts från DISC (stora databaser) för att anpassa resurserna till internationella standarder. Bilaga 7 ger en översikt av standarder och pågående standardiseringsarbete vad gäller text och tal.

Kostnader och finansiering

Kostnaderna för uppbyggnad, utveckling och drift av existerande databaser inom Sverige samt databasverktyg av olika slag uppgår till mycket stora belopp. Det finns få uppgifter om kostnader, i kronor räknat, angivna i de inkomna svaren. Dock anges från CTT summor på 1,5 milj och 2,5 miljoner kronor i inspelningskostnad för i sammanhanget relativt begränsade taldatabaser och en summa om (sannolikt) mer än 40 miljoner svenska kronor totalt för uppbyggnaden av Språkbanken. Men att kostnaderna i övrigt är betydande och uppgår till många tiotals miljoner sammantaget framgår av de uppgifter om databasernas storlek och innehåll som ges. Svårigheterna att precisera beloppen står att finna i bakgrunden till och finansieringen av resurserna.

Databaserna och verktygen har till stor del skapats inom ramen för olika forskningsprojekt bekostade med nationella offentliga medel (statliga såväl som icke-statliga finansiärer), huvudsakligen från Vetenskapsrådet och tidigare HSFR, Riksbankens Jubileumsfond, VINNOVA och tidigare NUTEK och STU, SIDA samt andra finansiärer som t ex Knut och Alice Wallenbergs Stiftelse och Svenska Akademien. Ytterligare en stor del av finansieringen har åstadkommits genom medel för EU-projekt, där i viss omfattning projektens syfte varit att ta fram jämförbara databaser för olika språk (t ex PAROLE, SpeechDat och SpeeCon). Mindre summor har också kommit från Nordiska Ministerrådet för just databasutveckling (standarder för trädbanker[15]). Ytterligare andra finansieringskällor är fakultets- och institutionsanslag och då oftast till enskilda forskare som tid för forskning.

[15] Datasamlingar med syntaktiskt analyserad text i form av trädstrukturer.

Framtida behov

Databaser och verktyg

Trots att de tillgängliga databaserna innehåller relativt omfattande material, täcks inte behoven på långt när. I de ingivna svaren understryks snarare behoven av mycket utökade infrastrukturresurser mot bakgrund av den internationella utvecklingen av språkteknologin och Sveriges roll i denna utveckling. Svensk språkteknologi har hittills varit framgångsrik och varit med och drivit denna utveckling, inte minst inom talteknologin. Eftersom forskningen i stor utsträckning är datadriven behövs stora mänger nya data, om Sverige fortsättningsvis ska kunna bidra med internationellt intressant forskning.

Följande lista är inte fullständig men ger en grov bild av de olika behoven i ett femårsperspektiv:

- Omfattande svenska textkorpusar representerande många olika textstilar, totalt minst 100 milj ord
- Textkorpusar för minoritetsspråk i Sverige
- En stor svensk trädbank baserad på texter med åtminstone 1 milj löpord
- Lexikon av olika slag, bl a lexikal semantisk korpus för t ex informationsextraktion
- Redskap för storskalig annotering av textmaterial för utveckling av lexikon och trädbanker
- Dialogkorpusar (människa-människa och människa-maskin)
- Taldatabank med olika typer av svenskt tal representerade, däribland 1000 timmar spontantal (berättande, samtal, etc)
- Multimodala databaser (kombinationer av ljud och bild för analys av samtal)
- Redskap för annotering av tal (bl a semi-automatisk segmentering och annotering av akustiska data)
- Översättningskorpusar för svenska och andra språk (europeiska men även turkiska, slaviska språk etc) och som går i båda riktningarna (från eller till svenska)
- Utbildnings- och inlärarkorpusar (bl a felkorpusar)
- Redskap för storskalig annotering och bearbetning av parallella texter på olika språk
- Modernisering, förädling, homogenisering av befintligt korpusmaterial
- Digitalisering av för forskningen värdefulla analoga data (ännu ej digitaliserade delar av Göteborg Spoken Language Corpus)
- Utrustning för datainsamling (bl a HD videoutrustning), bearbetning (bl a digitalisering anpassad till akustiska analyser) och lagring av korpusar (bl a servrar med stor minneskapacitet)
- Driftskostnader

Den samlade bilden visar omfattande behov av större och bättre annoterade text- och taldatabaser och resurser i form av lexikon än de som finns för närvarande. Det behövs också videodatabaser och metoder för att hantera simultan ljud- och bildinformation. I första hand krävs resurser för svenska, där bl a en stor svensk trädbank framhålls som en helt nödvändig språkteknologisk resurs. Men det räcker inte med svenska. För att utveckla språkoberoende metoder och för översättningsforskning krävs också material som täcker andra språk och även svenska i kombination med andra språk (inklusive minoritetsspråk i Sverige) i form av parallellkorpusar.

De beräknade kostnaderna i kronor/personår som anges i svaren är höga. Till exempel behövs översättningskorpusar till ett beräknat belopp av minst 10 miljoner kronor. Utvecklingen av en svensk trädbank beräknas kosta 10-30 milj kronor och en svensk talspråkskorpus ca 25 milj kronor. Detta är emellertid bara en del av de belopp som anges.

Det som framförallt bidrar till de höga kostnaderna är arbetet med uppmärkning, vare sig det gäller text eller tal. Den tid det tar att märka upp spontana taldata anges t ex till 50 gånger den inspelade tiden. 1000 timmar inspelat tal (ett uppskattat behov för forskningen inriktad på det talade språket) skulle då kräva 50 000 timmar för uppmärkning (i den mån det inte är möjligt att nedbringa tiden genom än mer avancerade automatiska metoder än vad som idag finns tillgängliga).

Inom den samlade svenska språkteknologin finns dock en stor samsyn vad gäller såväl vilka slags resurser som behövs samt också hur de ska åstadkommas. Delar av vad som tas upp i de svar som inkommit från de olika institutionerna/enheterna ryms inom ramen för den nämnda planeringsansökan till VR[16] som beviljades medel i november 2006. Ansökan beskriver behovet av en *nationell* infrastrukturresurs i flera olika delar. Där aviseras ett behov av dels omfattande svenska tal- och textkorpusar med stor variation i tal- och textstilar, en Svensk Nationell Korpus (SNK)[17], dels verktyg för uppbyggnad, inmatning, uppmärkning, sökning, uppdatering och underhåll av databaser (en sk BLARK = Basic Language Resource Kit for Swedish).

De komponenter som denna basresurs (BLARK) bör innehålla är dels insamlade språkliga data (både råmaterial och annoterat material) och verktyg. Följande listade komponenter är hämtade direkt från ansökan:

- corpora and speech databases (monolingual and multilingual)
- terminology resources, ontologies
- lexical resources (monolingual and multilingual)
- grammatical resources (e.g. formal morphological and syntactic descriptions)
- basic tools (e.g. part-of-speech taggers[18], lemmatizers[19], parsers[20], text-to-speech modules, speech alignment and annotation tools, speech recognizers, named entity taggers, parallel corpus sentence and word aligners)
- reference resources (“gold standards”) for evaluation of annotation formats and tools
- metadata and application programming interfaces (APIs) for accessing, searching and combining components

De två delarna – basresursen och den nationella korpusen – behövs båda. De överlappar och kompletterar varandra.

Den nämnda samsynen kommer till uttryck i att denna ansökan är utformad gemensamt av huvuddelen av de olika noderna i det nätverk som tillsammans bildar svensk språkteknologi

[16] *En infrastruktur för svensk språkteknologi*, planeringsansökan med Lars Borin som koordinator, se http://vrproj.vr.se.
[17] Storleken på existerande nationella korpusar för andra språk varierar. Några exempel: danska (30 milj ord), engelska/British National Corpus (100 milj ord), finska (180 milj ord), tjeckiska (> 200 milj ord).
[18] Dataverktyg som taggar, dvs märker upp orden som substantiv, adjektiv, verb, adverb, konjunktion, preposition, osv.
[19] Dataverktyg som identifierar ett ords grundformer samt respektive ordklass.
[20] Dataverktyg som etiketterar orden i en sats med grammatiska roller t.ex. olika satsdelar (subjekt, predikat, objekt mm).

och som bl a hålls ihop genom forskarskolan GSLT och ett sedan lång tid tillbaka utvecklat samarbete inom olika forskningsprojekt och de tidigare nämnda språkteknologiprogrammen.

Dokumentation, standardisering, öppenhet, spridning och lagring

Utvecklingen och användningen av databaser ställer krav som de svenska språkteknologierna är väl medvetna om och också är beredda att uppfylla. Databaserna ska vara väldokumenterade och standardiserade enligt gemensamma nordiska/internationella överenskommelser. Detta är något som svenska spåkteknologer redan praktiserar i stor utsträckning (se Bilaga 7). Databaserna ska också vara öppna och enkelt tillgängliga för forskningen via Internet och webbgränssnitt. För detta krävs internationellt nätverkande.

Som tidigare nämnts innebär emellertid tillgängligheten ett problem. Upphovsrättsliga bestämmelser lägger hinder i vägen för att utnyttja mycket av det textmaterial som behövs inom den språkteknologiska forskningen. Korpusuppbyggnad har hittills varit förenad med stora ansträngningar för att få tillgång till det önskade materialet. Ofta har det också inneburit stora kostnader att köpa loss copyright-skyddat material. Problemet är stort inte bara i Sverige utan även inom Norden och i övriga världen. Det har nyligen tagits upp till diskussion inom ramen för det samarbete som sker mellan nordiska språkteknologer. En genomlysning av dessa problem och ansträngningar att undanröja hindren för forskningens vidkommande vore något som i högsta grad skulle gagna den vetenskapliga utvecklingen.

Det finns här också uppenbara paralleller till de tillgänglighetsproblem som finns när det gäller bilder och som regleras genom Bildkonst Upphovsrätt i Sverige (BUS). En annan parallell är personuppgiftslagen (PUL) och dess konsekvenser för samhällsvetenskaplig forskning.

Spridning och lagring hanteras idag huvudsakligen av forskarna eller forskargrupperna själva. Det främsta undantaget här är Språkbanken, som också rymmer material som utvecklats av andra forskare, t ex de inlärarkorpusar som tagits fram vid olika lingvistikinstitutioner. Möjligheten att utnyttja SND, när det etablerats, nämns dock som en möjlighet för framtiden t ex i samverkan med Språkbanken.

SLUTSATSER OCH STRATEGIER FÖR FRAMTIDEN

Språkteknologisk forskning håller hög klass både nationellt och internationellt. Den utgör också en viktig nationell resurs när det gäller utveckling av kunskap om det svenska språket i samverkan med andra språk i det nuvarande och framtida svenska samhället. För att göra det är svensk språkeknologi beroende av infrastrukturella resurser.

De inkomna svaren visar på ett stort behov av omfattande text- och taldatabaser. Till behovet av resurser hör också utveckling av redskap för databashanteringen, redskap som är minst lika betydelsefulla som databaserna i sig och som ska ses som integrerade komponenter i skapandet av databaserna. Sådana redskap behövs bl a för sökning och för annotering, dvs uppmärkning av tal och text, för utveckling av lexikon och uppbyggnad av trädbanker och för bearbetning av parallella texter på olika språk.

Utöver de behov som framkommer i svaren från de olika forskargrupperna har grupperna gemensamt utarbetat en preliminär plan och också beviljats planeringsmedel från Vetenskapsrådet (KFI) för uppbyggnaden av en nationell infrastrukturell resurs. Den ska vara allmänt tillgänglig för den språkteknologiska forskningen och innehålla omfattande tal- och textkorpusar samt verktyg för uppbyggnad, inmatning, uppmärkning, sökning, uppdatering och underhåll av databaserna.

Det finns tydliga viljeinriktningar inför framtiden för den språkteknologiska forskningen i de inkomna svaren liksom en uppenbar samsyn i hur målen ska uppnås. Här spelar de infrastrukturella resurserna en nyckelroll. Utan sådana resurser kan inte forskningen leva upp till de höga kvalitativa mål som man eftersträvar. Den nationella satsning som den nämnda planeringsansökan skisserar är således ett viktigt medel i denna målsättning.

Språkteknologins framåtsyftande verksamhet reflekteras också i övrigt i dess agerande på olika plan. Ett uttryck för detta är den nationella forskarskolan (GSLT) som med dess nuvarande 40-tal doktorander möjliggör en god försörjning av språkteknologer för lång tid framåt. Synen på framtiden kommer också fram i olika strategidokument under senare år. I en skrivelse till regeringen 2004 (Bilaga 8) påtalas behovet av en sammanhållen strategi för svensk språkteknologi mot bakgrund av olika utredningar, om maskinöversättning (NUTEK) och den tidigare nämnda utredningen *Mål i mun*. Dokumentet föregick den mer utförliga beskrivning av svensk språkteknologi som utformades på regeringens uppdrag och som ingår som Bilaga 1 (se inledningen).

Den språkteknologiska forskningen är vidare en förutsättning för många olika tillämpningar med relevans i ett samhälleligt perspektiv liksom för högteknologisk industriell utveckling. Målsättningar för den språkteknologiska forskningen i dessa avseenden formuleras i den rapport som gavs ut av IT-kommissionen[21] med ”förslag om angelägna insatser för språkteknologin som ett led i utvecklingen av ett informationssamhälle för alla”. Rapporten, som är ett resultat av en hearing med ledande intressenter och företrädare för det språkteknologiska området i Sverige, redovisar mångfalden av de möjligheter som språkteknologin erbjuder. En förutsättning är dock att den språkteknologiska forskningen stärks. Speciellt fokuseras behovet av en språkteknologisk infrastruktur.

[21] *Svensk språkteknologi – vadan och varthän*?
http://www.itkommissionen.se/dynamaster/file_archive/020219/32aeb58f417ec9c2714c01db7f0a7116/Rapport%20Spr%e5kteknologi.pdf

Detta behov är idag uppenbart. Det framgår tydligt i de svar som lämnats på rundskrivelsen Det är vidare tydligt framhävt i den tioårsplan for nordisk språkteknologi som nyligen tagits fram. Möjliga satsningar inom Europa diskuteras också. I dokumentet *Human Language Technology for Europe*[22] är utgångspunkten ”det mångkulturella Europa” och språkteknologins möjligheter att överbrygga språkgränser. I den europeiska roadmap för forskningsinfrastruktur som utarbetats av ESFRI (European Strategy Forum on Research Infrastructures)[23] ingår det språkteknologiska initiativet CLARIN (Common Language Resources and Technologies Infrastructure)[24] ett paneuropeiskt projekt med mål att tillgängliggöra språkteknologiska resurser via webben. I detta, liksom i andra internationella sammanhang, har svensk språkteknologi en roll att spela.

Det har inte legat i uppdraget att föreslå åtgärder. Det är emellertid naturligt att se kartläggningen som det första steget i en process för att skapa en framtida strategi för svensk språkteknologi och inte minst att tillgodose behovet av goda infrastrukturella resurser. Vidare steg i det arbetet måste ske i dialog med forskarna på området.

[22] www.tc-star.org/pubblicazioni/D17_HLT_ENG.pdf
[23] http://www.vr.se/ansvarsomraden/forskningspolitiskafragor/forskningsinfrastruktur/esfrieuropeanstrategyforumonresearchinfrastructures.4.7bea596910e36c19cbc8000927.html
[24] www.mpi.nl/clarin/pdf/clarinmission-1.pdf

BILAGOR

Språkteknologi för Sverige

Lars Ahrenberg, Rolf Carlsson, Olle Josephson
11 mars 2004

Med bidrag från Robin Cooper, Björn Granström, Jussi Karlgren, Joakim Nivre, Anna Sågvall Hein, Bengt Waernulf

Inledning

Språkteknologin spelar en betydande roll för utvecklingen av nya IT-tjänster och för utvecklingen av svenska språket. Syftet med denna text är att belysa språkteknologins betydelse för dagens samhälle och näringsliv. Vi vill också föreslå ett antal åtgärder som vi i dagsläget anser vara de viktigaste för utvecklingen av svensk språkteknologi.

Vad är språkteknologi?

Språkteknologi är informationsteknologi som utvecklas för att hantera mänskligt språk i dess olika former. Till *talteknologin* räknas tekniker som konverterar mellan tal och text eller som känner igen en individuell röst. Tekniker som hanterar språket i dess skrivna former, t.ex. för stavnings- och grammatikkontroll, kan kallas *textteknologi*, men en betydande del av språkteknologin handlar om *tekniker som är gemensamma för tal och text*. Detta gäller exempelvis tekniker för att komma åt information som är lagrad i databaser eller tekniker för översättning mellan olika språk.

Ett språkteknologiskt system är i bästa fall uppbyggt med generell programvara och språkspecifika data. Ofta kan en och samma grundteknik användas för många olika språk och de tekniker som utvecklats för de stora språken kan med viss framgång tillämpas på de mindre. Förutsättningen är då givetvis att nödvändiga data för språket i fråga finns till hands i den form som tekniken kräver. Att få fram dessa data kan dock vara svårare än det i förstone kan verka. God täckning av ett språk kräver databaser med tiotals miljoner ord. Det handlar oftast inte heller om rådata i form av text eller inspelat tal utan om analyserade data som märkts upp med språklig information, eller bearbetade data i form av lexikon. Sådana analyser och bearbetningar kräver utvecklade verktyg och språkteknologisk expertis. Därtill kommer att mycket av de data man vill använda är upphovsrättsligt skyddat.

Tekniken är inte heller helt oberoende av språk. Även mellan så närbesläktade språk som engelska och svenska finns stora skillnader. Svenskans sammansättningar är hopskrivna till ett ord medan engelskans i regel är uppdelade på flera. Denna enkla skillnad innebär t.ex. att om man söker med en ordbaserad sökmotor som Google på ordet ’stuga’ så får man ingen träff i dokument som bara innehåller ordet ’sommarstuga’ medan en motsvarande sökning med det engelska ’cottage’ ger träffar vare sig det står ’summer’ framför eller något annat. Svenskan har ordtoner, vilket innebär att ett talsyntessystem ska kunna uttala ordet ’tomten’ olika beroende på om det syftar på ett markområde eller en jultomte. Omvänt bör ett taligenkänningssystem kunna höra skillnaden, men sådana prosodiska fenomen är svåra att hantera på grund av dialektvariationer och brytning. Större språk som engelska och franska, vilka saknar ordtoner, kan delvis bortse från sådana problem.

En viktig trend i dagens språkteknologi är utveckling av metoder som utifrån stora datamängder automatiskt skapar språkmodeller, s.k. maskininlärning. Maskininlärning fick sitt genombrott inom språkteknologin som en kraftfull metod inom taligenkänning. En vägledande princip för den typen av system har varit att ”mer data är bättre data”. Detta innebär att man kunnat se förbättringar av

systemen för varje gång man utvidgat sin databas. Maskininlärning tillämpas i dag på många språkteknologiska problem inklusive översättning. Men varken taligenkänning eller s.k. statistisk maskinöversättning är perfekta teknologier. Det finns därför ett behov av metodutveckling till vilken svensk forskning har kapacitet att bidra.

Hur är läget i Sverige idag?

Sverige har genom tidigare satsningar byggt upp en framgångsrik forskning och forskarutbildning inom språkteknologi. Flera pågående satsningar som VINNOVAs språkteknologiprogram och kompetenscentrum CTT kommer dock att avslutas inom kort. Utvecklingen inom språkteknologin står dock inte stilla. Det är viktigt för Sverige och svenska språket att forskning och utveckling kan bibehållas på en hög nivå. Ambitionen borde vara (1) att Sverige kan delta i den internationella utvecklingen och fortsatt vara ledande på några viktiga områden, (2) att resurser för svenska utvecklas i en takt som i stort följer den internationella utvecklingen. Vi illustrerar i följande avsnitt vikten av forskning och språkresurser med flera exempel.

När det gäller tillgång på data är svenskan är relativt väl försedd med textdatabaser för skriftspråket, t.ex. Språkbankens vid Göteborgs universitet, men för att skapa bättre modeller krävs att data märks upp med information som visar en fras grammatiska funktion eller vad den står för. Dessa båda aspekter hör ofta nära samman. Namn kan förekomma som subjekt, objekt eller adverbial men namn på personer är oftare subjekt än exempelvis geografiska namn eller namn på händelser som ’Olympiaden i Aten’ eller ’Andra världskriget’. Vad gäller data som märkts upp på detta sätt släpar svenskan efter de större språken. För minoritetsspråken och många invandrarspråk är bristen på tillgängliga textdatabaser i sig ett problem.

För svenskans del är behovet av nya bearbetade databaser för talspråket stort, liksom för minoritetsspråken. För närvarande existerar ett antal mindre databaser framtagna för speciella behov: dialogdata för talbaserade dialogsystem, upplästa listor för träning av taligenkänningssystem, dialektdata från Swediaprojektet och en samtalsdatabas i Göteborg. Man bör här också observera att variationen i talad svenska är större än för det skrivna språket och att taligenkänningssystem inte kan användas av personer med utpräglad dialekt, stark brytning eller avvikande tal beroende på hörselskada eller dysartri.

För översättning och annan kommunikation över språkgränser utgör s.k. parallella korpusar eller översättningsminnen en viktig resurs, vilket vi utvecklar längre fram.

Sverige har god tillgång på kompetenta forskare inom språkteknologi. Språkteknologiska forskargrupper finns vid de flesta svenska universitet och på SICS (Swedish Institute of Computer Science) med olika profiler och representerande olika discipliner som datavetenskap, lingvistik eller svenska språket. Kompetenscentrum CTT med KTH som värd utgör ett särskilt initiativ för samarbete mellan akademisk forskning och industriell utveckling. Inom forskarskolan GSLT (National Swedish Graduate School of Language Technology) samarbetar dessa grupper för en nationell forskarutbildning. Nästan alla svenska grupper deltar nu, eller har tidigare deltagit, i europeiska och nordiska samarbetsprojekt inom sina specialiteter.

Vad kan språkteknologin åstadkomma?

Dagens system har en begränsad språkförmåga, i synnerhet om vi jämför med människans, men är ändå tillräckliga för att skapa tjänster som underlättar kommunikation och ökar produktiviteten i många arbeten. Vi illustrerar här med två olika tillämpningsområden där svensk forskning håller sig

väl framme: informationsåtkomst inklusive dialogsystem och kommunikation över språkgränser. Vi visar också på tillämpningar där språkteknologi är en insatsteknologi bland flera.

1. Informationsåtkomst

Det är omöjligt att beräkna hur mycket tid som vanliga svenskar i arbete, studier, politik, föreningsliv m.m. i dag ägnar åt att söka upp central information i intranät eller över Internet. Men det är mycket, och ju bättre söksystem man kan utveckla desto precisare och snabbare svar får den som söker informationen. Avancerade söksystem ökar alltså kunskapsutnyttjande och kunskapsnivå inom alla delar av samhället.

Även internationella standardsystem ger förbättrad effektivitet om de anpassas till svenska. Databastillverkaren Oracle vill kunna hänvisa sina kunder i Sverige, varibland nästan alla kommuner, till en svensk morfologi som kan integreras med deras produkter. De vill inte bygga en själv, eftersom svenska inte anses vara ett tillräckligt stort språk. De system som i dag kan köpas på marknaden är dyra. Principerna för svenskans morfologi är väl kända och implementerade på flera universitetsinstitutioner i Sverige. Dock finns det inte något standardsystem att hänvisa till som är tillgängligt för en rimlig kostnad.

1.1 Svar på direkta frågor

Dagens sökmotorer baseras på fritextsökning och ger träffar i form av rangordnade länkar till webbdokument. Man skriver ett ord eller en fras och får en lista på dokument där ordet eller frasen är nämnd. I många fall söker dock användaren svar på en specifik fråga och får ägna tid åt att klicka fram de dokument som verkar mest lovande och läsa igenom dem. Hittar man inte svaret ger man upp eller försöker ge nya söktermer.

En mer avancerad typ av söksystem, kallade frågebesvarande system, ger användaren möjlighet att uttrycka sitt informationsbehov med en fråga, t.ex. ”Hur många naturreservat finns det på Gotland?”, ”Vad säger svensk affärstidslagstiftning om öppettider?”. Systemet levererar då ett textutsnitt som innehåller svaret på frågan. Vid utvärderingar i USA har det visat sig att de system som ger bäst resultat vid sådana sökningar dels använder avancerade grammatiska analysatorer, s.k. parsers, dels utnyttjar stora semantiskt organiserade lexikon (ordnät). Ett ordnät innebär att orden inte sorterats i bokstavsordning, som i ett vanligt lexikon, utan efter betydelsenärhet. Orden ”naturreservat”, ”fridlysning”, ”växter” och ”miljöpolitik” ligger ganska nära varandra i databasen. Orden ”övertid”, ”arbetsmiljölagstiftning” och ”kollektivavtal” ligger också rätt nära varandra. Eftersom en ordbetydelse alltid ligger nära betydelsen av flera andra ord (”växter” ligger nära ”fridlysning” men också ”djur” eller ”blommor”) blir hela databasen ett gigantiskt nätverk. För svenska finns partiella ordnät, men inget som har en tillräckligt stor täckning.

1.2 Dialogsystem

Dialogsystem har primärt utvecklats som användargränssnitt mot strukturerade data i databaser, men är också relevanta som en utveckling av frågebesvarande system mot textsamlingar. I dag kan vi använda system för exempelvis nummerupplysning eller tidtabellupplysning via telefonen även på svenska. Dagens system har en begränsad funktionalitet; de är i regel utvecklade för ett enda ämnesområde och kräver en väl strukturerad dialog som i hög grad styrs av systemet genom att användaren frågas ut.

Begränsningarna i dagens system ligger till inte obetydlig del i taligenkänningen, som allmänt betraktas som en av flaskhalsarna i utvecklingen av nya talbaserade informationstjänster. Ett vanligt sätt att kringgå denna begränsning är att använda flera samtidiga interaktionskanaler, s.k. multimodala system, vilket innebär att talad inmatning kan kombineras med pekdon eller gestigenkänning och talad utmatning med text och grafik och, även, en animerad figur som utgör

användarens motpart i dialogen. Ett annat sätt är att ge systemen ökad dialogförmåga, exempelvis förmåga att reda ut uppkomna missförstånd och ge relevant hjälpinformation till användaren. En annan begränsning finns i de metoder som används för tolkningen av användarens yttranden, som är anpassade till det förhållandet att systemet bara känner till ett enda ämne som det kan tala om.

1.3 Informationssökning i taldatabaser

Mycket av den information vi har omkring oss föreligger i form av inspelat tal. Det kan gälla TV- och radioprogram, röstbrevlådor, inspelade diskussioner och föreläsningar. För att kunna ta tillvara all denna information behövs hjälp med att söka i stora och i många fall ostrukturerade taldata. För att utveckla sådana system behövs nya akustiska sökmetoder kombinerade med stor vokabulärigenkän-ning av svenska. Dessa saknas idag. För att denna typ av tjänster krävs uppmärkta och representativa taldatabaser samt lexikala resurser.

2. Kommunikation över språkgränserna

2.1 Maskinöversättning och datorstödd översättning

Ett klassiskt tillämpningsområde för språkteknologi är översättning. Man bör här skilja på automatiska översättningssystem och interaktiva översättningsstöd som hjälper en översättare att producera redigeringsfärdiga texter. De senare systemen kallas i dag ofta översättningsminnen eftersom den centrala dataresursen i dem består av tidigare översättningar där varje mening eller stycke parats ihop med motsvarande mening eller stycke i originalet. Sådana översättningsminnen är allmänt använda vid översättning av brukstexter, t.ex. produktinformation och manualer. Den kanske största användaren av automatisk översättning är EU:s officiella översättningsbyrå som utför hundratusentals översättningsuppdrag med hjälp av översättningssystemet SYSTRAN. En version av SYSTRAN för översättning av EU-texter från svenska till engelska har nyligen utvecklats.

Ett automatiskt system är uppbyggt av en översättningsmotor och språkliga data som beskriver källspråket, målspråket och, i synnerhet, hur ord och konstruktioner i de två språken motsvarar varandra. Det hävdas från kommersiellt håll, att man genom maskinöversättning med efterföljande redigering kan minska översättningskostnaden med mellan 50 och 70 % jämfört med mänsklig översättning. Generellt sett kan man säga att översättningskvalitet i ett automatiskt system beror av kvaliteten på data i systemet och deras relevans för den text som ska översättas. Precis samma sak gäller om den effektivitetsvinst som ett översättningsminne kan ge. Därför har teknikutvecklingen på översättningsområdet allt mer kommit att inrikta sig på problemet att på automatisk eller semiautomatisk väg utvinna data ur parallellställda original och översättningar. På detta område ligger Sverige långt framme vilket verksamt bidragit till att ett stort lexikon på kort tid kunnat tas fram inom jordbruksområdet för det ovan nämnda svensk-engelska SYSTRAN-systemet.

En skillnad mellan automatiska system och dagens översättningsminnen är att de senare betraktar en mening endast som en följd av tecken, medan de förra i allmänhet har tillgång till grammatiska och semantiska klassificeringar. Men översättningsminnena kommer att utvecklas så att de får en ökad språkkunskap. De kommer att kunna känna igen fraser och konstruktionsmönster för satser och meningar. Därmed kommer en betydligt större del av en ny text att kunna matchas mot enheter i översättningsminnet och översättningen att kunna utföras snabbare. Ett nystartat franskt företag i branschen lovar fyra till fem gångers effektivisering i utnyttjandet av frasanalyserade översättningsminnen jämfört med vanliga. En sådan stor effektivitetsvinst förutsätter verktyg som på automatisk väg kan känna igen fraser och deras motsvarigheter i godtyckliga översättningsminnen.

2.2 Översättnings- och terminologiprocesser i exportföretag

Sverige är ett land som för sin utveckling är beroende av handel och starka exportföretag. Många exportföretag, i synnerhet programvaruföretag, lägger ner stora belopp för att översätta sin dokumentation och lokalisera sina system till många främmande marknader. En kostnad på 5 MSEK per språk och år är inte ovanlig, vilket kan innebära 100-150 MSEK per år. Svenska företag skiljer sig åt i hur mogna de är i fråga om sitt språkarbete, men i många fall beror en betydande del av de höga översättningskostnaderna på att man inte tagit kontroll över sin egen dokumentation, t.ex. i form av stilguider och översättningsminnen och inte systematiskt utvecklar och underhåller sin terminologi. Om man gjorde detta kunde befintlig översättningsteknologi användas betydligt effektivare och med högre kvalitet samtidigt som en enhetlig terminologi underlättar kommunikation externt och internt.

2.3 Flerspråkig informationssökning

Allt mer information görs i dag tillgänglig på allt fler språk bl.a. över Internet. Med maskinläsbara lexikon eller parallellkorpusar ur vilka ord- och fraskorrespondenser kan utvinnas utvidgas den mängd av dokument som man kan söka i. Man kan alltså använda ett språk för sökfrågan, men få tillbaka hela dokument eller textstycken på flera språk. Med automatisk översättning kan man också få resultatet översatt. Med en tillräcklig passiv språkkunskap på målspråket kan det vara en tillräcklig hjälp att kunna formulera sökfrågan på modersmålet. Detta gäller både svenskar som kan få hjälp att hitta information i engelska, franska eller spanska texter och invandrare som behöver hitta information i svenska texter med användning av sitt modersmål.

Flerspråkig informationssökning innebär också ökad tillgänglighet för medieproduktion på svenska. Det material som svenska mediehus förfogar över, inte bara text utan även bild och film som indexeras med svenska nyckelord, är i dag inte synligt för sökningar som görs på andra språk än svenska, vilket är en nackdel i en global konkurrenssituation.

3. Integrerade system

För de applikationer som exemplifierats ovan kan språkteknologin sägas vara den centrala eller bärande tekniken. I utvecklingen av nya IT-tjänster är dock språkteknologin i många fall bara en av flera tekniker som ska samverka för ökad funktionalitet och användbarhet.

Affärskommunikation sker i ökad utsträckning elektroniskt. Inte minst viktigt är här kommunikationen mellan företag och kunder. Språkteknologi spelar en ökande roll i utvecklingen av kundtjänster av olika slag. Det kan exempelvis gälla hantering av inkommande epostmeddelanden. Om man bygger upp en databas av vanliga meddelanden skulle vissa e-brev kunna hanteras helt automatiskt medan andra kan sorteras för att hamna hos rätt mottagare. Detsamma gäller inkommande samtal till en telefontjänst där vidarekopplingen av ett samtal kan ske beroende på kundens första yttrande. Samma förhållande gäller ju för övrigt kommunikation mellan myndigheter och medborgare.

Multimodala system kan vara system där tal och text samverkar med andra ”naturliga” mänskliga kommunikationssätt som gester och blickar. Det kan också vara system där tal och text förenas med de mer etablerade teknikerna för människa-dator-interaktion: musklickning, tangentbord, pekskärmar, animeringar, dialogrutor och så vidare. Sådana system har sin tillämpning inte bara för dialog i samband med informationssökning, utan också exempelvis i dataspel eller interaktiva system för lärande. Av betydelse är också tvärmodala eller intermodala system, d.v.s. system vars interaktionssätt kan anpassas till användarens förutsättningar och behov. En snabbare eller långsammare uppläsning av en text kräver exempelvis en förändring av ordens uttal för att låta naturlig. En kombination av visuell och akustisk information kan vara en kraftfull kombination t.ex. i utbildningssammanhang.

Språkutbildning. Språkteknologin har hittills inte utnyttjats så mycket för språkutbildning. Men språkteknologin erbjuder många möjligheter att utveckla interaktiva övningar av olika slag som till exempel uttals- och dialogträning och övningar på texter som den studerande i stor utsträckning skulle kunna välja själv, eller som hon skriver själv, på ett helt annat sätt än den traditionella övningsboken. Även översättningsteknologin kan användas i utbildning i främmande språk.

Hjälpmedel. Språkliga funktionshinder är ett problem för många. Det gäller inte bara diagnostiserade funktionshindrade som dyslektiker utan även barn som inte lärt sig läsa och skriva eller äldre som förlorar mycket av sin syn eller hörsel. Den som har svårt att läsa en textremsa eller blir trött av det borde kunna få den uppläst. Den som har svårt att skriva kan ha god hjälp av ordpredicieringssystem, och så vidare.

Vad behöver svensk språkteknologi?

Språkteknologi som FoU-område har drygt femtio år på nacken. Under denna tid har stora genombrott ägt rum och resulterat i de system vi använder idag. I och med att allt mer av vår språkliga kommunikation sker med eller via datorer blir språkteknologin allt viktigare.Det är också, som utredningen SOU 2002:27 Mål i mun visar, ett område med betydande kulturella och språkpolitiska implikationer eftersom de tjänster som kan erbjudas står i proportion till de resurser i form av kunskap, utbildade människor, databaser och programvara som en språkgemenskap förfogar över. Av dessa skäl är det oerhört väsentligt att Sverige har en långsiktig strategi för språkteknologi med stöd i Sveriges riksdag.

I en sådan strategi är det viktigt att ha en balans mellan olika insatser för forskning, utveckling, utbildning och uppbyggnad av infrastrukturella resurser i form av databaser och programvara. De insatser som i dagsläget är mest angelägna är följande:

1) Fortsatta satsningar på avancerad forskning och forskarutbildning. De naturliga statliga huvudmännen för forskningen är Vetenskapsrådet för grundforskningen och VINNOVA för den behovsmotiverade forskningen. Språkteknologi bör ses som en väsentlig kunskapsplattform för utvecklingen av nya IT-tjänster.
2) Accelererad uppbyggnad och utveckling av de bastillgångar i form av databaser och programvara som krävs för forskningens och samhällets behov.
3) Att bygga organiserad samverkan mellan näringsliv, forskare och statsmakterna inom området.

Uppgiften att bygga upp bastillgångar och organisera samverkan skulle kunna hanteras av ett språkteknologiskt sekretariat. Några grundläggande krav för ett sådant sekretariat är hög beställarkompetens på språkteknologins område, vilket innebär att både forskarsamhället, industrin och myndigheter bör vara representerade. Det bör också ha en nära anknytning till forskningsråden. Kostnadsmässigt borde det ha resurser motsvarande en rådssekreterartjänst, en halv informatörstjänst, en halv assistenttjänst samt arvoderingar för sammanträden och utlåtanden. Därtill borde det ha erforderliga medel för att finansiera utvecklingen av språkteknologisk infrastruktur enligt punkt 2 ovan.

En första uppgift för sekretariatet, eller annan lämplig projektgrupp, är att genomföra en inventering av vilka resurser som i dag finns för svenska, på vilket sätt, och till vilken eventuell kostnad de är tillgängliga och på den grunden göra prioriteringar av vilka som är de mest angelägna att genomföra. En sådan inventering har gjorts för nederländska (Binnenpoorte et al., 2002) med syfte att definiera vad man kallar ett Basic Language Resource Kit (BLARK) för nederländska inklusive flamländska. Den metodologi som detta projekt utvecklade kan tillämpas även på svenska förhållanden. Den skulle även kunna tillämpas på våra minoritetsspråk och viktigaste invandrarspråk.

Utan att vilja föregripa en sådan noggrannare inventering, vill vi ändå peka på några viktiga resurser som i dag saknas för svenska språket helt eller delvis, och som vi ovan visat leder till förbättrade system:

1) Ett grundläggande lexikon för svenska med morfologi, som kan användas fritt eller till självkostnadspris i storleksordningen 50,000 grundformer.
2) Ett vältäckande semantiskt ordnät för svenskan. Detta bör ske i samarbete med europeiska initiativ på området.
3) Databaser med talspråk i olika situationer, både monologiskt tal och vardagliga samtal. Databaserna bör innehålla ett representativt urval dialekter och brytningar bland annat som grund för robust taligenkänning och språkinlärning.
4) Uppmärkta textdatabaser, även kallade trädbanker, dvs. olika typer av texter som samlats in systematiskt och märkts upp med information om grammatiska egenskaper och av betydelse. Här kan man i ett första skede utgå från texter som redan samlats in.
5) Parallelkorpusar, dvs textdatabaser bestående av originaltexter och översättningar, vilka är avgörande hjälpmedel för at utveckla maskinöversättningsprogram (se ovan). De bör byggas upp för svenska, stora främmande språk och även minoritetsspråk och stora invandrarspråk i Sverige.

Mycket av detta arbete kan givetvis göras inom ramen för nordiskt samarbete och EU-samarbete. Det är därför viktigt att det språkteknologiska sekretariatet ges uppdraget att utveckla samarbetet på det språkteknologiska området i Norden och EU.

Vad kostar forskning och resursutveckling?

Under 2004 satsar VINNOVA ca 13 MSEK på tal- och språkteknologi fördelat på CTT, det språkteknologiska programmet och några mindre projekt. Det språkteknologiska programmet tar slut med utgången av 2004 och CTT avslutas 2006. Vetenskapsrådet finansierar ett begränsat antal projekt med mindre belopp (0,5 – 1 MSEK/år) som söks i öppen konkurrens. Detta är låga belopp i en internationell jämförelse och målet bör vara en fördubbling. Därigenom skulle större projekt med tydlig fokusering möjliggöras, så att utvecklingen på åtminstone några av de områden som nämnts ovan kan ta rejäla steg framåt.

Genom att vara en språkteknologiskt avancerad nation kan Sverige lättare uppnå mål som är väsentliga för vårt samhälle. Det är svårt att föreställa sig att den strukturering av information som målet om tjugofyratimmarsmyndigheten förutsätter skall ske på manuell bas. Effektiviseringsvinster som kan göras genom talinmatning inom olika samhällssektorer, t.ex. patientjournaler inom sjukvården, får inte gå om intet på grund av att det saknas svenska taldatabaser att tillgå. Det är viktigt att texter på svenska språket kan översättas och nås med samma grad av automatik som texter på andra språk.

Kostnader för framtagning av resursdatabaser och verktyg är svåra att uppskatta, bland annat beroende på att det saknas en inventering av svenska resurser. Vi utgår här ifrån de uppskattningar som gjorts i utredningen om en norsk språkbank (Svendsen m.fl., 2002). Där har man försökt göra uppskattningar med hjälp av olika nyckeltal för inspelningskostnader, uppmärkningskostnader, lönekostnader, färdiggjorda ord per timme och så vidare. De siffror som anges i tabellen nedan är modellerade efter den utredningens sätt att räkna, men anpassade till angivna storlekstal. Norska och svenska kronor har antagits motsvara varandra ett till ett. Ingen hänsyn har tagits till skillnader mellan Norge och Sverige i fråga om vad som finns att tillgå i dagsläget. Siffrorna bör alltså tas för vad de är: uppskattningar i avsaknad av säkra data. Men de kan ändå ge en fingervisning om behoven.

När det gäller uppmärkta textdatabaser finns en projektansökan från 2003 till Riksbankens Jubileumsfond om ett femårigt projekt. Ansökan innebar samarbete mellan en majoritet av språkteknologiska forskargrupper i Sverige. Totalkostnaden uppskattades till ca 25 MSEK. I den kostnaden ingick då också lön för tre doktorander och analys av transkriberade taldata. Med ett projekt som enbart inriktar sig på utvecklingen av själva dataresursen och verktygen för dess fortsatta utbyggnad kan en totalsumma om 15 MSEK vara mer rimlig.

Tabell 1: Uppskattning av kostnader för några viktiga språkteknologiska resurser.

Typ av resurs	**Storlek**	**Kostnad**
Grundläggande svenskt lexikon med morfologi	50,000 lexem	1 MSEK
Ordnät för svenska	50,000 lexem	3-3,5 MSEK
Taldatabas	1000 timmar	25 MSEK
Analyserade textdata (”trädbank”)	10,000,000 ord varav 1,000,000 expertgranskade	15 MSEK
Parallellkorpus	5,000,000 ord (2 riktningar)	8-10 MSEK

Referenser

D. Binnenpoorte, F. De Vriend, J. Sturm, W. Daelemans, H. Strik, C. Cucchiarini. A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch. Third International Conference on Language Resources and Evaluation (LREC’2002), Las Palmas, Canary Islands, 29-31 May, 2002. http://lands.let.kun.nl/literature/binnenpoorte.2002.3.pdf

Torbjorn Svendsen, Torbjorn Nordgård m.fl. Samling og tilgjengeleggjering av norske språkteknologiressursar. Endeleg rapport, oktober 2002, från Projektgruppe oppnemnd av Kultur- og kyrkjedepartementet, 19.3 2002.
http://www.sprakrad.no/sbank2.htm

Rickard Domeij, Språkrådet 2006-12-14

Språkpolitik och språkteknologi i Sverige och Norden

Nyckeln till delaktighet i samhället är språket. Det öppnar dörrarna till social och kulturell gemenskap. Det ger oss tillgång till nödvändig samhällsinformation och möjlighet att påverka vår situation. Det ökar möjligheterna till framgång i arbetslivet. Den som inte behärskar det eller de språk som samhället baseras på ställs obönhörligen utanför.

Samma gäller den som inte har tillgång till den teknik som i allt större utsträckning förmedlar den språkligt burna kulturen. Dagens flerspråkiga informationssamhälle kräver inte bara språkliga kunskaper, utan också nätuppkoppling och grundläggande datorfärdigheter. Den som har det finner nya sätt att söka information och delta i kommunikativa gemenskaper oavsett nationsgränser.

Ett forskningsområde som på ett väsentligt sätt kan bidra till att förbättra den språkliga kommunikationen och tillgängligheten till information är språkteknologi. Därför är språkteknologi något som uppmärksammas inom svensk språkpolitik. Sverige har sedan ett år tillbaka en av riksdagen antagen språkpolitik som fastställer medborgarnas språkliga rättigheter. De fyra övergripande målen för svensk språkpolitik är att:

- svenska språket ska vara huvudspråk i Sverige
- svenskan ska vara ett komplett och samhällsbärande språk
- den offentliga svenskan ska vara vårdad, enkel och begriplig
- alla ska ha rätt till språk: att utveckla och tillägna sig svenska språket, att utveckla och bruka det egna modersmålet och nationella minoritetsspråket och att få möjlighet att lära sig främmande språk.

Väl fungerande språkteknologi på svenska är en förutsättning för att Sverige ska uppnå målen. Det gör språkteknologi till en språkpolitisk angelägenhet i Sverige, liksom i våra nordiska grannländer och inom EU. Vad är språkteknologi och varför är den språkpolitiskt betydelsefull? Vad görs och behöver göras för att stärka språkteknologin i Sverige och de nordiska länderna? Det är vad det här dokumentet handlar om.

Vad är språkteknologi?

Inom forskningsområdet språkteknologi utvecklar man metoder för att analysera och bearbeta mänskligt språk både i skriven och i talad form. Syftet är att förstå vad språklig kommunikation är och skapa språkteknologiska hjälpmedel som gagnar den. Några stora kommersiella tillämpningsområden är:

- Översättning: terminologiska databaser, översättningsminnen och maskinöversättning.
- Informations- och kunskapshantering: indexering, informationssökning, informationsextraktion och textsammanfattning.
- Talteknologi: konstgjort tal (talsyntes), taligenkänning, dialogsystem och ”talande huvuden”.
- Textframställning: stavnings- och grammatikkontroller, diktering, avstavningsfunktioner och elektroniska ordböcker.

Talteknologi är det tillämpningsområde som utvecklats och expanderat mest de tio senaste åren. Idag kan man t.ex. få en text uppläst i webbläsaren på konstgjord väg av ett konstgjort talande huvud med näst intill naturlig röst och mänskliga munrörelser. Man kan också själv tala med ett datorsystem, t.ex. för att efterfråga och få information över telefon. Tekniken är inte helt problemfri men fungerar bra för många tillämpningar.

Informationshanteringstekniken har också fått ett stort genombrott, inte minst med den ökade användningen av webben. Den pågående utvecklingen av den semantiska webben ställer tekniken inför nya utmaningar med att hantera informationsinnehåll. Samtidigt ökar behovet av flerspråkig teknik som kan överbrygga gränserna mellan olika språk så att man t.ex. kan söka information på flera språk samtidigt. Och helst också få de eftersökta dokumenten direkt översatta med maskinöversättning – ett område som är på stark frammarsch just nu. Resultatet blir långt ifrån lika perfekt som med mänsklig översättning, men tillräckligt bra för många situationer där mänsklig översättning inte är ett alternativ, t.ex. när man direkt behöver en grovöversättning för att få en uppfattning om vad som skrivs eller sägs på ett främmande språk. Tekniken öppnar oanade möjligheter till kommunikation över språkgränserna (se t.ex. *Human language technologies for Europe*, 2006).

Språkteknologisk forskning och utveckling är resurskrävande. Empiriskt material i form av omfattande representativa text- och taldatabaser, så kallade korpusar, är oundgängliga för att utveckla och testa ny teknik, som ofta involverar datakrävande statistiska modeller. Likaså behövs grundläggande verktyg för att analysera och märka upp korpusarna – hel- eller halvautomatiskt – med information om t.ex. ordklass, ordböjning, frastillhörighet, grammatisk funktion, betydelse och uttal. Ett maskinöversättningssystem behöver t.ex. stora mängder uppmärkt text med länkade översättningar på olika språk att träna på, s.k. parallellkorpusar.

Ord- och textdatabaser används dessutom inom språkforskningen och lexikografin, liksom inom andra forskningsdiscipliner som har behov av databaser med språkligt material och avancerade metoder för att hantera dem. Därigenom kan språkteknologin på ett väsentligt sätt bidra till utvecklingen av framtidens human- och samhällsvetenskapliga forskning och bevarandet av vårt kulturarv.

Språkteknologins språkpolitiska betydelse

Den svenska utredningen *Mål i mun* (2002) konstaterade att Sverige behöver en samlad språkpolitik för att hantera språksituationen i dagens och framtidens samhälle. Det ledde fram till propositionen *Bästa språket* (2005) som formulerade målen för svensk språkpolitik, och bidrog till att Språkrådet bildades som en del av myndigheten Institutet för språk och folkminnen med ansvar att genomdriva politiken. Språkteknologisk forskning och utveckling är en viktig del i arbetet med att uppnå de språkpolitiska målen. Därför står det i instruktionen för språkmyndigheten att den särskilt ska främja språkteknologiskt arbete.

Huvudspråket i Sverige är svenska. Det ska vara ett komplett och samhällsbärande språk. Det säger de två första språkpolitiska målen. Det innebär att svenskan måste kunna erbjuda sina användare ett rikt utbud av språkteknologiska tillämpningar. Annars förlorar det mark gentemot språk som är bättre teknologiskt rustade, som t.ex. engelskan. Om det t.ex. inte finns talteknologi för svenska, leder det till att svenskar tvingas tala engelska när de använder sig av sådan teknik. För att svenskar ska kunna använda svenska i alla sammanhang måste vi se till att det finns språkteknologi för informationssökning, textframställning, översättning m.m. vare sig det gäller skriven eller talad svenska. Med elektroniska ordböcker, termbanker och språkkontroll kan svenskans ordförråd säkras och språkriktigheten stärkas, vilket i viss mån också bidrar till det tredje målet: att den offentliga svenskan ska vara vårdad, enkel och begriplig.

Åtgärder som stärker språk och språklig kommunikation, stärker också människors delaktighet i det samhälle de lever i. Det sista språkpolitiska målet syftar just på detta: att alla ska ha

rätt till språk för att inte hamna utanför språkliga gemenskaper. Medborgarna ska inte bara ha rätt till svenska, utan också till modersmål, minoritetsspråk och främmande språk. Därför bör det också finnas språkteknologi för de svenska minoritetsspråken och övriga språk i Sverige, så att alla åtminstone kan få tillgång till viktig samhällsinformation på det egna språket.

Många grupper i samhället med behov av särskilt stöd har stor nytta av språkteknologiska hjälpmedel. Människor med kommunikativa funktionshinder kan t.ex. få text uppläst med hjälp av konstgjort tal, eller omvänt få talet omvandlat till text. För personer med läs- och skrivsvårigheter finns andra användbara hjälpmedel. Hjälpmedlen är en viktig del i arbetet med att göra information tillgänglig för alla – en central tanke i utvecklingen av myndigheternas nätverksamhet, den s.k. 24-timmarsmyndigheten.

Med utvecklingen av maskinöversättning och annan flerspråkig teknik ökar alla medborgares möjligheter att kommunicera på det egna språket i en flerspråkig värld. Inte minst är det en viktig fråga för EU med för närvarande 20 officiella språk som ständigt kräver översättning. Utvecklingen av den europeiska gemenskapen förutsätter en god kommunikation över språkgränserna. Ministerrådets rapport *En ny ramstrategi för flerspråkighet* (2005) pekar på att språkteknologin har en nyckelroll i en sådan utveckling och understryker därför behovet av att stärka ”forskning om och teknisk utveckling av språkrelaterad teknik i informationssamhället, med särskilt fokus på ny maskinöversättningsteknik”. Det förutsätter i sin tur en väl utbyggd språkteknologisk infrastruktur: ”Ett flerspråkigt informationssamhälle behöver tillgång till standardiserade och driftskompatibla språkresurser (ordböcker, terminologi, textkorpusar osv.) och programvara för alla språk, också för EU:s mindre utbredda språk.”

Språkteknologin i Norden

Liksom inom EU uppmärksammas språkteknologins betydelse inom nordisk språkpolitik. Nordiska rådet antog nyligen en deklaration om en gemensam nordisk språkpolitik som ska se till att Norden är en föregångsregion för internationellt språkpolitiskt arbete (*Deklaration om nordisk språkpolitik*, 2006). Deklarationen tar sin utgångspunkt i att alla nordbor har rätt att

- tillägna sig ett samhällsbärande språk i tal och skrift, så att de kan delta i samhällslivet
- tillägna sig förståelse av och kunskaper i ett skandinaviskt språk och förståelse av de övriga skandinaviska språken, så att de kan ta del i den nordiska språkgemenskapen
- tillägna sig språk med internationell räckvidd, så att de kan delta i utvecklingen av det internationella samfundet
- bevara och utveckla sitt modersmål och sitt nationella modersmål.

För att öka språkförståelsen och språkkunskaperna i Norden vill man bl.a. att ”maskinöversättning för Nordens samhällsbärande språk och program för flerspråkig sökning i nordiska databaser utvecklas” samt att ”internordiska ordböcker i pappersform och i elektronisk form utarbetas”.

Den nordiska språkdeklarationen är ett uttryck för en större medvetenhet i de nordiska länderna om behovet av språkpolitik i dagens mångkulturella och flerspråkiga samhälle. Under senare år har de nordiska länderna ett efter ett börjat ta fram nationella, språkpolitiska och forskningspolitiska handlingsplaner där språkteknologins roll uppmärksammas (se t.ex. *Handlingsplan for norsk språk og IKT*, 2001; *Sprog på spil – et udspil til en dansk sprogpolitik*, 2003; Maegaard m.fl, 2004).

De nordiska språknämnderna samarbetar om språkteknologiska frågor i Arbetsgruppen för språkteknologi och språkvård i Norden med stöd av Nordens språkråd, som är en del av det Nordiska ministerrådet. Syftet är att stärka det språkpolitiska samarbetet om språkteknologiska frågor i Norden och främja nordisk språkteknologi. Arbetsgruppen anordnar bland

annat seminarier för att diskutera nordisk språkteknologi med forskare, industrirepresentanter och andra viktiga aktörer på området.

Arbetet har bland annat resulterat i att Nordiska ministerrådet låtit ta fram en s.k. vismansrapport (*Språkvis*, 2006) med en tioårsplan för att utveckla språkteknologin i Norden med visionen att göra Norden till en ledande region på området. I rapporten framhålls behovet av och fördelarna med att ta fram gemensamma språkteknologiska resurser för de nordiska länderna. Där föreslås bl.a. att ett samordnande nordiskt organ etableras som ser till att inventera befintliga resurser och resursbehov på området. Utifrån inventeringen bör en samnordisk plan upprättas för finansiering och framtagande av språkteknologiska resurser för de nordiska länderna.

Förutsättningarna för ett nordiskt samarbete måste anses vara goda. Det råder som vi sett en bred samsyn såväl inom Norden som inom EU om betydelsen av språkteknologisk forskning och utveckling. Man är också överens om att stora satsningar behöver göras för att bygga ut den språkteknologiska infrastrukturen, såväl nationellt som internationellt. Det största problemet är att en sådan satsning är förenad med omfattande kostnader som de enskilda länderna har svårt att finansiera fullt ut.

Därför vore ett samarbete mellan de nordiska länderna med stöd från EU den bästa lösningen, särskilt med tanke på ländernas politiska samsyn, språkliga och kulturella gemenskap och långa tradition av nära kontakter och samarbete på många områden. Dessutom har flera av de nordiska huvudspråken stora likheter. Vissa språk har också status som huvudspråk eller minoritetsspråk i flera länder, t.ex. finskan i Finland och Sverige (minoritetsspråk), svenskan i Sverige och Finland, och samiskan i Norge, Sverige och Finland. Det gör att inte bara teknikresurser (t.ex. grundläggande språkanalysverktyg), utan också vissa språkresurser (t.ex. korpusar) kan delas mellan de nordiska länderna. Det finns alltså mycket att vinna på ett samarbete, såväl ekonomiskt som kulturellt.

Organisatoriskt sett finns redan befintliga strukturer att bygga vidare på. Sedan ett halvt sekel tillbaka anordnas vartannat år den nordiska språkteknologikonferensen Nodalida. Mellan 2000-2004 pågick ett nordiskt samfinansierat forskningsprogram för språkteknologi som bland annat resulterade i en nordisk forskarskola, NGSLT, och uppbyggandet av språkteknologiska dokumentationscentrum för de nordiska länderna på webben, med Språkteknologi.se som svensk representant. Webbplatserna bildar ett nätverk för kontakt och informationsspridning om språkteknologi inom och mellan länderna. På terminologiområdet finns ett liknande nätverk, Nordtermnet, som samarbetar inom nordisk terminologi bl.a. i arbetet med en nordisk termbank. Nyligen har dessutom språkteknologiorganisationen NEALT bildats, med representanter från de nordiska länderna, samt de baltiska länderna och delar av Ryssland. Målet är att ytterligare stärka forskningssamarbetet mellan länderna och bredda det.

Med den språkpolitiska utvecklingen i de nordiska länderna och bildandet av Nordens språkråd och Arbetsgruppen för språkteknologi i Norden finns nya möjligheter att samordna och påverka språkteknologiutvecklingen i Norden. På senare år har Nordens språkråd finansierat några samnordiska språkteknologiska projekt. Bl.a. för att ta fram en nordisk nätordbok innehållande ordböcker för de nordiska språken och en flerspråkig sökfunktion som gör det möjligt att söka på ett svenskt ord och samtidigt få träffar på motsvarande ord i de andra språken. I oktober 2006 arrangerades ett nordiskt seminarium i Göteborg i där vismansrapportens förslag och möjligheterna till samarbete om en språkteknologisk infrastruktur i Norden diskuterades.

Språkteknologiskt arbete i Sverige

I Sverige har man framför allt under 1990-talet satsat en hel del offentliga medel till språkteknologisk forskning och utveckling, främst från Verket för näringslivsutveckling

(Nutek) och dåvarande Humanistisk-samhällsvetenskapliga forskningsrådet i det s.k. Språkteknologiprogrammet. Satsningarna har bidragit till att svensk språkteknologi är relativt välutvecklad och har god organisation, vilket den nationella forskarskolan i språkteknologi, GSLT, är ett exempel på. Språkrådet och GSLT samarbetar sedan några år om att driva webbplatsen Språkteknologi.se, en portal för svensk språkteknologi med information om aktiviteter, resurser, produkter och aktörer på området. Dock saknas fortfarande mycket av den infrastruktur i form av språkteknologiska grundresurser som skulle behövas för att påtagligt driva utvecklingen framåt.

I övriga nordiska länder är utvecklingen någorlunda jämförbar, men man kan notera att Norge har satsat stort på språkteknologi under 2000-talet med forskningsprogrammet KUNSTI, medan det inte funnits något motsvarande program i Sverige. Norge är också det land som kommit längst i planerna på att samla, ta fram och tillgängliggöra nationella språkteknologiska resurser i ”en norsk språkbank”. Språkrådet i Norge har på uppdrag från Kultur- och kirkedepartementet låtit utreda vad ett sådant arbete skulle medföra och kosta (*Samling og tilgjengeleggjering av norske språkteknologiressursar*, 2002). Det politiskt fastslagna målet är att på sikt bygga upp en norsk språkbank med språkteknologiska resurser till nytta för norsk forskning och industri. Arbetet med att lösgöra och samla in befintliga resurser har påbörjats.

I Sverige finns en politiskt uttalad vilja att göra motsvarande. I propositionen *Bästa språket*, som banade vägen för den svenska språkpolitiken, uttrycks den så här:

> ”Centralt för att främja en god utveckling på språkteknologiområdet är att systematiskt bygga upp stora text- och taldatabaser och att utveckla programvaror. I text- och taldatabaser lagras mycket stora mängder autentiskt tal- och skriftspråk på ett sätt som gör det åtkomligt för datoriserad, språkvetenskaplig analys. En sådan analys är i sin tur en förutsättning för att utveckla program för automatisk översättning, för överföring av text till tal (och vice versa), för datoriserad taligenkänning m.m. Uppbyggnaden av text- och taldatabaser är kostsamt och arbetskrävande samt fordrar långsiktig planering och handlar om att skapa språkteknologiska basresurser för att utveckla välfungerande språkteknik. Det är således inte möjligt för den nya språkvårdsorganisationen att själv genomföra detta arbete, men den bör ha kompetens att inventera och överblicka behoven samt ta initiativ till nödvändiga samarbetsprojekt. […] Vi anser därför att en funktion för samordning av språkteknologi bör finnas hos den nya språkvårdsorganisationen så att resurser bättre kan samordnas och förutsättningarna för att medverka inom större samverkansprogram inom Norden och EU förbättras. Språkvårdsorganisationen bör exempelvis långsiktigt verka för att uppmärkta och representativa text- och taldatabaser utvecklas. En första uppgift i det arbetet kan vara att inventera dagens resurser för svenska språket, på vilket sätt och till vilken eventuell kostnad de är tillgängliga och därefter göra angelägna prioriteringar. En sådan inventering bör även göras för våra nationella minoritetsspråk och vanligaste invandrarspråk.”

Nyligen har Vetenskapsrådet beviljat ett tvåårigt planeringsprojekt med syfte att inventera behovet av svenska språkteknologiresurser och ta fram en plan för framtida utveckling av nödvändiga resurser. Projektet, som startar 2007, är ett samarbete mellan ledande språkteknologer knutna till den svenska forskarskolan för språkteknologi (GSLT), Språkbanken i Göteborg och Språkrådet. Projektet gör att Sverige kan följa Norge i spåren och utarbeta en plan för att ta fram språkteknologiska resurser för språken i Sverige. I arbetet ingår att

- undersöka behovet av resurser för svensk språkteknologisk forskning och utveckling, samt för språkvetenskaplig och näraliggande humanvetenskaplig forskning

- inventera redan befintliga resurser, deras status och tillgänglighet
- planera för att lösgöra befintliga resurser och för att utveckla nya resurser utifrån framtagna kostnadsberäkningar och prioriterade behov.

Återstår sedan att sätta planerna i verket. För att åstadkomma detta måste flera centrala frågor lösas, bl.a. följande:

Samordning och finansiering. Hur arbetet ska samordnas och finansieras måste klargöras. Aktuella parter för ett samarbete är GSLT, Språkbanken, Språkrådet, Vetenskapsrådet (KFI, DISC och SND), Vinnova och intressenter från näringslivet. Det är också viktigt att företrädare för minoritetsspråken och för människor med särskilda behov är inblandade. Även om samfinansiering från företag är eftersträvansvärt, måste troligen den huvudsakliga finansieringen komma från samhälleligt håll. Inget hindrar dock att utvecklingen av forskningens infrastruktur kombineras med strategiska satsningar på tillämpningar, t.ex. maskinöversättning, med stöd från både Vetenskapsrådet och Vinnova. Möjligheterna till samarbete i Norden och stöd från EU måste undersökas.

Juridiska frågor. Upphovsrättslagen ställer till stora problem vid insamling och spridning av språkresurser, t.ex. korpusmaterial. Detta gäller även om materialet bara används som tränings- och utvärderingsmaterial i konstruktionen av språkteknologiska system och inte görs tillgängligt i klartext. Det bör undersökas hur man kan tackla de juridiska problem som uppstår i olika situationer. Det behövs juridisk rådgivning och mallavtal som underlättar vid insamling och spridning av resurser.

Öppna resultat. De resurser som finansieras med samhälleliga medel bör komma hela samhället till del, såväl forskarsamhället som i möjligaste mån även företagen. I Sverige krockar den principen med det så kallade lärarundantaget som ger forskare rätt till de egna resultaten. Det bör därför finnas juridiskt bindande avtal som klargör äganderätten till resurserna och säkrar spridningen av dem. Med avtal om öppen källkod blir det lättare att såväl sprida resurserna, som att tillåta att de modifieras och vidareutvecklas av andra.

Standarder och kvalitetssäkring. Tydliga riktlinjer bör tas fram för hur resurserna ska dokumenteras, utvärderas och kvalitetssäkras. Språkresurserna ska vara uppmärkta enligt föreskrivna format. Teknikresurserna bör göras modulära med standardiserade gränssnitt så att de är lätta att använda och lätt kan kopplas samman med varandra och med andra befintliga resurser. Riktlinjerna ska baseras på internationellt framtagna standarder och bästa praxis.

Lagring och spridning. Färdiga resurser bör finnas lätt tillgängliga på webben i ett gemensamt gränssnitt, vilket inte hindrar att lagringen distribueras över flera datorer. Språkresurser för humanvetenskaperna bör vara sökbara on-line. Andra frågor som bör diskuteras och lösas är de som rör underhåll, driftssäkerhet, åtkomst, informationsspridning, användarinstruktioner m.m. Lösningar bör diskuteras med tanke på de möjligheter som erbjuds av bl.a. DISC, SND, Språkbanken, Språkteknologi.se och Humanistlaboratorierna i Lund och Umeå.

De frågor som arbetet med en språkteknologisk infrastruktur väcker är visserligen komplexa, men fullt hanterbara. Det finns färdiga resultat, metoder och erfarenheter att falla tillbaka på. Nya möjligheter står för dörren. Det planerade EU-projektet CLARIN kan bli en vägvisare med sin målsättning att bygga en europeisk infrastruktur för tillgängliggörande av språkteknologiska resurser för human- och socialvetenskaperna via webben (CLARIN, 2006).

Språken i Sverige och det svenska samhället har mycket att vinna på att vi ser till att Sverige ligger långt framme i den språkteknologiska utvecklingen och har en väl utbyggd språkteknologisk infrastruktur, gärna i samarbete med övriga nordiska länder. Ska Norden vara en föregångsregion för språkpolitiskt arbete måste det vara med och visa vägen in i framtiden.

Källor

Bästa språket – en samlad svensk språkpolitik. Proposition 2005/06:2. Utbildnings- och kulturdepartementet. 2005. <www.regeringen.se/sb/d/5359/a/50761>

CLARIN – Common Language Resources and Technologies Infrastructure. 2006. <www.mpi.nl/clarin/pdf/clarinmission-1.pdf>

Handlingsplan for norsk språk og IKT. Norsk språkråd. Oslo 2001. <www.sprakrad.no/iktrev.htm>

Human language technologies for Europe. Information Society and Media. 2006. www.tc-star.org/pubblicazioni/D17_HLT_ENG.pdf

Maegaard B., Bick E., Dalsgaard P., Kirchmeier-Andersen S., Togeby O., Henriksen B.H.: *Strategisk satsning på dansk sprogteknologi*. Statens Humanistiske Forskningsråd, København 2004. <www.cst.dk/dandokcenter/sprog/STRATEGISK_SATSNING.PDF>

En ny ramstrategi för flerspråkighet. KOM(2005) 596 slutlig. Meddelande från kommissionen av den 22 november 2005. <eur-lex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!DocNumber&lg=sv&type_doc=COMfinal&an_doc=2005&nu_doc=596>

Deklaration om nordisk språkpolitik. Nordiska ministerrådet, 13. september 2006. <http://www.norden.org/sagsarkiv/sk/sag_vis.asp?vis=2&id=335>

Mål i mun. Förslag till handlingsprogram för svenska språket. SOU 2002:27. 2002. <www.regeringen.se/sb/d/108/a/1443>

Samling og tilgjengeleggjering av norske språkteknologiressursar. Norsk språkråd, Oslo 2002. (Prosjektplan for norsk språkbank). <www.sprakrad.no/upload/1308/sprakbankrapport-2002.pdf>

Sprog på spil – et udspil til en dansk sprogpolitik. Kulturministeriet 2003. http://www.kum.dk/sw6576.asp

Språkvis - Vismansrapport - Expert Panel Report. The Nordic Countries - A Leading Region in Language Technology. 2006. <www.ling.helsinki.fi/~klinden/pubs/Spr%E5kVisFullReport.pdf>

2007-01-22

Vetenskapsrådet

Database Infra-Structure Committee (DISC)

Till institutioner/enheter med språkteknologisk forskning/egna språkdatabaser

Kartläggning av databasresurser inom språkteknologi – läget idag och framtida behov

KFI (Kommittén för forskningens infrastrukturer) vid Vetenskapsrådet har som uppdrag att *främja och stödja uppbyggnad och utnyttjande av infrastruktur för svensk forskning av högsta vetenskapliga kvalitet*. DISC (Database Infra-Structure Committee) har under KFI ansvar för den del av infrastrukturen som utgörs av databaser.

Som ett led i detta arbete har DISC beslutat göra en kartläggning av tillgängliga forskningsdatabaser och framtida behov av databasresurser inom språkteknologi. Arbetet ingår som en del i underlaget för en långsiktig planering av svenska forskares tillgång till forskningsinfrastruktur.

I den strategiska rapporten ”Vetenskapsrådets guide till infrastrukturen”[1] har ämnet språkteknologi utpekats som ett av de viktiga områden där utredningar snarast behöver göras för att klargöra hur infrastrukturen kan byggas upp, förbättras eller effektiviseras. Svensk språkteknologi utgör ett område med hög vetenskaplig kvalitet både nationellt och internationellt och ”svensk forskning inom språkteknologi befinner sig i en unik situation i och med att ett välfungerande samarbete har vuxit fram mellan svenska universitet och tekniska högskolor. Det finns ett stort behov av att se över den nationella infrastrukturen för språkteknologi och verka för samordning av databaser och analysverktyg.” (s.13)

Den aktuella kartläggningen, som ska ses som ett komplement till en tidigare inventering av databasresurser inom humaniora och samhällsvetenskap i Sverige[2], genomförs av Eva Strangert på uppdrag från DISC. I arbetet deltar också Merle Horne, Lund. Resultatet ska föreligga i en skriftlig rapport, där de inkomna svaren från denna rundskrivelse ska medfölja som bilaga.

Rapporten ska spegla förhållandena vad gäller *nu existerande databaser* inom språkteknologi (främst digitala men också ev ännu ej digitaliserat material) inklusive infrastruktur för databashantering (t ex verktyg för

VETENSKAPSRÅDET
SWEDISH RESEARCH COUNCIL

Postadress/Postal address
SE-103 78 Stockholm
Sweden

Besöksadress/Visiting address
Regeringsgatan 56

Tel: +46-(0)8-546 44 000
Fax: +46-(0)8-546 44 180

Org.nr.
202100-5208

vetenskapsradet@vr.se
www.vr.se

[1] http://www.vr.se/download/18.4b3ca0f810bf51c92278000164/VRsguidetillinfrastruktur060608_prel.pdf

[2] http://www.vr.se/download/18.320a86de108dcd98cbc80004975/Om+forskningens+infrastrukturer_KFI-HS_slutrapport.pdf

uppbyggnad, inmatning, uppmärkning, sökning, uppdatering samt underhåll). Rapporten ska också visa på *framtida behov av resurser* för planering, utveckling, uppbyggnad, drift och avveckling av databaser för forskningen. Eftersom behovet av resurser för databaser bör ses i relation till den forskning som sker inom språkteknologi ska kartläggningen också omfatta *en beskrivning av området som sådant*.

Mot denna bakgrund ombeds ni ge en beskrivning om *totalt 3-5 sidor* som för er verksamhet innefattar följande:

1) Den språkteknologiska forskning som sker/planeras inom den egna institutionen/enheten och som kräver tillgång till stora databaser och databasverktyg. Ange:

    - Inriktning/nuvarande projekt
    - Forskningsplaner
    - Gruppens storlek och sammansättning

2) De forskningsdatabaser (inklusive verktyg) som finns vid institutionen/enheten. Ange där så är relevant och möjligt:

    - Ändamål som databaserna skapades för
    - Nyttjandefrekvensen inom den egna institutionen/enheten samt bland andra forskare i Sverige och/eller andra länder
    - Samarbete med andra forskare i Sverige och/eller andra länder kring uppbyggnaden av databaserna
    - Samordning med en eller flera andra nationella och/eller internationella databaser
    - Samordning ifråga om anpassning av dokumentation och standarder till liknande nationella och/eller internationella databaser
    - Kostnader för databasuppbyggnad och kostnader för drift, underhåll och support (personal, hård- och mjukvara)
    - Finansiering av databasuppbyggnad samt drift
    - Ev annan utrustning som behövs som komplement i verksamheten som rör databaser

3) Det framtida behovet av resurser för planering, utveckling, uppbyggnad, drift och avveckling av databaser samt ev annan utrustning som behövs som komplement i verksamheten som rör databaser. Uppgifter som bör anges är:

    - Ändamål
    - Förväntad nyttjandefrekvens samt samarbete/samordning nationellt och internationellt av verksamheten
    - Beräknade kostnader i ett 5-årsperspektiv

Vetenskapsrådet

- Ev utsikter till finansiering (t ex ansökningar under beredning, samfinansiering inom nordiska eller europeiska nätverk etc)

Det är av största vikt för forskningen inom språkteknologi att kartläggningen beskriver området och visar omfattningen och behoven av investeringar vad gäller forskningsinfrastruktur. Kartläggningen bör vara så fullständig som möjligt. Där är därför viktigt att rundskrivelsen besvaras. Ange kontaktperson.

Era svar önskas **senast onsdagen den 15 november**. De sänds **som bifogat dokument (pdf)[3] per e-post** till undertecknad som också besvarar eventuella frågor med anledning av denna rundskrivelse.

Med vänliga hälsningar,

Eva Strangert, DISC, e-post: eva.strangert@nord.umu.se, tel: 090/7865680

SÄNDLISTA

Centrum för talteknologi (CTT), Skolan för datavetenskap och kommunikation, KTH
Datalingvistik, Högskolan i Skövde
Datorlingvistikgruppen, Institutionen för lingvistik, Stockholms universitet
MALT (Models and Algorithms for Language Technology Research), Matematiska och systemtekniska institutionen, Växjö universitet
NLPlab, Institutionen för datavetenskap, Linköpings universitet
Swedish Institute of Computer Science (SICS), Stockholm
Språkdata/Språkbanken, Institutionen för svenska språket, Göteborgs universitet
Språkteknologigruppen, Göteborgs universitet
Språkteknologigruppen, Skolan för datavetenskap och kommunikation, KTH
Språkteknologigruppen, Institutionen lingvistik och filologi, Uppsala universitet
Ledningsgruppen för Graduate School of Language Technology (GSLT)
Prefekter (motsv) inom lingvistik vid universiteten i Stockholm, Göteborg, Lund, Uppsala och Umeå

[3] För att uppnå enhetlighet i slutdokumentet önskas svaren utskrivna med Times Roman, 12 pt.

Databasresurser inom språkteknologi: CLT

1 CLT: Språkteknologiforskning i Göteborg

Språkteknologiforskningen i Göteborg kännetecknas av samarbete över traditionella ämnes-, fakultets- och universitetsgränser. I samarbetet ingår fyra institutioner, Institutionen för lingvistik, Institutionen för svenska språket, Filosofiska institutionen (Göteborgs universitet, Humanistiska fakulteten) och Institutionen för data- och informationsteknik (Chalmers tekniska högskola/ Göteborgs universitet, IT-fakulteten).

För att markera detta samarbetes stabila karaktär, bedriver vi det under ett gemensamt namn, *Centre for Language Technology* (CLT), ett slags virtuellt centrum som för närvarande fungerar som en samlingsrubrik för samarbetet, men vi bedriver också ett aktivt arbete för att skapa en mer formaliserad organisation för CLT. För det ändamålet lämnade vi in en Linnéansökan som trots lysande omdömen av de internationella bedömarna dock till slut inte fick något Linnéstöd. Vi räknar med att lämna in en ny Linnéansökan i nästa ansökningsomgång.

Den språkteknologiforskning som bedrivs i Göteborg går tillbaka till det korpusarbete som Sture Allén påbörjade på 1960-talet, vilket så småningom ledde till Språkbankens tillkomst. 1984 startades ett fyraårigt grundutbildningsprogram i datalingvistik, det första i sitt slag i Sverige och nu ett av två i landet, och faktiskt relativt unikt i Europa än idag. Humanistiska fakulteten hyser den svenska nationella forskarskolan i språkteknologi (GSLT – *Graduate School of Language Technology*). GSLT är ett samarbete mellan alla de svenska institutioner där högre utbildning i språkteknologi eller närliggande ämnen bedrivs och en GSLT-doktorand kan vara placerad vid vilken som helst av dem. Göteborgs universitet och Humanistiska fakulteten har garanterat fortsatt finansiering av GSLT åtminstone till 2012, efter det att det regeringens ursprungliga finansiella åtagande upphör år 2007.

Inom CLT kan man urskilja tre huvudforskningsteman:

1. Empiriskt baserad språkteknologi och korpuslingvistik
2. Formalismer för språkteknologi
3. Språkteknologiska system för kommunikation och undervisning

CLT omfattar i skrivande stund ungefär 20 disputerade forskare och ungefär 10 doktorander.

1.1 Not om Språkbanken och Institutionen för svenska språket

Språkbanken är en sedan länge etablerad institution för lagring, tillhandahållande och förädling av språkliga och språkteknologiska resurser för svenska.

Forskning, resurser, ekonomi och framtida behov för Språkbanken och Institutionen för svenska språket beskrivs dock utförligt i ett separat dokument som skickats in till DISC, varför vi hänvisar till detta för mer information.

2 Pågående språkteknologiforskning i CLT

I detta avsnitt beskrivs främst sådan pågående språkteknologiforskning som är beroende av och/eller avsätter språkteknologiresurser och -verktyg (se även avsnitt 3 nedan). I CLT bedrivs också forskning som har en mer indirekt relation till dessa och som av naturliga skäl inte kommer att nämnas ytterligare i detta sammanhang. Hit hör t.ex. teoretisk forskning om formalismer för beskrivning av olika aspekter av språket och dessa formalismers formella egenskaper. Sådan forskning bedrivs bl.a. vid Filosofiska institutionen och Institutionen för data- och informationsteknik.

2.1 Institutionen för data- och informationsteknik

Institutionen för data- och informationsteknik vid Chalmers tekniska högskola och Göteborgs universitet har en forskningsgrupp i språkteknologi med 5 seniora medlemmar och 5 doktorander. Gruppen deltar i CLT-samarbetet och i diverse externfinansierade projekt tillsammans med institutionerna för lingvistik och svenska språket. Ett av projekten, *Grammatiker som mjukvarubibliotek* (VR 2006–2008) har till syfte att skapa verktyg och resurser för tillämpningar av språkteknologi som en del av mjukvaruteknik.

2.1.1 Språkteknologisk forskning

Forskning inom språkteknologigruppen har två karaktäristiska drag: att vi ser språkteknologi som en integrerad del av mjukvaruteknik, och att vi intresserar oss för flerspråkiga tillämpningar. Det förra draget innebär att vi vill skapa resurser som är återanvändbara i olika tillämpningar, på samma sätt som mjukvarubibliotek. Det senare draget innebär dels att vi vill hitta lösningar som fungerar oberoende av språk, dels att vi samlar data om och bygger resurser för många språk, inte minst "exotiska" språk sådana som arabiska, finska och urdu.

En stor del av forskningen handlar om skapandet av mjukvara som möjliggör denna samling, organisering, och återanvändning av lingvistisk data. Den största arbetsinsatsen har lagts på GF (Grammatical Framework), som är ett programspråk för specifiering av grammatiker, i synnerhet grammatiker som relaterar flera språk med varandra. FM (Functional Morphology) är ett mjukvarupaket som möjliggör snabb utveckling av morfologiska lexika, och Extract är ett verktyg som kan användas för att skapa lexikala databaser halvautomatiskt från ostrukturerat textmaterial.

2.2 Institutionen för lingvistik

2.2.1 Skrift hos barn och ungdomar, IKT och datorbaserat skrivstöd

En projektgrupp bestående av FD Ylva Hård af Segerstad, FD Sylvana Sofkova Hashemi och professor Robin Cooper bedriver språkteknologisk forskning inriktad på barns och ungdomars skrivande, datormedierad kommunikation och utveckling av teknologier för automatisk lokalisering av målgruppens skrivproblem. Arbetet har bl.a. innefattat insamling av diverse textmaterial och systematisk lexikal och syntaktisk analys av det skrivna materialet. Forskargruppen har utvecklat en ny metod för att söka automatiskt efter grammatiska fel. Grammatikstödet FiniteCheck har utvecklats speciellt för barntexter och den nuvarande prototypen hittar subtantivfraser med kongruensfel och verb eller verbsekvenser med formfel. Systemet uppvisar lovande resultat vad gäller täckning av fel hos barn jämfört med system utvecklade för vuxna. Projektet *Att lära sig skriva i IT-samhället* startade i januari 2003 och finansieras av Vetenskapsrådet. Syftet är att undersöka skrift hos barn och ungdomar i olika skrivsituationer och eventuella effekter på deras texter som kan ha samband med deras användning av olika former av informations- och kommunikationsteknologier (IKT).

3 Språkteknologiresurser och -verktyg i CLT

Inom alla de tre uppräknade CLT-forskningstemana är man beroende av tillgång till språkteknologiresurser – data i form av korpusar, lexikon, grammatiker, taldatabaser, etc. – och språkteknologiverktyg, t.ex. ordklasstaggare, parsrar, namntaggare, tal-till-text- och text-till-tal-system, etc. Till denna forsknings natur hör dessutom att den avlägger sådana resurser (helt nya resurser eller ”förädlade” versioner av befintliga resurser) som behöver åtminstone ett minimum av underhåll ifall de ska förbli tillgängliga för forsknings- och utbildningssyften.

3.1 Institutionen för data- och informationsteknik

GF-resursgrammatikbiblioteket är ett mjukvarupaket som beskriver den grundläggande grammatiska strukturen av tio språk (danska, engelska, finska, franska, italienska, norska, ryska, spanska, svenska, tyska). Denna struktur innefattar dels ett fullständigt system för böjningsmorfologi, dels en språkoberoende beskrivning av syntax. Därutöver har vi i GF och FM skapat morfologiska beskrivningar av arabiska, fornsvenska och urdu. Med hjälp av Extract-verktyget och/eller anpassning av tillgängliga resurser har vi även skapat morfologiska lexika av flera av dessa språk, med storlek från 3000 till 20000 lemman.

Ett kännetecken av våra resurser är att de distribueras som fri mjukvara med öppen källkod.

3.2 Institutionen för lingvistik

På institutionen finns infrastrukturella resurser i form av talspråkskorpora, verktyg för att bearbeta korpora och uppbyggda nätverk till andra forskare som tillhandahåller korpora och nätverk. Våra korpora används kontinuerligt för både forsknings- och undervisningssyften.

3.2.1 Korpusar

Lite mer specifikt består talspråkskorpusen *Göteborg Spoken Language Corpus* (GSLC) av flera olika subkorpora:

- En kärnkorpus med ca. 1.5 miljoner ord svenskt talspråk inspelat i olika sociala verksamheter* (ca hälften videoinspelat, resten audioinspelat), allt transkriberat enligt GTS (Göteborg Transcription Standard) och MSO (Modifierad Standardortografi för svenska)
  Syfte: Skapad under ca 30 år inom olika projekt med olika syften, men med det övergripande syftet att åstadkomma en stor, verksamhetsbaserad korpus möjlig att analysera med dator.
  Aktivitetstyper: diskussion, återberättande av artikel, intervju, uppgiftsorienterad dialog, informellt samtal, rollspel, mässa, arrangerad diskussion, fomellt möte, konsultationssamtal, affärssamtal, middag, marknad, auktion, samtal i fabrik, fest, spel, telefornsamtal, resebyrå, rättegång, predikan, föreäsning, hotellsamtal, terapisamtal, busschaufför-passagerare
- En inlärarkorpus med vuxna invandrare från ESF-projektet ”Ecology of Adult Second Language Acquisition”, inspelad audio eller vido och transkriberad
- ett antal mindre subkorpusar med olika språk, patologiskt tal mm
- Insamlande av större jämförbara korpusar sker just nu i SIDA- och EU- projekt i Sydafrika och Nepal

Skrift hos barn och ungdomar, IKT och datorbaserat skrivstöd: Forskargruppen har samlat in texter skrivna med hjälp av papper och penna, ordbehandlare, e-post, SMS, chatt och dagböcker skapade på internet. Det insamlade materialet omfattar över 600 skoltexter och över 400 fritidstexter, varav drygt en tredjedel om 97 433 ord är transkriberat och analyserat. Korpusen följer transkriptions- och kodningsformatet för minCHAT (Codes for the Human Analysis of Transcripts) och analysverktyget CLAN (Computerized Language Analysis). Utöver textmaterialet finns även ett videoinspelat material med observationer av 14 elever skrivande på dator.

3.2.2 Nätverk

Inom NorFA-nätverket NORDTALK, som letts från institutionen, har korpusinsamlande och analys samt verktygsutvecklande samordnats mellan de nordiska länderna. Samordning sker även med forskningsgrupper i Sydafrika och Nepal (SOUTHTALK) och ytterligare samordning sker i WORLDTALK (nystartat nätverk).

Kostnaderna har legat huvudsakligen på externfinansierade projekt genom åren. Det är dock inte möjligt att i längden tillhandahålla och uppdatera korpusen utan mer kontinuerliga medel. Ett omfattande arbete med digitalisering av video- och audioinspelningar har pågått under flera år med viss finansiering från externa projekt och institutionen. Mer medel och uppdaterad utrustning för detta skulle behövas.

3.2.3 Verktyg

Det finns även ett antal verktyg för kodning, automatisk bearbetning och multimodal transkription, som utvecklats inom korpusprojekt: Corpus Browser, Gorallt statistiska mått, Multitool transkriptions- och kodningsverktyg etc. Även för dessa verktyg saknas idag pengar.

Nyttjandegraden är hög – flera avslutade, pågående och planerade projekt inom institutionen, samt studentprojekt och undervisning. GSLC används regelbundet av ett antal andra forskare i Sverige och andra länder.

4 Finansiering

Som nämnts ovan, är CLT som helhet en än så länge ofinansierad virtuell organisation. Vi kommer att söka Linnéstöd igen och vi utforskar även andra möjligheter (vi ligger inne med en ansökan om att få ta del av Göteborgs universitets strategiska medel för att kunna utvidga CLT:s verksamhet).

4.1 Institutionen för data- och informationsteknik

Under gruppens 7-åriga existens har dess forskning finansierats dels i form av fakultetsanslag och dels med externa anslag (ca. 50% av varje). Den externa finansieringen har kommit från VR (ca. 3 Mkr), Vinnova (4 Mkr), EU (2 Mkr) och GSLT (3 Mkr). Från och med början av 2007 är den enda säkra externa finansieringen VR-projektet *Grammatiker som mjukvarubibliotek*, ca. 600 tkr per år.

4.2 Institutionen för lingvistik

4.2.1 Skrift hos barn och ungdomar, IKT och datorbaserat skrivstöd

Insamlingen av skoltexter och fritidstexter har dels skett under de medverkande forskarnas avhandlingsarbeten och dels inom ramen för projektet *Att lära sig skriva i IT-samhället* som finansieras av Vetenskapsrådet sedan 2003 (c:a 3 Mkr). Projektet lider mot sitt slut och förväntas vara avslutat i februari 2007.

5 Framtidsplaner och förutsedda behov

5.1 Institutionen för data- och informationsteknik

Arbetet på flerspråkiga resurser som kan användas som mjukvarukomponenter har givit bra resultat, inte minst i form av samarbete med andra projekt som behöver sådan teknik som en del av sina tillämpningar. Speciellt bör nämnas EU-projektet WebALT (Web Advanced Learning Technology), som använder våra resursbibliotek för översättning av matematiska övningar till 7 språk. Detta projekt har lett till grundandet av ett företag, WebALT Inc, för att kommersialisera tekniken.

Vår närmaste plan är dels att utvidga våra språkresurser, dels att hitta samarbetspartners i industrin. Vi tillhandahåller våra resurser i en form som gör dem direkt an-

vändbara i mjukvarusystem sådana som teknisk översättning, mjukvarulokalisering, och människa-dator-interaktion. Dessutom har vi utvecklat verktyg för resursernas integrering som moduler i flera olika programspråk.

En konkret framtidsplan är utökningen av GF-resursgrammatikbiblioteket till samtliga officiella EU-språk, vilket gör sammanlagt 23 språk år 2007. 15 av dessa är inte ännu representerade i biblioteket. För att uppnå detta planerar vi att organisera en sommarskola med deltagare från alla involverade språk, kombinerad med självständigt arbete i allas hemländer. Detta görs i sammarbete med EU-nätverket JEM (Join Educational Mathematics), med det speciella syftet att täcka det matematiska språkbruket. Vi söker för närvarande finansiering för denna sommarskola, med beräknad budget 30 tEUR.

5.2 Institutionen för lingvistik

5.2.1 Talspråkskorpusar

Ändamål: Personalresurser för fortsatt digitalisering, drift och vidareutveckling av GSLC och korpusverktygen samt hjälp till forskare att utnyttja korpusen. Uppdaterad utrustning för digitalisering och nyinspelning.
Förväntad nyttjandefrekvens: Som nu — de flesta av institutionens forskare på alla nivåer samt ett ökande antal externa forskare, t ex blir talspråkskorpusen allt mer använd för utveckling av dialogsystem. Fortsatt och ökande samarbete med forskare i Norge, Danmark och övriga nordiska länder, Tyskland (Bielefeld), Österrike (Wien), Sydafrika (Pretoria) och Nepal samt med flera forskare i USA. Planerad EU-ansökan om projekt.

5.2.2 Skrift hos barn och ungdomar, IKT och datorbaserat skrivstöd

Forskargruppens mål är att via språkteknologiska analyser av barns texter utveckla ett språkligt, pedagogiskt och funktionellt anpassat datorbaserat skrivstöd som möter skribenterna i deras språk- och skrivutveckling. Arbetet innefattar en vidareanalys och transkription av det insamlade textmaterialet och en vidareutveckling av metoder för automatisk analys av de skrivproblem som texterna kännetecknas av. Planer finns att studera vidare skrivprocessande på dator via observationer och videoinspelning och dessutom att utöka korpusen med textmaterial från barn och ungdomar med svenska som andraspråk.

6 Beräknade kostnader i ett 5-årsperspektiv

6.1 Institutionen för lingvistik

- 2–5 heltidsekvivalenter — programmerare/ systemutvecklare, forskare/ korpus- och insamlingsansvariga, ledning av arbetet med insamling, utveckling och användning av korpusarna
- 100–200% amanuens/assistent, bl.a. för diverse analys- och kodningsarbete
- Utrustning: ca 50 000 för lagringsmedia, utrustning för digitalisering, inspelning mm.

Författare av detta dokument

- Lars Borin, professor i språkvetenskaplig databehandling, Språkbanken, Institutionen för svenska språket
- Aarne Ranta, professor i datavetenskap, Institutionen för data- och informationsteknik
- Jens Allwood, professor i allmän språkvetenskap, Institutionen för lingvistik
- Sylvana Sofkova Hashemi, forskare i språkteknologi, Institutionen för lingvistik

Databasresurser inom språkteknologi: Språkbanken

1. Introduktion

Språkbanken <http://spraakbanken.gu.se> är en administrativ enhet inom Institutionen för svenska språket, Göteborgs universitet. Språkbankens budget fördelas dock direkt av Humanistiska fakulteten. Språkbanken inrättades 1975 som ett nationellt centrum med nationell finansiering och uppdraget att samla in, bearbeta och lagra (svenska) textkorpusar (stora textsamlingar, systematiskt sammanställda för att kunna användas i språkvetenskaplig och språkteknologisk forskning) och vidare, för bruk i forskning och utbildning liksom för den intresserade allmänheten, tillgängliggöra lingvistiska data utvunna ur korpusarna och även tillgängliggöra andra språkteknologiresurser, såsom elektroniska lexikon och termlistor. Även om Språkbanken inte längre finansieras nationellt, så uppfattas den i mångt och mycket fortfarande som en nationell resurs med konstant hög nyttjandegrad i svensk forskning och i viss utsträckning även utanför Sverige (se nedan).

Idag besitter Språkbanken en unik kompetens inom områdena svenska textkorpusar, parallella textkorpusar, svenska elektroniska lexikon samt språkteknologiverktyg för bearbetning, annotering och presentation av textkorpusar och elektroniska lexikon, parad med en stabil organisation för permanent lagring, underhåll och tillhandahållande av dessa resurser. Av den anledningen fungerar Språkbanken som tillhandahållare även av resurser som har tillkommit utanför Göteborg, t.ex. SUC (Stockholm Umeå Corpus; Stockholms/ Umeå universitet), inlärarkorpusarna ASU (Andraspråkets strukturutveckling; Stockholms universitet) och SVANTE (Svenska andraspråkstexter; KTH/ Stockholms/ Göteborgs universitet), grammatikövningsplattformen ITG (IT-baserat kollaborativt lärande i grammatik Uppsala/ Stockholms/ Göteborgs universitet) och FTS, en stor färöisk tidningskorpus (Göteborgs universitet/ Fróðskaparsetur Føroya). Likaså har Språkbanken av samma skäl anförtrotts att svara för utveckling, underhåll och drift av Litteraturbanken <http://litteraturbanken.se>.

2. Språkteknologisk forskning nära och inom Språkbanken

I Språkbanken och dess närområde bedrivs språkteknologisk forskning på flera håll. I de flesta fall använder och/eller skapar denna forskning språkteknologiresurser:

(1) I själva Språkbanken pågår kontinuerligt ett arbete med att ”förädla” de befintliga resurserna (tillföra lingvistisk information till dem) så att de ska kunna användas för nya forskningsuppgifter, liksom även med att skapa nya resurser, en verksamhet som i sig till stora delar är att betrakta som forskning. Likaså arbetar vi på att göra resurser från olika projekt kompatibla med varandra samt på att göra språkteknologiverktyg utvecklade vid institutionen (och även verktyg utvecklade på andra håll) allmänt tillgängliga för annotering av Språkbankens korpusar, för annotering av Litteraturbankens texter, samt för annotering av externa användares egna textmaterial.

(2) Annars bedrivs den språkteknologiska forskningen vid Institutionen för svenska språket huvudsakligen inom forskningsgruppen för språkvetenskaplig databehandling (Språkdata), där projekten typiskt både använder befintliga språkteknologiska resurser i Språkbanken och avsätter nya eller förädlade resurser som i många fall kan göras tillgängliga för andra forskare genom Språkbanken. I skrivande stund (november 2006) pågår bl.a. följande projekt: (2a) Semantic Mining, ett EU-NoE inom (flerspråkig) biomedicinsk informationssökning och informationsextraktion, som bl.a. använder ett elektroniskt lexikon (*Svenska ord*) och som resultat kommer att avsätta ett svenskt elektroniskt medicinskt lexikon, en annoterad korpus över svenskt medicinskt språk och ett flertal språkteknologiverktyg; (2b) ett planeringsprojekt finansierat av Humanistiska fakulteten, där tesen är att språkteknologiverktyg kan användas för att skapa nya möjligheter i textbaserad humanistisk forskning och utbildning, specifikt språklärande; (2c) flera av de moderna svenska referensordböckerna har tillkommit och tillkommer vid Språkdata, och kring dessa finns en livlig lexikologisk forskning som använder framförallt Språkbankens korpusar. Den har även resulterat i att elektroniska lexikon har kunnat göras tillgängliga i Språkbanken (t.ex. det tidigare nämnda *Svenska ord*). Mer specifikt språkteknologiska lexikologiska projekt finns också, t.ex. ett pågående arbete på ett svenskt ramsemantiskt lexikon, samt ett arbete som syftar till att göra en uppdaterad elektronisk

version av *Svenskt associationslexikon* (Uppsala universitet/ Universitettet i Tromsø/ Göteborgs universitet) tillgänglig; (2d) forskare från Språkdata samarbetar i några projekt vid Institutet för svenska som andraspåk (ett nationellt finansierat forsknings centrum placerat vid Institutionen för svenska språket), där språkliga material undersöks utifrån ett andraspråksperspektiv och där Språkbankens korpusar används som jäm förelsebas i sådana undersökningar. Dessa projekt förväntas vidare avsätta nya korpusar i Språkbanken, bl.a. s.k. inlärarkorpusar, alltså texter på svenska producerade av icke-modersmålstalare.

(3) Språkteknologiforskning bedrivs i Göteborg vid flera institutioner (Filosofi, Lingvistik, Svenska språket vid Göteborgs universitet och Datavetenskap vid Chalmers). Humanistiska fakulteten driver dessutom den nationella forskarskolan i språkteknologi (GSLT). I Göteborg pågår således en mycket livlig forskning och utbildning inom området. Av den anledningen och för att avspegla det faktum att samarbetet mellan de inblandade institutionerna är både djupt och går långt tillbaka i tiden finns ett t.v. informellt forum för samarbete och samverkan mellan de inblandade forskarna som vi har döpt till *Centre for Language Technology* (CLT). Språkdata och datavetenskap sam arbetar f.n. om att skapa två morfologiska resurser som kommer att både användas internt i Språkbanken och göras tillgängliga för andra forskare via Språkbanken. Det handlar om datamaskinella morfologiska beskrivningar (som alltså kan användas för automatisk morfologisk analys eller generering av textord) för modern svenska (för ord förrådet i *Svenskt associationslexikon*; se ovan) och för fornsvenska (för ordförrådet i Söderwalls och i Schlyters fornsvenska lexikon). I de vidare planerna ingår att skapa liknande resurser för färöiska och 1800-talssvenska (ordförrådet i Dalins lexikon), där den sistnämnda resursen blir särskilt värdefull med tanke på den stora mängden äldre text som kommer att finnas i Litteraturbanken.

Storleken på de aktuella forskningsgrupperna är:

(1) Språkbanken (inkl. Litteraturbanken): ½ forskare, c:a 3 systemutvecklare
(2) Språkdata: c:a 5 forskare, c:a 3 doktorander, c:a 3 systemutvecklare
(3) CLT: c:a 20 forskare, c:a 10 doktorander (inklusive dem i Språkdata/Språkbanken)

3. Språkteknologiresurser och -verktyg

Tabellen på sista sidan av denna rapport ger en översikt över korpus- och lexikonresurser i Språkbanken (och Litteraturbanken). När det gäller språkteknologiverktyg, så har Språkdatas projekt genom åren resulterat i en rad sådana verktyg: tokeniserare, flera ordklasstaggare, heuristisk lemmatiserare, syntaktisk parser för modern svenska, svensk namnigenkännare som senare har utvidgats till att även känna igen och märka upp medicinska termer (organismer, anatomiska termer, sjukdomar), böjningsmorfologiska fullformslexikon.

I tillägg till detta kan nämnas att vi bedriver ett ständigt arbete med att förbättra våra korpussökverktyg, åtkomliga via Språkbankens webbplats <http://spraakbanken.gu.se/>, eller för experimentella verktyg och gränssnitt på <http://demo.spraakdata.gu.se/>.

Språkbankens resurser har tillkommit huvudsakligen till stöd för mer traditionell språk forskning, och det är fortfarande deras huvudanvändning. Under september 2006 sände Språk bankens föreståndare Lars Borin ut en e-postförfrågan till institutioner för svenska/nordiska språk vid universitet i Sverige och Finland. Förfrågan handlade om institutionernas använd ning av Språkbanken i forskning, undervisning och för tredje uppgiften (t.ex. populär vetenskap). De flesta institutionerna svarade på förfrågan. Av svaren framgår att Språkbanken används flitigt för forskning (ett antal publikationsreferenser bifogades svaren; bland annat har flera doktorsavhandlingar använt Språkbanken som källa) och för tredje uppgiften (svar på språkbruksfrågor hänvisar ofta till Språkbanken), men inte i så stor utsträckning i under visning. Man påpekade också i svaren att det är fantastiskt att tillgången till Språkbanken är fri, både i bemärkelsen kostnadsfri för brukaren, men också i den bemärkelsen att för åtkomst till huvuddelen av resurserna krävs inget inloggningsförfarande (vilket krävs t.ex. i den finska språkbanken). Språkbanken används också i viss, men av naturliga skäl relativt liten utsträck ning utanför Norden, vilket framgår av frågor som kommer in till Språkbankens hjälp-e-postlåda. Här handlar det nästan uteslutande om forskare i svenska/nordiska språk, men även en och annan person med någon kontrastiv problemställning.

Språkbanken har således en väl etablerad roll som ”lagerhållare” och ”leverantör” av empiriska (skrift)språkliga data för forskning om svenska språket ur alla synvinklar (inklusive kontrastiva). Språkbankens traditionella användning är ”inspektion”, alltså att användare kan ställa sökfrågor till korpusarna och få tillbaka ett eventuellt resultat i form av en s.k. rad konkordans, alltså sökträffarna visade mitt på varsin rad (uppräknade i textordning på skärm en) med ett litet stycke kontext på ömse sidor. med den vidare möjligheten att genom ett mus klick beställa fram en större kontext för utvalda intressanta träffar. Inspektion kan vara tillfyllest för somliga typer av språkteknologiforskning, men i allmänhet kräver språktekno logiforskningen tillgång till resurserna (korpusar, lexikon, etc.) i deras helhet, för träning av språkteknologiverktyg genom s.k. maskininlärning, men också som allmänt tillgängliga referensdatamängder för reproducerbar och jämförbar utvärdering av språkteknologiverktyg. I viss mån kan vi tillhandahålla resurser även på det viset, t.ex. kan den svenska Parole-korpusen (19 miljoner ord ordklasstaggad text) och baslexikonet *Svenska ord* (c:a 20 000 ord) laddas ner från Språkbanken till svenska universitetsdomäner (definierade som vissa IP-nummer) för användning i forskning och undervisning. Speciellt Parole-korpusen åtnjuter en stadig efterfrågan som en av de få resurser som är tillgängliga på detta sätt för svenska. Vi arbetar på att utöka mängden material som ska vara fritt tillgängliga på detta sätt, men här är framförallt de upphovsrättsliga frågorna besvärliga. Närmast på tur för tillhandahållande som fri resurs är ett stort svenskt lexikon (c:a 72 000 ingångar) med semantisk och morfologisk information (*Svenskt associationslexikon*; se ovan).

Som framgått av föregående avsnitt är många av resurserna resultat av samarbeten, där en av parterna har varit Språkdata oich det därigenom har fallit sig naturligt att Språkbanken skulle svara för lagring och tillhandahållande av de resurser som projekten resulterade i; detta gäller åtminstone SVANTE, ITG och FTS av de tidigare nämnda resurserna, samt ytterligare många av resurserna i tabellen på sista sidan i denna rapport, Speciellt kan där nämnas Parole- och MEDLEX-korpusarna samt AVENTINUS, som är resultat av EU-projektsamarbeten där Språkdata har medverkat.

Som kanske också har framgått av det föregående, är Språkbanken mer än ett ”rent” data arkiv, på grund av sin nära koppling till den forskning som ger upphov till resurserna. Språk bankens föreståndare (50% av heltid) är tillika ämnesföreträdare i Språkvetenskaplig data behandling (50% av heltid), och den konstruktionen är avsiktlig. Det betyder att Språkbanken ibland kan framträda som forskningsinrättning och ibland som databasarkiv, ett arrangemang som passar ovanligt bra inom ett område som språkteknologi, där språkteknologiresurser och metod- och standardiseringsfrågor kring dessa har kommit i skarpt fokus på senare år. Språk banken/ Språkdata deltar i diskussionen av dessa frågor, i Sverige (där vi samordnade en KFI-planeringsansökan för en svensk nationell korpus och en svensk basresursuppsättning för språkteknologi), i Norden (där vi verkar inom det nybildade *North European Association for Language Technology*, som har språkteknologiresurser för språken i Norden och Baltikum som en prioriterad fråga) och i Europa (där vi deltar i PAROLE-samarbetet för skapande och underhåll av europeiska språkteknologiresurser). Vi har ännu ingen direkt representation i de fora där internationella standarder på detta område utarbetas (TEI, XCES, ISO TC37/SC4, OLAC, ISLE, etc.), men vi är väl medvetna om detta arbete, vi tar hänsyn till det i vårt pågående arbete med att ”framtidssäkra” Språkbanken, och kommer att i möjligaste mån sträva efter att också kunna bidra till det mot bakgrund av vårt praktiska arbete med svenska resurser. Språkbanken har sökt medel från DISC-utlysningen 2006 för att mer koncentrerat kunna arbeta med att anpassa den brokiga floran av resurser och verktyg inbördes och till dessa framväxande internationella standarder. För detta sökte vi motsvarande 1,75 anställ ningar på två år (3,5 personår).

4. Finansiering

De enskilda resurserna i Språkbanken har i stor utsträckning finansierats med nationella offentliga medel, antingen universitetens fakultetsanslag eller projektmedel från statliga forsk ningsråd. Somliga resurser har också kommit till med finansiering från svenska icke-statliga forskningsfinansiärer eller med EU-medel. Det är svårt att uppskatta hur mycket pengar som sammanlagt har lagts på dessa resurser till dags dato; en snabb och grov uppskattning ger vid handen åtminstone 40 MSEK över de senaste 40 åren, men förmodligen är den verkliga summan betydligt högre. För drift och underhåll av Språkbanken avsätter Humanistiska fakul

teten vid Göteborgs universitet en årlig summa, som f.n. täcker knappt 2,5 heltidsanställningar och en del kringkostnader. För Litteraturbankens uppbyggnad, drift och underhåll får vi f.n. motsvarande 1,3 heltidsanställningar från den ideella stiftelsen Litteraturbanken (i praktiken kommer pengarna från Svenska Akademien), plus en del kringkostnader. När det gäller fakultetens finansiering för Språkbanken kan man förvänta sig att den kommer att ligga kvar på samma nivå under överskådlig framtid. Däremot kommer förmodligen andelen från Litteraturbanken att minska, till någonstans mellan halv och en hel anställning över de närmaste åren, allteftersom uppbyggnadsfasen övergår i en permanent verksamhet.

Språkbankens datorutrustning (servrarna: utvecklings-, databas- och publiceringsmaskiner) finansieras delvis med fakultetsmedel. År 1999 fick Språkbanken ett utrustningsanslag från Knut och Alice Wallenbergs Stiftelse, och vi kommer inom kort att lämna en ny ansökan till dem. Om den inte beviljas kommer Språkbanken i ett bekymmersamt läge, eftersom vår basutrustning börjar uppvisa ålderstecken och det rör sig om en investering som är för stor för Humanistiska fakultetens redan hårt ansträngda budget.

5. Framtidsplaner och förutsedda behov

Språkbanken har växt ”organiskt” över fyra decennier. Det betyder som redan nämnts att resurser och verktyg har olika format som inte alltid är kompatibla med varandra, liksom att olika resurser är förädlade i olika grad. Det betyder också att det finns lakuner i Språkbankens täckning av det svenska skriftspråket. De största lakunerna är diakrona: Det behövs mer material från många perioder i den den svenska språkhistorien. Detta är ett problem huvudsakligen för mer traditionella språkvetare. För en språktenologiforskare är det mer bekymmersamt att täckningen av moderna svenska (skrivna) genrer är bristfällig. Här behövs systematiska insatser. Vår KFI-planeringsansökan handlar bland annat om detta, nämligen den del av denna som benämns Svensk Nationell Korpus (SNK), som tänks bestå av minst 100 miljoner ord balanserad svensk text, varav minst 10 miljoner ord skall vara transkriberat tal, och hela korpusen skall vara ordklassuppmärkt och 10% syntaktiskt uppmärkt. Detta kan vi i Språkbanken inte åstadkomma ensamma (i synnerhet inte talspråksdelen men heller inte det omfattande arbete som kan förutses för att hantera upphovsrättsliga frågor) och i KFI-ansökan är följaktligen samtliga svenska språkteknologiforskningsmiljöer representerade.

Vi bedriver i blygsam takt det förbättringsarbete som har nämnts ovan, men inom ordinarie budget kommer det att ta lång tid innan några resultat blir synliga, eftersom det som till slut syns på ytan förutsätter ett omfattande arbete med den underliggande infrastrukturen.

Drift och visst utvecklingsarbete av Språkbanken och Litteraturbanken klarar vi inom befintlig budget (c:a 3,5 heltidsekvivalenter per år under de närmaste åren), men för en grundläggande modernisering av Språkbankens infrastruktur och för att göra samtliga (moderna svenska) korpusresurser likvärdiga m.a.p. lingvistisk förädlingsgrad kommer vi att behöva ett tillskott av medel motsvarande 3,5 personår över två år. För en utveckling av nya resurser enligt KFI-ansökan talar vi om helt andra siffror: Där beräknade vi preliminärt (en viktig uppgift för det sökta planeringsprojektet blir att ta fram säkrare siffror) att en svensk nationell korpus skulle kosta i storleksordningen 50-75 MSEK och en basuppsättning svenska språkteknologiresurser 40-50 MSEK att förverkliga, men vi såg också stora potentiella synergieffekter i att arbeta med dessa två resurser samtidigt, så att de sammanlagda kostnaderna skulle kunna nedbringas. Vi kommer också inom kort att behöva förnya vår maskinpark till en kostnad av i storleksordningen 1,5-2 MSEK (se föregående avsnitt),

Lars Borin
Föreståndare för Språkbanken
Professor i språkvetenskaplig databehandling
Institutionen för svenska språket
Göteborgs universitet
tel. 031 773 4544, 070 747 8386
<lars.borin@svenska.gu.se>

Svenska korpusar i Språkbanken		**storlek (löpord)**
Press 65 (dagstidningstext)		990 989
Press 76 (dagstidningstext)		1 156 958
DN 1987 (dagstidningstext)		4 132 784
Press 95 (dagstidningstext)		6 769 649
Press 96 (dagstidningstext)		5 755 168
Press 97 (dagstidningstext)		11 900 570
Press 98 (dagstidningstext)		9 239 336
SVD 00 (dagstidningstext)		13 131 043
GP 01 (dagstidningstext)		15 257 883
GP 02 (dagstidningstext)		18 434 005
GP 03 (dagstidningstext)		16 663 701
GP 04 (dagstidningstext)		19 406 813
Stockholm Umeå Corpus (balanserad; modernt publicerat skriftspråk)		1 166 590
SYNTAG (syntaktiskt annoterad dagstidningstext)	c:a	100 000
TB/Bruksprosa (syntaktiskt annoterat blandat publicaert skriftspråk)	c:a	87 000
PAROLE corpus (ordklasstaggat blandat publicerat skriftspråk)	c:a	19 000 000
SVANTE (skriftlig inlärarkorpus)		204 398
ASU (inlärare + infödda; tal + skrift)	c:a	730 000
Forskning & Framsteg (populärvetenskap)		669 893
Äldre svenska romaner (sent 1880- och tidigt 1900-tal)		3 702 748
Bonniersromaner I (moderna romaner, 1976/77)		5 626 348
Bonniersromaner II (moderna romaner, 1980/81)		3 715 690
Strindbergs brev		1 223 288
Strindbergs romaner och dramer		2 461 426
SAOL 11 (Svenska Akademiens ordlista, 11:e upplagan, som text)		404 596
Svenska dagbladets årsbok 1923-1958 (dagstidningstext)	c:a	1 500 000
Psalmboken (1937)		111 304
Svensk författningssamling 1978-81		612 688
Bellmans samlade verk	c:a	360 000
Riksdagens snabbprotokoll 1978-79		4 420 767
Källtext (fornsvenska)		1 096 244
Manuductio (skrift om poesi från 1651)		28 202
MEDLEX-korpusen (medicinsk text)	c:a	10 000 000
Litteraturbanken (skönlitteratur)	c:a	1 500 000
SAOB (Svenska Akademiens ordbok som text)		28 375 720
Totalt svenska korpusar	*c:a*	*209 935 801*
icke-svenska korpusar i Språkbanken		
Panegyrici (latin)		86 442
The Elder Edda (fornisländska)		36 764
FTS (modern färöiska)		11 000 000
SOL (modern spanska)	c:a	3 800 000
Totalt icke-svenska korpusar	*c:a*	*14 923 206*
Totalt i Språkbanken	***c:a***	***224 859 007***

lexikondata i Språkbanken (svenska och två/flerspråkiga lexikon och termbaser))	**storlek (ung. antal uppslag)**
Söderwall (formsvenska)	23 000
Söderwall, supplement (fornsvenska)	20 000
Schlyter (fornsvenska lagarnas ordförråd)	16 000
SAOB	51 000
Svenska ord (svensk basvokabulär)	20 000
Svenskt associationslexikon (ett slags tesaurus)	72 000
TERMIN (tvåspråkiga samhällstermlistor)	4 400
AVENTINUS (flerspråkig narkotikabekämpningsrelaterad terminologi)	15 000
Totalt i Språkbanken	***221 400***

Svar för Allmän språkvetenskap
Institutionen för lingvistik
Göteborgs universitet

1)

Nuvarande projekt inom ASV:

Projekt som använder institutionens talspråkskorpus GSLC (Göteborg Spoken Language Corpus),

dvs
Olika doktorandprojekt, bl a om utvecklande av situationsanpassade talhjälpmedel (Allm. Arvsfonden)
Projekt som studerar multimodalitet och embodiment i mänsklig kommunikation för att utveckla ECAs (Embodied Communicative Agents) (VR)
Projekt som gör jämförelser mellan talspråk i olika nordiska länder – SweDanes, NORDTALK (NorFA)
Projekt om talspråk i Sydafrika och Nepal (SIDA, EU)
Projekt som jämför talspråk med patologiskt tal vid afasi (FAS, VR)
Projektet ”Språk och språkbruk bland ungdomar i flerspråkiga storstadsmiljöer” har samlat in en stor talspråkskorpus av spontant och eliciterat material med ungdomar i olika miljöer.
Ett antal studentuppsatser inom kurser och som C- och D-uppsatser som studerar talspråk.
Projekt om emotioner och kommunikation (GU forskarassistenttjänst)
Projekt om talspråk och dialogsystem (GU forskarassistenttjänst)

Talspråkskorpusen används kontinuerligt av 10-20 personer vid institutionen eller med nära anknytning till den. Det gäller forskare på alla nivåer: professorer, forskarassistenter, lektorer, doktorander, C- och D-studenter, Breddmagisterstudenter i Kommunikation, studerande på kurser i pragmatik, ickeverbal kommunikation m fl kurser.

2)

Talspråkskorpusen Göteborg Spoken Language Corpus, som består av flera olika sub.korpusar:
Se även: http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3

- en kärnkorpus med 1.5 ord svenskt talspråk inspelat i olika sociala verksamheter* (ca hälften videoinspelat, resten audioinspelat), allt transkriberat enligt GTS (Göteborg Transcription Standard) och MSO Modifierad Standardortografi för svenska. Det totala materialet består av ca 2 000 timmars tal, varav uppskattningsvis 400 timmar är digitaliserade.
  Syfte: Skapad under ca 30 år inom olika projekt med olika syften, men med det övergripande syftet att åstadkomma en stor, verksamhetsbaserad korpus möjlig att analysera med dator.

*** Activity Types**

- Discussion
- Retelling Of Article
- Interview

    - Task-Oriented Dialogue
    - Informal Conversation
    - Role Play
    - Trade Fair
    - Arranged Discussions
    - Formal Meeting
    - Consultation
    - Shop
    - Dinner
    - Market
    - Auction
    - Factory Conversation
    - Party
    - Games & Play
    - Phone
    - Travel Agency
    - Court
    - Church
    - Lecture
    - Hotel
    - Therapy
    - Bus Driver-Passenger
- en inlärarkorpus med vuxna invandrare från ESF-projektet ”Ecology of Adult Second Language Acquisition”, inspelad audio eller vido och transkriberad
- ett antal mindre subkorpusar med olika språk, t ex 90 timmar inspelningar av finska talad i Sverige
- ett antal mindre subkorpusar med patologiskt tal mm
- insamlande av större jämförbara korpusar sker just nu i SIDA- och EU-projekt i Sydafrika och Nepal
- Nyttjandegraden är hög – flera avslutade, pågående och planerad projekt inom institutionen, samt studentprojekt och undervisning. Korpusen används regelbundet av ett antal andra forskare i Sverige och andra länder.
- Inom NorFA-nätverket NORDTALK, som letts från institutionen, har korpusinsamlande och analys samt verktygsutvecklande samordnats mellan de nordiska länderna. Samordning sker även med forskningsgrupper i Sydafrika och Nepal.
- Kostnaderna har legat huvudsakligen på externfinansierade projekt genom åren. Det är dock inte möjligt att i längden tillhandahålla och uppdatera korpusen utan mer kontinuerliga medel. Driftskostnaderna har idag ingen täckning.
- Det finns även ett antal verktyg för kodning, automatisk bearbetning och multimodal transkription, som utvecklats inom korpusprojekt: Corpus Browser, Gorallt statistiska mått, Multitool transkriptions- och kodningsverktyg etc. Även för dessa verktyg saknas idag pengar.
- Ett omfattande arbete med digitalisering av video- och audioinspelningar har pågått under flera år med viss finansiering från externa projekt och institutionen. Mer medel och uppdaterad utrustning för detta skulle behövas.

3)

Behov av resurser:

Ändamål:
Personalresurser för fortsatt digitalisering, drift och vidareutveckling av korpusen och korpusverktygen samt hjälp till forskare att utnyttja korpusen.
Uppdaterad utrustning för digitalisering och nyinspelning.

Förväntad nyttjandefrekvens:
Som nu – de flesta av institutionens forskare på alla nivåer samt ett ökande antal externa forskare, t ex blir talspråkskorpusen allt mer använd för utveckling av dialogsystem.

Fortsatt och ökande samarbete med forskare i Norge, Danmark och övriga nordiska länder, Tyskland (Bielefeld), Österrike (Wien), Sydafrika (Pretoria) och Nepal samt med flera forskare i USA. Planerad EU-ansökan om projekt.

Beräknade kostnader i ett 5-årsperspektiv.
1-2 heltidsekvivalenter – programmerare/korpusansvariga och utvecklare
50-100% amanuens/assistent
Utrustning: ca 50 000 för lagringsmedia, utrustning för digitalisering, inspelning mm.
ca 20% av forskartid för att leda och delta i fortsatt utvecklande av korpusen

Ev utsikter till finansiering:
Vi har ansökt om medel för långsiktigt stöd till databaser från VR. I övrigt saknas finansiering med undantag för mindre del i tjänst för en lönebidragsanställd.

Linköpings universitet

Till Vetenskapsrådet/DISC,
att: Eva Strangert

Databasresurser inom språkteknologi – läget vid NLPLab/Institutionen för datavetenskap, Linköpings universitet

DISC (Database Infra-Structure Committee) har påbörjat en kartläggning av databasresurser inom språkteknologi och som ett led i detta begärt information från ett antal institutioner som bedriver språkteknologisk forskning. Detta brev speglar situationen vid Institutionen för datavetenskap vid Linköpings universitet och särskilt forskningsgruppen för databehandling av naturligt språk, NLPLab.

1. Pågående och planerad forskning

NLPLab bedriver forskning inom två större språkteknologiska tillämpningsområden, översättningsteknologi och dialogsystem/interaktiva frågebesvarande system, och på grundläggande algoritmer och metoder av relevans för dessa tillämpningsområden. Dessa områden kommer att vara våra huvudområden även för de närmaste åren.

På översättningsområdet arbetar två seniora forskare och två doktorander och på dialogsystem likaledes två seniora forskare, en forskarassistent och tre doktorander.

Forskningen förutsätter tillgång på empiriska data. Mycket av den grundläggande forskningen handlar om metoder för att generera sekundära data, t.ex. tvåspråkiga lexikon eller begreppshierarkier, utifrån primärdata i form av textkorpusar. I dagsläget använder vi i stor utsträckning data som vi samlat in i olika projekt med begränsad tillgänglighet för andra forskare. Vi använder även allmänt tillgängliga data, som ofta tagits fram av utländska forskare och forskningsfinansiärer, t.ex. i samband med olika kampanjprojekt där forskare från hela världen inbjuds att utveckla system för en given uppgift. Exempel på sådana data inom översättningsområdet är ARCADE-korpusen för

Institutionen för datavetenskap
HCS/Lars Ahrenberg
581 83 Linköping
tel: 013 - 282422
E-post: lah@ida.liu.se

franska-engelska (Véronis & Langlais, 2000)[1] och Europarl för 11 EU-språk (Koehn, 2004)[2].

2. Befintliga forskningsdatabaser vid NLPLab

De data vi förfogar över föreligger främst i form av XML-formaterade textfiler. Ibland utgörs datan av rena, dvs oannoterade, textfiler men oftast är de uppmärkta också med lingvistisk information som anger lemma, ordklass med mera. I de flesta fall är den lingvistiska informationen automatgenererad och ej manuellt granskad.

Data som samlats in i dialogrelaterade projekt utgörs i några fall av transkriptioner från taldata, t.ex. trafikupplysning, men är i de flesta fall skrivna dialoger insamlade med Wizard-of-Oz-teknik.

Det mesta av dessa data samlades in med specifika projektändamål i åtanke, alltså inte i syfte att bygga större databaser med lång livslängd. Vi har dock i en del fall gjort data tillgängliga via vår hemsida[3]. Det finns ett större undantag och det är det vi kallar för Linköpings översättningskorpus, som började samlas in i mitten på 90-talet och sedan utökats efter hand i flera projekt som Plug, Transmap och KOMA. I projektet ”Mikro- och makroanalys av en översättningskorpus” har en del av detta korpusmaterial använts för ett pilotprojekt med en engelsk-svensk parallell trädbank (LinES).

Det är främst gruppens egna forskare som nyttjar våra data. När projekten inneburit samarbete med andra forskningsgrupper i landet har också dessa använt materialet. Omvänt har då också vi använt data som samlats in av andra grupper.

Uppbyggnaden av översättningskorpusen har också inneburit samarbete med utomstående. Exempelvis är romanmaterialet i korpusen tillhandahållet av Språkbanken i Göteborg, medan manualmaterialet erhållits via samarbete med Microsoft och IBM.

[1] Jean Véronis and Philippe Langlais: Evaluation of parallel text alignment systems. In Jean Véronis (ed.) *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer Academic Publishers, 2000.
[2] Philipp Koehn: Europarl: A Multilingual Corpus for Evaluation of Machine Translation. http://people.csail.mit.edu/koehn/publications/europarl/
[3] Se http://www.ida.liu.se/~nlplab/resources/corpora.shtml

Institutionen för datavetenskap
581 83 Linköping

Samordning med andra i fråga om formatering, dokumentation och standarder har endast skett i samband med projektsamarbete men då på temporär basis, dvs för det aktuella syftet. Kostnader har belastat pågående projekt. Någon fast finansiering för översättningskorpusen finns inte utan det arbete som utförs sker inom ramen för pågående projekt eller på arbetstid för forskning.

3. Framtida behov av resurser

Vår egen forskning, och, som vi tror, även annan svensk språkteknologisk forskning med inriktning på översättningssystem resp. dialogsystem skulle ha mycket att vinna på en uppbyggnad av omfattande, lingvistiskt uppmärkta dataresurser, som är allmänt tillgängliga, ev. för självkostnadspris. Detta skulle vara till stor gagn inte bara för språkteknologin utan också för annan forskning med inriktning på översättning resp. dialog.

En nationell översättningskorpus

Linköpings översättningskorpus är i dag begränsad till ett språkpar, engelska-svenska, en översättningsriktning och fyra genrer. En översättningskorpus av nationellt värde borde omfatta åtminstone ett tiotal språkpar (där svenska är ett av språken), båda översättningsriktningarna, och ett tiotal genrer samt vara uppmärkt med manuellt granskad lingvistisk information. Omfattningen av materialet är beroende av genre och språk, men för de språkpar och genrer där så är möjligt borde en omfattning av 500,000 källord totalt fördelat på ca 15,000 källord per verk/delgenre vara eftersträvansvärd.

Vi känner till flera andra översättningskorpusar som byggts upp i Sverige. Ett stort problem är att dessa, liksom Linköpings översättningskorpus, inte är allmänt tillgängliga utan tillhandahållna med begränsningar av copyright-innehavare, eller insamlade utan dessas tillstånd och därför inte sprids alls. I princip är det annars lätt att göra data tillgängligt och sökbart, t.ex. via Internet och webbgränssnitt, vilket också möjliggör en distribuerad lagring.

Kostnaden för en översättningskorpus av denna storlek är svår att uppskatta, men 1 mkr per språkpar (1,5-2 årsarbeten) förefaller inte orealistiskt. Den är beroende bl.a. av tillgången på material för vilket spridningsrättigheter kan utverkas, granskare och befintligheten av effektiva analys- och uppmärkningsverktyg.

Institutionen för datavetenskap
581 83 Linköping

En nationell dialogkorpus
Även dialogsystemforskningen skulle vinna på tillgång till en större dialogkorpus med data både från människa-dator situationen och människa-människa-situationen. I en sådan korpus skulle NLPLabs korpusar kunna ingå, men det vore synnerligen värdefullt att på ett samlat sätt få tillgång till data som samtalsforskare på olika håll i Sverige samlat in under årens lopp. I detta sammanhang vill jag särskilt nämna Tema Kommunikation i Linköping som ju studerat en mångfald olika samtalsgenrer med utgångspunkt i empiriska data.

Lexikon och ordnät
För språkteknologisk systemutveckling med inriktning på svenska är också utvecklingen av allmänt tillgängliga en- och flerspråkiga lexikon och ordnät väsentlig.

Linköping, 16 november 2006

Lars Ahrenberg

Institutionen för datavetenskap
581 83 Linköping

Lund University: Department of Linguistics and Phonetics

The department of linguistics and phonetics in Lund has a broad research profile. One of the areas of specialization is the study of different aspects of speech- and language-processing. Within this area, production and perception studies of prosody as well as their speech technology applications (speech synthesis and speech recognition), constitute a central area of research. Language learning, as well as psycholinguistic aspects of the language acquisition process constitute another area where the department's researchers are very active. The reading process, as well as the writing process, comparisons between spoken and written language production as well as the cognitive processes that control them constitute another theme where research focus lies on the understanding of human language processing and its acquisition. Another profile area represented at the department is language typology and the description of less well known languages and minority languages. Particular focus lies on the description of languages and language families in Asia. Research dealing with the relationship between language, gestures and pictures constitutes another, cross-disciplinary research area that the department is currently involved in together with the departments of cognitive science and semiotics.

In 2006, the department become a part of the Language and Literature Center (http://www.sol.lu.se/) at Lund University. The humanities laboratory located in the new center makes it possible for the department of linguistics and phonetics to offer students and researchers a modern and well-equipped facility for pursuing interdisciplinary studies on language and speech. A number of digitalized databases are being developed there (see below); the laboratory is also a partner in the European initiative DAM-LR which proposes to develop and deploy an infrastructure that enables easy management of and access to linguistic resources of all kinds such as large (multimedia) corpora, lexicons, grammar description, etc.
The department of Linguistics and Phonetics is also a partner in the national Graduate School of Language Technology, a collaboration between leading centres in language technology in Sweden (http://www.gslt.hum.gu.se/). The school aims to integrate research on speech and language and to provide a sound basis in both theoretical foundations and applications oriented research.

Ph.D. dissertations in linguistics and phonetics since 2003:

- Frid, Johan. 2003. *Lexical and acoustic modelling of Swedish prosody*. Travaux de l'institut de linguistique de Lund 45.
- Hansson, Petra. 2003. *Prosodic phrasing in spontaneous Swedish*. Travaux de l'institut de linguistique de Lund 43.
- Mattson Anna Flyman. 2003. *Teaching, learning, and student output*. Travaux de l'institut de linguistique de Lund 42.
- Karlsson, Anastasia Mukhanova. 2005. *Rhythm and intonation in Halh Mongolian*. Travaux de l'institut de linguistique de Lund 46.
- Schötz, Susanne. 2006. *Perception, Analysis and synthesis of speaker age*. Travaux de l'institut de linguistique de Lund 47.
- Uppstad, Per Henning. 2005. *Language and literacy. some fundamental issues in*

research on reading and writing. Lund: Dept. of Linguistics and Phonetics.
- Zetterholm, Elisabeth. 2003. *Voice imitation. A phonetic study of perceptual illusions and acoustic success*. Travaux de l'institut de linguistique de Lund 44.

Ph.D. dissertations in progress:
- Ambrazaitis, Gilbert. Comparison of Swedish and German intonation.
- Andrén, Mats. Language evolution and gestures.
- Johansson, Victoria. The writing process.
- Roll, Mikael. Language and speech processing using ERP-studies.
- Sayehli, Susan. Second-language learning (German-Swedish).
- Segerup, My. Word accents and quantity in the Gothenburg dialect and other west-Swedish dialects
- Uneson, Marcus. Data-driven induction of phonological rules

Current research projects:
- Adverbiella verb i formosanska språk, RJ (Arthur Holmer)
- Att läsa medan man skriver VR (Åsa Wengelin)
- Kommunikationsflödet i texttelefonsamtal FAS (Åsa Wengelin)
- Att skilja intonation från ton RJ (Anastasia Karlsson)
- Den taktila läsprocessen (Sven Strömqvist)
- Distribuerad Access till Språkvetenskapliga Forskningsdata (Sven Strömqvist)
- Grammatik, Prosodi, Diskurs och Hjärnan. ERP-studier i Språkbearbetning VR (Merle Horne)
- Kammu-Engelskt lexikon (Jan-Olof Svantesson)
- Lingvistiska strukturers effekt på kognitionen VR (Sven Strömqvist)
- Språk, gester och bilder i ett semiotiskt utvecklingsperspektiv LU (Jordan Zlatev)
- Simulering av svenskans prosodiska dialekttyper VR (Gösta Bruce)
- Verbsyntax i Gulfarabiska dialekter VR (Maria Persson)

Frequently used databases in Lund:
- SweDia (Swedish Dialect spoken language corpus)
- SUC (Stockholm-Umeå Corpus)
- CHILDES (Child Data Exchange System) (http://childes.psy.cmu.edu/)
- The Kiel Corpus of Read/Spontaneous Speech
- CTHs uttalslexikon
- MBROLA Swedish diphone databases[1]

[1] Two Swedish MBROLA diphone databases have been developed at the department, and NLP components for text processing, pronunciation and prosody have been implemented using the Festival (www.festvox.org) framework. A demo is available (http://www.ling.lu.se/persons/JohanF/php/festival.php), but the current state of the system is more a test implementation than a practical product and many aspects of the system could be improved. Other synthesis-related demos are available as well: Letter-to-Sound rules for Swedish (http://www.ling.lu.se/persons/JohanF/php/ltsr.php) and an MBROLA interface (http://www.ling.lu.se/persons/JohanF/php/mbrola.php).

Language technology resources currently being developed in Lund:

- DAM-LR: Distributed Access Management of Language Resources (www.mpi.nl/dam-lr/lra-flyer)
- Swedish and Thai longitudinal child language corpora- approximately half a million running words each plus extensive video linkage
- Archive of Kammu Language and culture
- Recordings of reading and writing activity online (eyetracking, keystroke logging)
- Swedish (Festival based) Text-to-speech-conversion

VR Survey Regarding Language Technology Databases

Department of Speech, Music and Hearing, KTH, 2006-11-15

1. Language Technology Research

1.1. Research Topics and Plans

The Department of Speech, Music and Hearing has long experience in basic research in speech analysis, synthesis and understanding and in several language technology areas including multimodal spoken dialogue systems and communication aids for persons with disabilities. The work is characterized by a search for understanding spoken human-human interaction and how this knowledge can be used in human-machine interaction. Work on multimodal dialogue systems combines research in speech technology with linguistics, phonetics, cognitive science, psychology, and computer science. Below follows a very short description of our research topics and plans, divided into five areas all dependent on databases of various kinds. Data-driven methods have in practice proved to perform very well in our field of research.

Spoken dialogue systems. The research concerns the integration of speech technology in advanced interactive demonstrators and building multimodal conversational dialogue systems. The research area also involves development of new multimodal methods for second language acquisition utilizing speech technology. Obviously the development is dependent on data on Swedish spoken with and without accent.

Language models for spoken language, including dialogue models. This research deals with the creation of speech technology-motivated language and dialogue models for Swedish. It also includes the development and testing of data-driven solutions suitable for speech technology applications together with studies of robust linguistic analysis for spoken language, optimized for dialogue systems. This research is dependent on databases of linguistically annotated spontaneous dialog data.

Methods for automatic speech understanding. A major goal is the development of state-of-the-art automatic speech understanding for Swedish, e.g. to be used for speaker-independent recognition of large vocabularies. Another objective is the development of robust speech recognition that uses a relatively restricted vocabulary and is applicable in noisy environments. Databases covering spoken Swedish are mandatory for the development of speech recognition of Swedish rather than English.

Principles of speaker characterisation. This research direction includes the creation of models of speakers for use in systems for speaker verification, speech recognition with rapid speaker adaptation and individualized speech synthesis. It also includes the developments of methods for fast speaker adaptation. It is self evident that this research must be based on speech data from a large variety of individuals.

Speech production for multi-modal speech synthesis. This field concerns the development of articulatorily motivated, highly natural multi-modal parametric synthesis for different voices, speaking styles and emotional expressions. Included is a complete 3-D model of a face and speech organs that generate articulatory synthesis for use in animated speaking agents and moreover the modelling of non-articulatory facial gestures typical for interactive speech. This type of research depends heavily on specialised audio and video recordings enabling multimodal measurements.

1.2. Projects

Currently we are engaged in 6 EU projects, 8 national projects and one international, industrial co-operation. A listing of finished EU projects can be found in our 10 year report (http://www.speech.kth.se/ctt/publications/CTT_10_year_final_report.pdf).

Current projects are listed below together with the respective funding organisation. Most of these projects are carried out together with other academic research institutions and industrial groups.

MILLE, Modelling Interactive Language Learning, Riksbankens Jubileumsfond
ARTUR, Articulatory Tutor, VR
What makes conversation special?, VR
Separating intonation from tone, VR
Knowledge rich speaker adaptation for speech recognition, VR
Simulekt, Simulation of Swedish Prosodic Dialect Types, VR
ADEPT, Audiovisual Detection of Errors in Pronunciation Training, VR & SIDA
Dico, A Multimodal Menu-based In-vehicle Dialogue System, VINNOVA
CHIL, Computers in the Human Interaction Loop, EU
MonAmi, Mainststreaming on Ambient Intelligence, EU
HaH, Hearing at Home, EU
ASPI, Audiovisual to Articulatory Speech Inversion, EU
MUSCLE, Multimedia Understanding through Semantics, Computation and Learning, EU
COST 2102, Cross-Modal Analysis of Verbal and Non-verbal Communication, EU
Siemens Sound Classification, Siemens

1.3. Research Group

The Department of Speech, Music and Hearing, KTH, (http://www.speech.kth.se/) has been a prominent international centre of speech communication research for more than 30 years. The department is or has been engaged in numerous EU-projects. KTH is represented on the International Speech Communication Association (ISCA) board and the European Language and Speech Network (ELSNET) board. The department organized the International Congress on Phonetic Sciences (ICPhS) together with the Linguistic department at Stockholm University in1995, and the International congress Eurospeech in a joint Nordic effort in 2001. Both events attracted about 1000 participants. Moreover, it has organized several international workshops, including the ESCA workshop STiLL (Speech technology in language learning), the ISCA workshop “Error Handling in Spoken Dialogue Systems” and the European summer school MiLaSS, multimodality in language and speech systems.

CTT (Centre for Speech Technology) is associated to the department. It was created as a platform for co-operation between industry and academic research within the strategically important area of speech technology. In the 10 year report, see section 1.2, the 34 external co-operating partners that have been active during the period are presented. These include large Swedish industries such as ABB, TeliaSonera, SAAB, several SME’s and also the Swedish Radio, Swedish Television and the Swedish Handicap Institute. During the 10 year period more than 300 refereed contributions were published. Also 8 licentiates and 12 doctors were examined. CTT has been evaluated three times by international experts and the reports are available from VINNOVA.

As of November 15, 2006 the department has 2 Professors (Rolf Carlson and Björn Granström); 1 Associate Professor (David House); 11 PhD:s and 10 Graduate Students.

2. Language Technology Databases and Tools

2.1. Databases at TMH

First we list Swedish databases that we have been using during the last years. All of them except the last two have been collected at our department. The first three are part of large scale EU projects including many different European languages. They contain mostly read speech, but in the SpeeCon database we also recorded 10 spontaneous utterances per speaker and noted that annotation of these data required enormously more efforts and time than read speech or about 50 times the duration of the corresponding recorded speech signal. The costs for recording the SpeechDat database was around 2,5 MSEK while the cost for SpeeCon was around 1,5 MSEK. However, we estimate the cost for a creating 1000 hour spontaneous speech database to be around 25 MSEK, of which the orthographic annotation is the major part. This would give us comparable amounts of spontaneous speech to what exists for English and French.

Table 1. Swedish databases at the Department of Speech, Music and Hearing, KTH

Database	Year	Speakers	Purpose and type of recording	Usage	Stanardised
Speech, large scale					
SpeechDat, fixed telephone	1998	5000	Automatic speech recognition, ASR, for the fixed telephone network	frequent	International standard
SpeechDat, mobile telephone	1998	1000	ASR for mobile phones: in the office, pavement, vehicle or public place	rare	International standard
SpeeCon	2003	550 adults 50 children	ASR in the home, office, outdoors or car, 4 mics	frequent	International standard
Speech, project data					
Rafael	1994	1000	Speech recognition	rare	Nordic standard
Gandalf	1996	86	Speaker verification	frequent	
PER	2006	52	Speaker verification	frequent	
Waxholm	1992	68	Speech based dialogue system, audio	frequent	
August	1998	265	Dialogue system, audio and video	rare	
Adapt	2004	16 + 32	Dialogue system, speech and some video	rare	
PF-star	2003	198 children	ASR for children, audio, 2 mics	frequent	Partly
PF-star	2003	2	Multimodal Qualisys, video, audio	frequent	
Higgins	2006	16 + 16	Human dialogue system, speech and some video	frequent	
CHIL KTH Connector	2006	10 + 8 + 8	Dialogue system, speech	frequent	
Text and lexica		**Words**			
KTH-text	1990	150 M	Language models for Swedish	medium	
Onomastica		100 000	Name pronunciation Sweden	rare	
CENTLEX	2006	410 000	Pronunciation dictionary for Swedish	frequent	
National resources					
SWEDIA 2000	~2001	~1300	Recording of 110 different Swedish dialects	rare	
GSLC, Göteborg Spoken Language Corpus	From early 1980s	~1700, (all not different)	Transcriptions of Swedish in different social activities. Video and sound recordings.	rare	

Below we list databases for other languages, most of them English, we have acquired through membership in LDC, Linguistic Data Consortium, in USA and ELDA, Evaluations and Language resources Distribution Agency, in Paris. The reason we have got so many of these databases is that by paying membership for one year in order to get certain databases you get a lot of other databases for the same fee. Very few of these contain any Swedish material.

Table 2. Non-Swedish databases at the Department of Speech, Music and Hearing, KTH

Database	Language	Usage
Speech		
SpeechDat		
FDB614	Luxembourgish German, Luxembourgish French	
FDB1000	Flemish, French, Belgian French, Slovenian , Finland, German	rare
FDB2000	Welch	
FDB3000	Italian, Swiss French	rare
FDB4000	German, Swiss-German, Finnish, English, Spanish	rare
MDB1000	German, Dutch, English, Italian, Swiss French	
SDB1000	Swiss French	
Other Speech Databases		
TIMIT, NTIMIT, CTIMIT, HTIMIT, FFMTIM	American English, different conditions, phonetically labeled	frequent
Resource Management (RM1+RM2)	American English for speech recognition	
ICSI Meeting Speech	American English	
ICSI Meeting Transcripts	American English	
2002 NIST Speaker Recognition Evaluation	American English	rare
ISL Meeting Speech Part 1	American English	frequent
ISL Meeting Transcripts Part 1	American English	frequent
Switchboard Cellular Part 2 Audio	American English	
MDE RT-03 Training Data Speech	American English	
NIST Meeting Pilot Corpus Speech	American English	
Santa Barbara Corpus of Spoken American English III	American English	
2002 Rich Transcription Broadcast News and Conversational Telephone Speech	American English	
Fisher English Training Speech Part 1 Speech	American English	
Fisher English Training Speech Part 1, Transcripts	American English	
MDE RT-03 Training Data Text and Annotations	American English	
NIST Meeting Pilot Corpus Transcripts and Metadata	American English	
2000 Communicator Dialogue Act Tagged	American English	
2001 Communicator Dialogue Act Tagged	American English	
TIDIGITS	American English for speech recognition	
CSR, WSJ0, Training data and Nov-92 Test data	American English for speech recognition, training and test	
CSR-95, Radio Broadcast News, Nov-95 Hub 4	American English for speech recognition, training and test	
Boston University Radio Speech Corpus	American English	
SWITCHBOARD-1, Release 2	American English	
CALLHOME, American English	American English, German, Spanish, Japanese	
CALLFRIEND	15 different languages/dialects	
OGI Multi Language Telephone Speech Corpus	11 different languages	
Voice Across Hispanic America	Spanish	
SPIDRE	American English	
SWITCHBOARD Speaker ID Corpus	American English	
YOHO	American English	rare
SESP	Dutch	
SESP-II	Dutch	
HCRC Map Task Corpus	Brittish English	
HCRC Map Task Corpus, Sleep Deprivation Study	Brittich English	
Text and Lexica		
Spanish Language News Corpus	Spanish	
European Languages News Corpus	French, German, Portuguese	
Multilingual Corpus 1	29 languages and a total of 98 million words	
Onomastica	Name pronunciation lexica	rare
CELEX	Dictionaries for Dutch, English, German	

2.2. Tools at TMH

Over the years we have developed various tools for the handling speech research data. The WaveSurfer tool we use today is an Open Source tool for speech visualization and annotation. It has been designed to suit both novice and advanced users and can be used as a tool for a wide range of tasks in speech research and education. WaveSurfer can easily be extended to new applications by custom plug-ins or by embedding WaveSurfer visualization components in other applications. WaveSurfer may also be used for the annotation of multimodal signals, i.e. both sound and video signals may be annotated synchronously.

For the handling of our multi-purpose central lexicon database CENTLEX we have developed special tools. The lexicon is based on lexical resources of different types and formats. All information is stored in a relational database. CENTLEX is a full-form lexicon, with each entry minimally containing an orthographic word form and a grammatical analysis. An entry can have an arbitrary number of phonemic representations, ordered by their probability of use. An entry also contains information about the probability of a particular grammatical analysis. Presently, the database contains about 410,000 entries.

For general text-to-phoneme conversions we have a rule based system, RULSYS, that utilizes the CENTLEX transcriptions. It is used for speech synthesis and for deriving transcriptions to be used for training of automatic speech recognition systems.

We also use other public domain tools such as HTK, Hidden Markov Model Toolkit, for training and testing of speech recognition, and SRILM, The SRI Language Modelling Toolkit, for building and applying statistical language models.

3. Future Needs

It should be obvious from this report that speech research to a large extent is data-driven and thus needs substantial amounts of data to advance the speech science. If we want to continue internationally interesting research on Swedish there is consequently a need for more Swedish speech and video databases.

Recently there has emerged a common interest among Swedish researchers regarding the need for Swedish language databases. In an application to Vetenskapsrådet in the spring of 2006 we, together with other language research institutions, proposed a joint planning project *An infrastructure for Swedish language technology* (Research grant M and KFI 27 April 2006) concerning the creation of such databases. Another example of this engagement to promote the field among language researchers is the active feedback on the KFI web discussion on language databases concerning a roadmap proposal for Swedish infrastructure. In this context we can also refer to the governmental commission of inquiry “Mål i mun”, 2002, and the government bill “Bästa språket”, 2005, that both emphasize the need of Swedish speech databases. These documents have been discussed in a strategic letter to the Swedish government *Sverige behöver en strategi för språkteknologi*, see http://www.gslt.hum.gu.se/docs/strategiskrivelse.pdf, on the web page *Språkpolitiska dokument* at http://sprakteknologi.se/dokument.

It may be seen in our listing of databases that there exist an abundance of English speech databases that have been developed for various different research purposes, such as speech recognition of large vocabularies, in different noisy environments, of spontaneous speech and also for speaker verification and many more. The efforts needed to get a good linguistic insight into a language are of course not dependent of how many speakers the language has. Thus, one could claim that Swedish scientists need a corresponding amount of databases for their research. However, it is not realistic to collect comparable amounts of data for Swedish, especially when

considering the large English research community. Fortunately not all data are of the same importance. What we at our department see as vital is spontaneous speech data for our speech understanding research and video recorded spontaneous dialogues for our dialogue and multimodal speech synthesis research. Thus, we have recently applied for a grant from VR regarding *Multimodal database of spontaneous speech in dialog*. We have applied for around 1 MSEK per year in this 3 year proposal.

If we look at a future five year perspective we have described some of our needs in the previously mentioned planning project *An infrastructure for Swedish language technology*. There we propose to record 1000 hours of spontaneous speech, which mostly will be achieved by recording people in interaction, often in a dialogue context, but also in scenarios with more people, e.g. meetings. The recording will include different Swedish accents. Considering that around 10 % of the Swedish population is born outside Sweden and a further 10 % have at least one parent not born in Sweden there will also be a need for recordings of people that do not have a Swedish mother tongue. To perform the recordings and annotations of the recorded data we will need the definition of recording standards and development of recording and annotation software. These costs are on the order of 1 MEK. Regarding hardware we will need some extra storage facilities, computers and high-definition video cameras. These requirements are on the order of hundred kSEK. The total cost for a 1000 hour database is estimated to 25 MSEK, of which the most will be used for annotation as explained before. Concerning resources needed for the backup, distribution and support of our databases we estimate a need of 1 to 2 man-months per year. However, it must be pointed out that all facts regarding this database are very preliminary as they come from an application for a planning grant, which, if successful, will result in a complete and more detailed final application.

About the foreseen usage of the data it will be freely available and constitute a national resource for research and development. It will also be an extensive source of knowledge for phoneticians as well as linguists interested in various aspects of spoken Swedish. Speaking for ourselves we, of course, predict an intense use of the data. It will be used for improving speech recognition and multimodal speech synthesis as well as more user friendly spoken dialogue systems.

Regarding international collaboration we hope to continue that on the breadth we have today. Although the databases are language specific there are still very many aspects of our research that is not, e.g. algorithms for noise robust speech recognition, methods for speaker verification, methods for real-time speech understanding as well as improvements of spoken man-machine interfaces.

4. References

Mål i mun - Förslag till handlingsprogram för svenska språket, SOU 2002:27, April, 2002.

Prop. 2005/06:2, Bästa språket – en samlad svensk språkpolitik, September, 2005.

Språkteknologigruppen på Nada

Skolan för datavetenskap och kommunikation, KTH
http://www.csc.kth.se/tcs/humanlang/

Viggo Kann, viggo@nada.kth.se

1 Inriktning, sammansättning och planer

Gruppens forskningsinriktning kan sammanfattas: effektiva och resurssnåla metoder för språktekniksystem som hanterar text, i synnerhet på svenska.
Tre huvudområden:

- svensk språkgranskning och språkteknik inom utbildningsområdet
- informationssökning och -extraktion, särskilt klustring och sammanfattning
- ordboksbyggen

Vid vår utveckling av språktekniksystem har vi följande designmål:

- funktionellt och robust
- effektivt
- resurssnålt (i fråga om arbetstid, statistiska metoder föredras)
- evaluerbart, helst automatiskt
- fritt tillgängligt

De flesta av verktygen och tillämpningarna som vi har konstruerat finns tillgängliga för nerladdning från http://www.csc.kth.se/tcs/humanlang/tools.html Eftersom vi arbetar mycket med statistiska metoder och evaluering har vi stora behov av språkdatabaser.
Gruppens sammansättning 2006:

Viggo Kann, professor i datalogi
Ola Knutsson, fil.dr. i människa-datorinteraktion
Jonas Sjöbergh, tekn.dr. i datalogi
Martin Hassel, fil.lic. i datalogi
Magnus Rosell, tekn. lic. i datalogi
Ett antal exjobbare arbetar också i gruppen med sina projekt.

Gruppen har nära samarbete med bland andra

Kerstin Severinson-Eklundh, professor i MDI-gruppen
Hercules Dalianis, docent i data- och systemvetenskap, DSV KTH/SU
Tessy Cerratto, lektor i data- och systemvetenskap, DSV KTH/SU
Rolf Carlssons och Björn Granströms grupp vid KTH CSC

Två företag har hittills avknoppats från gruppen:

Euroling AB (http://www.euroling.se/)
Algoritmica Hollman och Kann HB (http://www.algoritmica.se/)

Nuvarande projekt:

- Nordisk nätordbok – Tvärslå (http://www.csc.kth.se/tcs/projects/netordbog/)
- Språkliga datorstöd och andraspråksinlärning (http://www.nada.kth.se/∼knutsson/call-en.html)
- Textsammanfattning – Martin Hassels doktorandprojekt (http://www.dsv.su.se/∼hercules/textsammanfattningeng.html)
- Textklustring – Magnus Rosells doktorandprojekt (http://www.nada.kth.se/∼rosell/research/forskning.html)

Planerade projekt:

- Utveckling av språkteknologiska program för utbildningsområdet (se nedan)
- Detektion och generering av humor (Jonas Sjöberghs tvååriga postdocprojekt i Japan som precis har börjat)
- Resurssnåla metoder för framtagande av språkteknologiska resurser för minoritetsspråk (sökt från VR 2006 men avslaget)

Språkteknologiska program för utbildningsområdet

Detta planerade projekt handlar inte enbart om språkutbildning utan också om olika utbildningsaktiviteter där språk spelar stor roll (vilket det gör i de flesta). Forskningen ska inriktas på en mångfacetterad bild av språk i linje med Tomasellos ”usage-based theory of language acquisition” och det som Lantolf & Thorne (2006) kallar en lingvistik för kommunikativ aktivitet. Inom detta fält är influensen av prefabricerat språk stor och viktig vilket bör avspeglas t.ex. i form av korpusbaserade idiomdatabaser och databaser med andra kommunikativa formler.

En utbildningsinriktning kräver också stora text- och talmaterial dels från andraspråksinlärare men också från modersmålstalare. Nivåer från grundskole- till högre universitetsutbildning krävs. Det är viktigt att material samlas in med hänsyn till uppgiftsbeskrivningar, elevbidrag och bakgrundsinformation, gärna med iterationer och inblandning av lärare och elever. Läroboksmaterial är också mycket önskvärt för framtagning av effektiva hjälpmedel och bakgrundstudier för teoriutveckling.

2 Våra forskningsdatabaser

2.1 Stavas ordbas

Vårt rättstavningsprogram Stavas ordbas är baserad på SAOL, upplaga 11, uppdelad i förled och efterled och kodad för användning i Stava. Den kodade versionen är fritt tillgänglig från http://www.csc.kth.se/tcs/humanlang/tools.html
Ordbasen byggdes upp i projektet Algoritmer för svenska språkverktyg 1993–1996 (http://www.csc.kth.se/tcs/projects/swedish.html)
Den har sedan dess förbättrats kontinuerligt av Viggo Kann. Flera manmånaders arbete ligger bakom ordbasen. Underhållskostnaden är låg. Stava används både självständigt och som en del av vår grammatikgranskare Granska som i sin tur är en del av skrivmiljön Grim. Varianter av Stava används som morfologianalysator och sammansättningsanalysator. Många människor har laddat ner Stava och dess ordbas.

2.2 Tagglexikon

Ett lexikon med POS-taggade ord och ordstatistik som används av Granskas ordklasstaggare och ordböjare samt vår sammansättningsuppdelare. Tagglexikonet är gjort för att kunna spridas fritt. Det utvecklades 2003 på cirka sex manmånader av Jonas Sjöbergh i projektet CrossCheck – svensk grammatikkontroll för andraspråksskribenter (http://www.csc.kth.se/tcs/projects/xcheck/) Lexikonet behöver inte underhållas. Både taggaren och sammansättningsuppdelaren används av flera andra forskare, och eftersom de ligger öppna för nedladdning troligen av många flera, se http://www.csc.kth.se/tcs/humanlang/tools.html

2.3 CrossCheck-korpusen – en elektronisk svensk inlärarkorpus

Uppbyggd 2001–2004 av i första hand Janne Lindberg och Gunnar Eriksson vid SU. Vi var inblandade i projektet och bidrog med delar av korpusen, bland annat morfosyntaktisk annotering. Har använts i CrossCheckprojektet för grammatikgranskning.
http://www.csc.kth.se/tcs/projects/xcheck/korpus.html

2.4 KTH News Corpus

Består av hundratusentals nyhetstexter nerladdade från tidningars webbsidor, totalt 13 miljoner ord. Skapades av Martin Hassel 2000–2001 (http://www.nada.kth.se/∼xmartin/papers/) och har används i många projekt i gruppen.

2.5 Nätordboksprojektets ordböcker

Flerspråkiga ordböcker på nordiska språk och engelska, insamlade och XML-kodade av nätordboksprojektet (bland annat ingår Lexinlexikonen). Sökbara i Tvärslå: http://ordbok.nada.kth.se/ XML-formatet (som huvudsakligen är en förenklad version av TEI dictionary base tag set) och

konverteringen till detta format har gjorts av Viggo Kann. Vissa av ordböckerna kommer att utvidgas inom projektet, som håller på minst hela 2007.

2.6 Folkets synonymlexikon

67 000 svenska synonympar betygsatta av internetanvändare 2005–2006. Synonymlexikonet förbättras ständigt. Underhållet består i att med jämna mellanrum föra in inskickade förbättringar. Lexikonet är helt fritt och kan laddas ner i XML från http://lexikon.nada.kth.se/synlex.html
Upphovsman och underhållsansvarig är Viggo Kann.

2.7 KTH Extract Corpus

Insamlad av Martin Hassel: en korpus med en större samling extraktbaserade sammanfattningar till ett mindre antal texter.

2.8 Mindre korpusar

- 99 ämnescentrala frågor och svar till lika många nyhetstexter, avsedd för Q&A-tester. Den har använts till textsammanfattning och informationssökning i vår grupp.
- Liten svensk trädbank skapad av Ola Knutsson och använd i våra parserutvärderingar.
- Svenska sammansättningar dels uppdelade i sina ordled och dels med ordleden POS-taggade. Även svensk löptext där sammansättningar delats upp i sina ordled. Användes för utveckling och utvärdering av sammansättningsuppdelaren, som i sin tur används av oss och många andra forskare.
- Två tvåspråkiga lexikon. Ett för svenska-japanska (som användes för att starta webblexikonet www.japanska.se) och ett för thailändska-svenska (som använts i utvärdering av andraspråksinlärares användande och behov av lexikon).

3 Framtida behov av forskningsdatabaser

3.1 En stor svensk råkorpus på en miljard ord

En stor råkorpus borde gå att åstadkomma i dagens läge när alla böcker och tidningar finns elektroniskt. Den kan användas till massor av spännande saker, som grammatikgranskning genom detektion av avvikelser från normalspråk, ords samspel i text, automatisk detektion av orsakssamband, träning av diverse unsupervisedmetoder m.m. Copyrightproblem lär dock finnas. För många tillämpningar räcker det med att få alla meningar i bokstavsordning eller liknande, vilket borde lösa sådant (men å andra sidan omöjliggör vissa saker också). Vi kan automattagga den på ordklass- och frasnivå.

3.2 Korpusar med skämt

Kommer att produceras i humorprojektet.

3.3 Lexikal semantisk korpus

En svensk korpus som tagits fram med fokus på lexikal semantik t.ex. med en taggning baserad på Fillmores FrameNet (användbar i humorprojektet och för informationsextraktion).

3.4 Felkorpus

Analyserad och rättad svensk felkorpus framtagen under så naturliga förhållanden som möjligt (datadriven granskning och utvärdering av stavnings- och grammatikgranskningsverktyg).

3.5 Utbildningskorpus

Stora mängder svenskt text- och talmaterial från olika utbildningsnivåer inom olika ämnen för att kunna göra diverse genreorienterade studier och språkteknologiskt baserade läromedel. Även lärobokstext. (För användning i projektet språkteknologi inom utbildningsområdet.)

3.6 Minoritetsspråkskorpus

Korpusar och lexikon för minoritetsspråk som sydsamiska och tornedalsfinska. Även parallellspråkskorpusar med dessa språk och svenska, finska eller engelska. (Samlas in i minoritetsspråksprojektet om vi får stöd för det.)

3.7 Svensk trädbank

Skulle användas i bland annat humorprojektet och för grammatikgranskning.

3.8 Sammansatta ord

En större korpus med sammansatta ord, helst med kontext, märkta med korrekt uppdelning (förbättrad sammansättningsanalys och ordklasstaggning). Vi har en mindre sådan korpus, se ovan.

MARTIN VOLK
Professor of Computational Linguistics
Inst för lingvistik
AVD FÖR DATORLINGVISTIK

2006-11-13

Bilaga 4h

Kartläggning av databasresurser inom språkteknologi – läget idag och framtida behov

Hej Eva

Thank you for your invitation to participate in this survey.

I have listed the databases that we are using today and our plans for the near future.

If you need any further information, please let me know.

Best regards, Martin

Postadress:
Inst för lingvistik
Stockholms universitet
106 91 Stockholm

Besöksadress:
Universitetsvägen 10C
Frescati

Telefon: +46 8 16 2334
Fax: +46 8 15 53 89

E-post
volk@ling.su.se

MARTIN VOLK
Professor of Computational Linguistics
Inst för lingvistik
AVD FÖR DATORLINGVISTIK

2006-11-13

1) Den språkteknologiska forskning som sker/planeras inom den egna institutionen / enheten och som kräver tillgång till stora databaser och databasverktyg. Ange:

1.1 Inriktning/nuvarande projekt

Our group participates in an industry-sponsored project to build a **machine translation system**. The system translates film subtitles from Swedish to Danish. Our industry partner is a major company in film subtitling (both TV and DVD productions) and has provided us with large amounts of already translated film subtitles (**several million subtitles in each language**). The translation system re-uses and re-assembles previous translations at various levels of granularity. Automatic translation of subtitles is promising because of the limited length of the sentences.

A first prototype has been built and produces good results. The output will be checked by a professional translator, but it is expected that a considerable percentage of the automatically translated subtitles need not be touched. The final version of the system shall be operational by the end of 2007.

All subtitles have been stored in a relational database with pointers to their translation, date, source and other administrative information. We also added linguistic information about named entities and word classes to improve the precision of the re-use. The database provides also for constant updates as new translations are coming in.

1.2 Forskningsplaner

We plan to investigate large amounts of parallel texts in multiple linguistic dimensions such as syntax, semantics, and across languages. We want to explore an innovative and promising direction of corpus annotation, **massive parallel text analysis**. Large amounts of translated texts have become available in recent years. For instance, the documents from the European Parliament (more than 20 million words per language) translated into 12 languages (termed "Europarl corpus") have triggered a lot of innovative research not only in statistical machine translation but also in word sense disambiguation and even automatic grammar induction. The parallel texts in such translation corpora are an important source of information for machine translation and for many other language technology applications.

Parallel texts can be seen as annotated texts. The (human) translation is a special type of annotation of the original. Often, syntactic or semantic ambiguities in the original must be decided in the translation and thus provide the basis for the automatic resolution of ambiguities and for other types of inferences. Consider the ambiguous sentence "She had known more attractive men". If the German translation is "Sie hatte attraktivere Männer

Postadress:
Inst för lingvistik
Stockholms universitet
106 91 Stockholm

Besöksadress:
Universitetsvägen 10C
Frescati

Telefon: +46 8 16 2334
Fax: +46 8 15 53 89

E-post
volk@ling.su.se

gekannt", then the computer may safely conclude that "more" modifies "attractive" (rather than "men").

Most initiatives in the past have exploited the parallelism only for one language pair or stayed within one language family. But there are now research results that show the advantages of processing more than two languages in parallel. We plan to use three Germanic languages (Swedish, English and German), plus at least one Romance language (French), and one non-Indo-European language (Finnish or Estonian) in order to evaluate whether clues from close or far-away languages are more or less helpful in automatic disambiguation. This selection covers some of the major languages of the EU but it avoids coding problems since all the above mentioned languages have writing systems based on the Latin alphabet.

As the starting point for this project we have annotated a parallel treebank in English, German and Swedish called SMULTRON (Stockholm MULtilingual Parallel TReebank). The name treebank is derived from the fact that syntax structures are mostly encoded as tree graphs. The trees in our treebank are aligned on the phrase level across languages.

1.3 Gruppens storlek och sammansättning

The Stockholm University group in Computational Linguistics currently consists of

1. Prof. Martin Volk
2. Sofia Gustafson-Capková (Lecturer)
3. Hans Hjelm (PhD student)
4. Kristina Nilsson (PhD student)
5. Henrik Oxhammar (PhD student)
6. Yvonne Samuelsson (PhD student)
7. Joakim Lundborg (Technician, Programmer)
8. Sören Harder (Researcher in the Subtitle Translation Project)
9. Jörgen Aasa (Researcher in the Subtitle Translation Project)

The Stockholm University group in Computational Linguistics is an active member of GSLT, The Swedish National Graduate School of Language Technology. Both Martin Volk and Sofia Gustafson-Capková participate regularly in teaching GSLT courses. And three of the Stockholm PhD students receive financing through GSLT.

Postadress:
Inst för lingvistik
Stockholms universitet
106 91 Stockholm

Besöksadress:
Universitetsvägen 10C
Frescati

Telefon: +46 8 16 2334
Fax: +46 8 15 53 89

E-post
volk@ling.su.se

MARTIN VOLK
Professor of Computational Linguistics
Inst för lingvistik
AVD FÖR DATORLINGVISTIK

2006-11-13

2) De forskningsdatabaser (inklusive verktyg) som finns vid institutionen/enheten. Ange där så är relevant och möjligt:

This table describes 3 databases that were (partially) created at our department in detail. It is followed by a list with a number of other databases which we are also using for our research.

	Database of Film Subtitles in Swedish and Danish	**SUC – The Stockholm Umeå Corpus (A balanced Swedish corpus)**	**SMULTRON – The Stockholm Multilingual Treebank (DE, EN, SV)**
- Storlek	> 10 million entries, constantly growing	1 million words with manually checked word class, morphology and base form information	~ 50.000 words with manually checked word and grammar information
- Ändamål som databaserna skapades för	Commercial – for building a machine translation system	Linguistic research and also Training and Evaluating Language Technology systems	Linguistic research and also Training and Evaluating Language Technology systems
- Nyttjandefrekvensen inom den egna institutionen/enheten samt bland andra forskare i Sverige och/eller andra länder	Project-specific	Widely used in Language Technology Research mostly in Sweden, more than 100 licenses signed (but also used abroad, e.g. in Switzerland and Germany)	Distribution starting soon
- Samarbete med andra forskare i Sverige och/eller andra länder kring uppbyggnaden av databaserna	Project-specific, cooperation with international researchers of the industrial partner	Regarded by many as one of the most important annotated Swedish language corpora	Development coordinated with researchers from Czech Republic, France, Germany, and Ireland
- Samordning med en eller flera andra nationella och/eller internationella databaser	-	Available in the widely-used TIGER-XML standard	Follows the widely-used TIGER-XML standard

Postadress:
Inst för lingvistik
Stockholms universitet
106 91 Stockholm

Besöksadress:
Universitetsvägen 10C
Frescati

Telefon: +46 8 16 2334
Fax: +46 8 15 53 89

E-post
volk@ling.su.se

- Samordning ifråga om anpassning av dokumentation och standarder till liknande nationella och/eller internationella databaser	Project Documentation, use of XML standards for data interchange	Corpus Documentation and Annotation Guidelines	Corpus Documentation, Alignment and Annotation Guidelines
- Kostnader för databasuppbyggnad och kostnader för drift, underhåll och support (personal, hård- och mjukvara)	Several million SEK	Several million SEK in various projects in the 1990s	~ 500'000 SEK
- Finansiering av databasuppbyggnad samt drift	Regular maintenance through system administrator (20% job)	-	Distribution and Maintenance and Enlargements (20% job)
- Ev annan utrustning som behövs som komplement i verksamheten som rör databaser	The processing of the data with the purpose of training statistical machine translation systems requires fast servers with large amounts of RAM	-	-

Postadress:
Inst för lingvistik
Stockholms universitet
106 91 Stockholm

Besöksadress:
Universitetsvägen 10C
Frescati

Telefon: +46 8 16 2334
Fax: +46 8 15 53 89

E-post
volk@ling.su.se

MARTIN VOLK
Professor of Computational Linguistics
Inst för lingvistik
AVD FÖR DATORLINGVISTIK

In addition, we work with large text collections (many of which are annotated with various types of linguistic information). For example we use

- Newspaper corpora.
  - 1 year of the Swiss daily Newspaper TagesAnzeiger (> 100 million words)
  - 5 years of the German weekly Computer Science newspaper ComputerZeitung (~ 7.5 million words)
- Europarl (written accords of the European parliament)
  - About 25 million words each in 12 languages (we use DE, EN, SV)
- Acquis Communautaire (the European legislative texts)
  - About 9 million words each in 20 languages (we use DE, EN, EE, FR and SV)
- The Swedish Parole corpus
  - About 19 million words annotated with morphological information, word classes, and shallow syntax structures
- Stockholm EkonomiKorpus (SEK)
  - The corpus consists of 2.800 Swedish economy documents. Part of the corpus (365 documents with ca. 122.000 words have been annotated with word class tags and name class tags. 66 of these documents (ca 22.000 words) furthermore have been annotated with co-reference relations between noun phrases. This corpus is used in a PhD project for the evaluation of automatic co-reference resolution methods.
- Bonnier Product Classification Corpus
  - An EU ontology with 8.300 product classes called the Common Procurement Vocabulary (we use Swedish, English and German)
  - More than 10.000 Crawled company and product documents annotated with codes referring to the product classification ontology (provided by Bonnier Affärsinformation AB). This corpus is used in a PhD project on automatically classifying product descriptions.
- Treebanks (syntactically annotated texts; which means that each word is a data unit with additional linguistic information attached to it)
  - English
    - Penn-Treebank (Wall Street Journal) ~ 1.2 million words
  - French
    - LeMonde Treebank ~ 630.000 words
  - German
    - NEGRA-Treebank (Frankfurter Allgemeine Zeitung) ~ 180.000 words
    - TIGER-Treebank (Frankfurter Allgemeine Zeitung) ~ 900.000 words
  - Swedish
    - Talbanken (written and spoken Swedish) ~ 330.000 words

Postadress:
Inst för lingvistik
Stockholms universitet
106 91 Stockholm

Besöksadress:
Universitetsvägen 10C
Frescati

Telefon: +46 8 16 2334
Fax: +46 8 15 53 89

E-post
volk@ling.su.se

MARTIN VOLK
Professor of Computational Linguistics
Inst för lingvistik
AVD FÖR DATORLINGVISTIK

2006-11-13

3) Det framtida behovet av resurser för planering, utveckling, uppbyggnad, drift och avveckling av databaser samt ev annan utrustning som behövs som komplement i verksamheten som rör databaser. Uppgifter som bör anges är:

Here I summarize a project proposal that I made earlier this year to Vetenskapsrådet. It sketches the kind of research we plan to do in the near future.

	Massive parallel text analysis (The text cube project)
Ändamål	Our goal is to develop a methodology for the large scale annotation and interpretation of parallel texts (DE, EN, SV + FR and EE) which is both fast and accurate. Such a methodology will lead to valuable resources for Language Technology, General Linguistics and Translation Studies as well as to commercially viable language technology applications.
Förväntad nyttjandefrekvens samt samarbete / samordning nationellt och internationellt av verksamheten	We are in close contact with research groups in Estonia (Tartu), Germany (Osnabrück, Saarbrücken), and Switzerland (Zurich) which work on language analysis tools (e.g. PoS taggers, chunkers and word aligners) and provide annotation tools (e.g. the SALSA tool for frame semantic annotation) for the chosen languages. In exchange they are using our TreeAligner and our Parallel Treebank SMULTRON for their activities. We will make both the collected language data and the annotation results available to the research community. Since we will be annotating data in 5 languages (including large languages like DE, EN and FR), our added-value will be of interest to a large number of researchers and system developers. If we extrapolate from the user frequency of SUC (see above), we can safely assume several 100 research and development groups interested in these resources.
Beräknade kostnader i ett 5 årsperspektiv	About 1.5 million SEK **per year** for data selection, preparation, annotation, plus database structuring and accessibility (these costs were estimated for a 2006 project proposal to VR)

Postadress:
Inst för lingvistik
Stockholms universitet
106 91 Stockholm

Besöksadress:
Universitetsvägen 10C
Frescati

Telefon: +46 8 16 2334
Fax: +46 8 15 53 89

E-post
volk@ling.su.se

Institutionen för lingvistik, Stockholms universitet

Tre svar har inkommit från institutionen.

I svar från Avdelningen för fonetik nämns dialektdatabasen SWEDIA, utvecklad i samarbete mellan fonetikavdelningarna i Stockholm, Lund och Umeå samt IRIS (Invandrarröster i Sverige) som innehåller inspelningar av invandrares tal. Vidare finns digitala databaser (audio och video) med joller och tidigt tal (upp till 30 månader) samt av patologiskt tal (efter glossektomi).

De övriga svaren, ett från Avdelningen för teckenspråk om en interaktiv teckenspråksdatabas och ett från Avdelningen för allmän språkvetenskap om två inlärardatabaser bifogas nedan.

Nedanstående text är ett utdrag ur Ahlgren, Inger & Bergman, Brita: Det svenska teckenspråket. SOU 2006:29 (s. 11-70). *Teckenspråk och teckenspråkiga. Kunskaps- och forskningsöversikt.* Delbetänkande av utredningen Översyn över teckenspråkets ställning.

”*1.3 Dokumentation av det svenska teckenspråket*

[…] Utmärkande för teckenböckerna från 60-talet och framåt är att urvalet av tecken i stor utsträckning bestämts av hörande inlärares behov. Man har utgått från listor med svenska ord och registrerat tecken med motsvarande betydelse. Tecknen är ordnade i ordens alfabetiska ordning, dvs. tecken som råkar översättas med ord som börjar på bokstaven ”A” hamnar först. Den enda sökvägen är alltså på de svenska orden.

”Svenskt teckenspråkslexikon” utgivet av Sveriges Dövas Riksförbund (1997) representerar ett helt annat arbetssätt i och med att man här utgått inte från ord, utan från tecken, som först dokumenterats med hjälp av transkription och fotografier och sedan översatts till svenska. Bokens uppläggning skiljer sig från de tidigare böckernas genom att tecknen här är ordnade i s.k. handformsklasser. (En tidig föregångare med liknande uppläggning är ”*Dictionary of American Sign Language*”, Stokoe et al. 1965.) Det innebär att tecken med samma handform beskrivs i samma avsnitt och ordnas sedan inom varje klass enligt läge och rörelseart. Givet att man vet ett teckens form kan man alltså slå upp det för att få veta vad det betyder.

”Digital version av Svenskt teckenspråkslexikon” (2001) är en vidareutveckling av ”Svenskt teckenspråkslexikon” [1997] i form av en interaktiv databas som innehåller det tryckta lexikonets information kompletterad med videosekvenser. Där visas tecknet dels i isolering, dels i en teckenspråksmening som belyser dess användning. Den digitala versionen av ”Svenskt teckenspråkslexikon” är alltså ett datorbaserat lexikon med rörliga bilder och med ett antal olika sökvägar såsom via svenskt ord, tecknets form, dess nummer i bokversionen och ämnesområde. Lexikonet, som framställts vid Avdelningen

för teckenspråk på Institutionen för lingvistik, Stockholms universitet, finns tillgängligt på internet via avdelningens hemsida (www.ling.su.se/tsp).

”Svenskt teckenspråkslexikon” omfattar 2968 uppslagstecken. Om detta säger SDR:

Dock upptar det endast en liten del av hela teckenförrådet, varför det endast är ett första steg i arbetet. Mycket forskning och arbete återstår till det stora målet: ett fullständigt lexikon över det svenska teckenspråket” (ibid. sid. vii).

Arbetet med fortsatt dokumentation av teckenförrådet, som bedrivs dels av SDR, dels av en fast redaktion vid Institutionen för lingvistik, SU, har hittills resulterat i ett antal speciallexika: ”Tecken från området idrott” (2003), som föreligger både i tryckt form och på CD-rom med 1056 uppslagstecken, ”Tecken inom området bridge” (2005) med 370 tecken, ”Tecken för matematiska begrepp” (2004) med 469 tecken. Under utgivning är ”Tecken inom området religion”, ”Tecken för begrepp inom språkvetenskap” och en samling med ca 500 geografiska egennamn. Därutöver har ca 3150 persontecken, ca 1000

vardagliga tecken, ca 1600 ålderdomliga och regionala tecken dokumenterats i form av transkriptioner, fotografier och videoinspelningar, men är ännu inte utgivna (Tomas Hedberg, pers. komm. 2005).

Sammantaget är idag ca 11 500 tecken dokumenterade i någon form. Hur stor del av det svenska teckenförrådet dessa utgör vet vi inte. Säkert är att många specialområden, såväl akademiska ämnen som mer vardagliga, inte är inventerade. Hur stor del av det mer centrala teckenförrådet som återstår att dokumentera är svårare att bedöma. Våra observationer tyder dock på att endast en liten del av detta dokumenterats, och att ett omfattade arbete med inventering och beskrivning återstår.

En svaghet med samtliga teckenspråkslexika är den ofullständiga betydelse-beskrivningen, som förlitar sig på svenska ord och meningar. Det digitala lexikonet har visserligen användningsexempel på teckenspråk, men ett enda exempel räcker i många fall inte för att belysa tecknets betydelse och användningsområde. Dessutom är exempelmeningarna konstruerade och sådana redaktionella, icke-autentiska, exempel kan vara mindre tillförlitliga. Detta problem kan endast avhjälpas med tillgång till ett omfattande datoriserat teckenspråksmaterial med olika typer av teckenspråkstexter, alltså en teckenspråkskorpus. En sådan korpus finns inte idag men är nu tekniskt fullt möjligt att bygga upp.

En annan brist hos teckenbeskrivningen är avsaknaden av grammatisk information, så som ordklasstillhörighet och morfologisk information. Även här skulle en teckenspråks-korpus möjliggöra den fördjupade analys som krävs för att få fram sådan ny kunskap.”

(SOU 2006:29, s. 15-18)

Databaser Allmän språkvetenskap, Stockholms universitet
BH 2001-12-01

Här nedan följer uppgifter om två språkkorpusar som har byggts upp vid Institutionen för lingvistik, Avdelningen för allmän språkvetenskap, Stockholms universitet. Båda har skapats för undersökningar av språket hos vuxna inlärare av svenska, men har olika uppläggning och forskningssyften.

ASU-korpusen
(Benämnd efter forskningsprojektet ”Andraspråkets strukturutveckling, ASU”)

Textkorpusen ASU består av ljudinspelade och transkriberade samtal och skrivna uppsatser på svenska producerade av unga vuxna inlärare, och därtill ett jämförbart språkmaterial insamlat från infödda svenskar. Korpusen omfattar totalt ca 490 000 löpord, varav ca 415 000 ord muntligt och ca 75 000 ord skriftligt material. Den är speciellt designad för longitudinella undersökningar av andraspråksutveckling och jämförelser av inlärares och inföddas språkproduktion. Den dokumenterar språket hos individuella inlärare med bestämda tidsintervall, så att man kan spåra och jämföra stadier i utvecklingen inom och mellan individer. Den ska också kunna tjäna som källa för att observera aspekter av tillägnandeprocessen i andraspråket. Tal och skrift har dokumenterats parallellt hos samma personer. Materialet från de infödda personerna utgör i sig en liten korpus av standardsvensk tal- och skriftproduktion.

Korpusen har en potential för forskning på flera områden av språket. I första hand är den tänkt för grammatiska och lexikala undersökningar, men den har också gett underlag för fonologi och samtalsanalys.

Materialet insamlades, transkriberades, taggades och redigerades åren 1990-93 och 1998 inom projektet *”Andraspråkets strukturutveckling (ASU)”* vid Institutionen för lingvistik, Stockholms universitet, under ledning av Björn Hammarberg och med finansiellt stöd från Humanistisk-Samhällsvetenskapliga Forskningsrådet och Stockholms universitets humanistiska fakultet. Korpusen lagrades från början i ASCII-format och redigerades och bearbetades med hjälp av korpusprogramvaran *PC Beta* i samarbete med prof. Benny Brodda. Denna korpusversion har legat till grund för artiklar och avhandlingar under 1990-talet och de första åren av 2000-talet.

Korpusen har senare genomgått en grundlig teknisk modernisering för att överensstämma med vad som idag vinner giltighet som standard för språkliga textkorpusar, och för att göra den användbar för en större krets av forskare och studenter. Detta har skett i samarbete med prof. Lars Borin, Språkbanken, Göteborgs universitet. Stöd för detta har erhållits från Magn. Bergvalls Stiftelse och Birgit & Gad Rausings Stiftelse för Humanistisk Forskning, samt genom samarbetet med Språkbanken. Den tekniska moderniseringen av korpusen har inneburit att hela korpusen har konverterats till ett XML-baserat lagringsformat och försetts med ett Java-baserat sök- och analysverktyg som har utarbetats i anslutning till projektet ”IT-baserat kollaborativt lärande i grammatik (ITG)”. Verktyget är speciellt utformat för att man ska kunna bearbeta framsökta data vidare i en interaktiv arbetsgång.

Korpusen kommer att göras tillgänglig på webben via Språkbanken (http://spraakbanken.gu.se/). Fortlöpande underhåll och vidareutveckling av verktygets funktioner sker också där.

Med den nya versionen förväntar vi ett ökat utnyttjande av ASU-korpusen både inom och utanför vår institution, vilket väntas aktualisera önskemål om utvecklingar. Det är dock ännu för tidigt att nu bedöma kostnader eller skissa på projektplaner.

Nu närmast ska en utförlig innehållsbeskrivning och en bruksanvisning kopplas till analysverktyget.

Ett projekt vi har övervägt men ännu inte satt i verket är att digitalisera ljudinspelningarna av korpusens taldel (ca 70 timmar) och samordna dem med textutskriften, vilket skulle möjliggöra fonetisk forskning på materialet.

SSM-korpusen
(Benämnd efter forskningsprojektet ”Svenska som målspråk, SSM”)

SSM-korpusen är den elektroniska versionen av en samling uppsatstexter producerade av studenter i preparandkursen i svenska för utländska studerande vid Stockholms universitet.
Korpusen omfattar i dagsläget ca 112 000 löpord, men texterna skall ses över och sovras något, i syfte att uppnå bättre proportioner i sammansättningen. Texterna är skrivna av personer med tio olika modersmål, vilket delar in korpusen i tio ungefär lika stora delar. Olika kursstadier är representerade inom varje del. Matrialet har tidigare givit underlag för språktypologiskt orienterade studier av problem i svensk syntax i ett andraspråksperspektiv.

Texterna insamlades under åren 1973-75 inom projektet *”Svenska som målspråk (SSM)”* under ledning av Björn Hammarberg och med finansiering från Skolöverstyrelsen. Senare, med avslutning 2005, har texterna skannats, redigerats och lagrats i ett XML-format i anslutning till ”CrossCheck”-projektet, Nada, KTH, samt försetts med ett sökgränssnitt utarbetat vid Språkbanken, Göteborg, där korpusen är åtkomlig. Den kan också nås från ITG-gränssnittet (se ovan under ASU-korpusen).

Det som närmast behöver göras med SSM-korpusen är den ovannämnda sovringen av texturvalet, samt att utarbeta en ordentlig dokumentation av korpusens innehåll och sammansättning.

Kartläggning av databasresurser inom språkteknologi – läget idag och framtida behov

Umeå University

1. Since Professor Eva Ejerhed resigned her position at Umeå University, there has been, if we apply a narrow definition of Speech and Language Technology, no Speech and Language Technology research conducted at Umeå University. However, there has been, and currently there is and will be in the future, a wide range of research conducted at Umeå University that may have direct, or indirect, relevance for Speech and Language Technology research and development. The database outcomes of our speech and language research are outlined in the answer to Question 2.

2. Five databases, four held by the Department of Philosophy and Linguistics, one in conjunction with the Faculty for Teacher Education, and one by the Department of Modern Languages. One of these is not-yet-digitized (The Archive of Accented Swedish). In addition to these databases we hold several collections of analogue material, for example, the corpus outcome of the project *Stadsmål i Övre Norrland*.

    a. VaKoS, Variation in Consonant Clusters in Swedish. A partially online database. The controlled labeled material is accessible by all researchers. The spontaneous material is yet to be labeled and published. This database was created within a project of the same name.

    b. Swedia 2000, The Phonetics and Phonology of Swedish Dialects around the year 2000. A central goal for this project, was to develop a digital database for research on Swedish dialects. The database was created on a national basis, in cooperation between the phonetic departments at Umeå, Stockholm and Lund universities and intended as a resource for research both within and outside the project. The corpus includes speech samples from 12 speakers (including younger and older men and women) from 107 locations across the Swedish-speaking areas of Sweden and Finland, altogether more than 1200 hours of collected speech. The database material of which a reasonable part has been annotated, has been used by researchers from several different areas.

    c. UC^3 Corpus, Umeå Child Consonant Cluster Corpus. A corpus of 22 children's productions of a set of control words. Each child was followed over a period of twelve months. This corpus is digital, yet not generally accessible. Only certain aspects are marked-up. This database was created as part of Fredrik Karlsson's PhD research; the findings of which are published in Karlsson (2006) *The Acquisition of Contrast: A longitudinal investigation of Initial s+plosive cluster development in Swedish Children.*

d. UDID, Umeå Disguise and Imitation Database. This is a digital and partially marked-up database contain voice imitations and attempts at voice disguise by 20 speakers. Work based on the database has been conducted together with Robert Rodman of North Carolina State University. This database is being constructed within Erik Eriksson's PhD project that is a part of the external project Imitated Voices awarded to Umeå University.

e. The Archive of Accented Swedish, the analogue database on which Robert Bannert's (1990) book *På väg mot svenskt uttal* is based. This database contains a rich set of data on accented Swedish.

f. Umeå Keystroke logged writing corpus. This is a digital collection of young writers' keystroke log files of English as a Foreign Language and Swedish writings that are yet to be indexed into a database. The collection is held jointly with the Faculty of Teacher Education, Umeå University.

g. UKC, Umeå Krio Corpus. An on-line (http://creole.mos.umu.se:8080) Krio corpus that is a unique resource and as such a core database resource for any Krio speech and language research. RJ supported the creation of this database.

3. Based on the review of the digital and the not-yet-digitized data that are available in Umeå University of relevance for Speech and Language research presented in 2 (above), there are several tasks that would enhance the usability of these databases including digitization, publishing on the internet, indexing, marking-up and documenting.

   There is a clear value in the development of databases for speech and language research in its broadest sense.

References

Bannert, R., (1990). *På väg mot svenskt uttal.* Lund: Studentlitteratur

Karlsson, F., (2006). *The Acquisition of Contrast: A longitudinal investigation of Initial s+plosive cluster development in Swedish Children*. (Umeå Studies in Linguistics No. 3). Umeå, Sweden: Umeå University, Department of Philosophy and Linguistics.

Anna Sågvall Hein
Professor i datorlingvistik
Dekanus för Språkvetenskapliga fakulteten
Uppsala universitet

2006-11-15

Tillgång till och behov av språkdatabaser vid Uppsala universitet samt några behov av nationellt intresse

Föreliggande redogörelse avser infrastruktur i form av språkdatabaser för såväl språkteknologisk som språkvetenskaplig forskning. Språkdatabaser utgörs vanligen av textsamlingar, s.k. korpusar, eller lexikala databaser, lexikon. Vidare upptas behov av språkdatabaser för svenska - med bearbetningsverktyg -, som bedöms vara av generellt intresse för hela den svenska språkteknologiska och språkvetenskapliga forskningsgemenskapen.

Språkteknologisk forskning i Uppsala bedrivs av den datorlingvistiska forskningsgruppen vid Institutionen för lingvistik och filologi. Språkvetenskaplig forskning som använder sig av språkdatabaser bedrivs såväl inom övriga delar av samma institution som inom Engelska institutionen, Institutionen för moderna språk och Institutionen för nordiska språk.

Ett nytt initiativ är Engelska parkens informationsteknologiska centrum för forskning och utbildning i humaniora, språk och teologi, EPARIT, inom vars ram såväl språkteknologisk som språkvetenskaplig, historisk-filosofisk och teologisk forskning samlas. Tanken är att skapa en gemensam infrastruktur för humanistisk och teologisk forskning och utbildning. Centret, som ännu befinner sig i sin linda, har tillkommit genom en strategisk satsning från Områdesnämnden för humaniora och samhällsvetenskap. En gemensam nämnare för den forskning som bedrivs inom EPARIT är användning av stora språkdatabaser och humanistiska databaser av andra slag för historisk-filosofisk och teologisk forskning.

1. Språkteknologisk forskning vid Institutionen för lingvistik och filologi

All seriös språkteknologisk forskning kräver tillgång till stora databaser och datoriserade verktyg för att bearbeta dem t.ex. analysera och märka upp dem med lingvistisk information för att göra dem effektivt sökbara. Det kan röra sig om allt från texter om c:a 50.000 ord för frågor om textattribution till miljontals ord för konstruktion av statistiska översättningssystem.

Inriktning

Forskningen är i första hand inriktad på det skrivna språket. Ur ett tillämpningsperspektiv har den sin tyngdpunkt på maskinöversättning, grammatikkontroll, textprediktion, automatisk summering, textkategorisering, textgradering, informationsutvinning, och automatisk summering. Forskning om datorstödd språkinlärning har också inletts.

Metodiskt inrymmer forskningen både regelbaserade och datadrivna ansatser, samt kombinationer av dem. Maskininlärning är en viktig del. Ett annat centralt forskningsområde är korpuslingvistik med fokus på uppbyggnad, analys och automatisk annotering av korpusar, enspråkiga såväl som flerspråkiga; här spelar maskininlärningsmetoder en stor roll. Korpusarna har flerfaldiga användningar, t.ex. som bas för utveckling av statistisk maskinöversättning, som bas för extraktion av översättningsekvivalenter för maskinöversättning och manuell översättning, som bas för extraktion av grammatisk kunskap om språk, som bas för forskning om informationssökning, informationsutvinning och automatisk summering, för flerspråkig lexikografi, som resurs för kontrastivt inriktad språkforskning och för undervisning i främmande språk.

Doktorandprojekt

Pågående doktorandprojekt rör maskinöversättning (3), evaluering av läsbarhet på mobila apparater, textsummering baserad på retorisk struktur och statistisk ordkategorisering. En nyantagen doktorand kommer att forska på psykolingvistiska frågor, t.ex. minnesbegränsningar och syntaktisk analys.

Pågående projekt

- Maskinöversättning av akademiska kursplaner från svenska till engelska /UU
- Utprovning av regelbaserad maskinöversättning från svenska till engelska/SÄPO
- Utprovning av regelbaserad och statistisk maskinöversättning för teknisk dokumentation/Scania CV AB
- Pilotstudie om maskinöversättning svenska-turkiska/VR:s småspråkssatsning
- Korpusuppbyggnad för svenska-turkiska, svenska-hindi/VR:s småspråkssatsning
- Automatisk grammatikextraktion/VR
- Automatsk textgradering av nationella prov i samarbete med Utbildningsvetenskapliga fakulteten och Umeå universitet/VR

Avslutade projekt

- A Swedish Systran module/EU, SYSTRAN, VINNOVA
- SCARRIE - Scandinavian proof-reading tools /EU
- FASTY – Fast typing for disabled people/EU
- Scania Checker – en språkgranskare för tekniska skribenter/Scania
- Scania Swedish – ett kontrollerat språk för teknisk dokumentation/Scania
- PLUG – Parallel corpora in Linköping, Uppsala, Göteborg/Svenska språkteknologiprogrammet (VINNOVA och VR)
- KOMA – Korpusbaserad maskinöversättning/Svenska språkteknologiprogrammet (VINNOVA och VR)

- MATS – en svensk maskinöversättningsmodul/VINNOVA
- MIA - Multra i arbete/ HSFR

Forskningsplaner

I första hand avser vi utvidga forskningen om maskinöversättning till att omfatta de orientaliska språk som ingår i institutionens utbildnings- och forskningsutbud, dvs. indiska språk (hindi, tamil), iranska språk (persiska), turkiska, semitiska språk (arabiska) och kinesiska. Vidare överväger vi att inkludera några slaviska språk (ryska, bulgariska, serbiska/kroatiska/bosniska) och några finsk-ugriska språk (finska, ungerska). Grunden för en maskinöversättningssatsning är tillgång till översättningskorpusar över käll- och målspråk. Sådana korpusar som länkats till varandra (alignment) på olika nivåer benämns parallellkorpusar. Ur dem kan översättningsekvivalenter utvinnas för såväl statistisk som regelbaserad maskinöversättning. Korpusarna används också för träning av maskinöversättning och automatisk utvärdering av resultaten. Länkningen understöds av att korpusarna annoteras med lingvistisk information. Parallellkorpusar av god kvalitet och tillräcklig omfattning utgör en bristvara, och uppbyggande och annotering av sådana korpusar är ett oundgängligt inslag i det fortsatta forskningsarbetet. Här utgör de studier och de experiment som hittills utförts på den turkisk-svenska korpusen en modell, som redan börjat tillämpas på parallellkorpusar över svenska-hindi och svenska-bulgariska.

Forskargruppens storlek och sammansättning

1 professor i datorlingvistik (Anna Sågvall Hein)
1 gästprofessor i datorlingvistik (Joakim Nivre)
3 tillsvidareanställda lektorer (Mats Dahllöf. Roussanka Loukanova, Beata Megyesi)
6 doktorander (Gustav Öquist, Eva Forsbom, Per Weijnitz, Ebba Gustavii, Markus Saers, Mattias Nilsson, Klas Prytz,)
1 forskningsassistent (Eva Pettersson)
2 forskningsingenjörer (Per Starbäck, Bengt Dahlquist)

Magisterstudenter på det fyraåriga språkteknologiprogrammet http://stp.lingfil.uu.se/

Forskare i de olika särspråken på institutionen i varierande utsträckning

2) Existerande forskningsdatabaser och verktyg

Scania corpus
Två samlingar tekniska manualer från Scania CV AB: Scania 95 och Scania 98. Scania 95 finns på 8 språk: svenska, nederländska, engelska, finska, franska, tyska, italienska och spanska. Omfånget på korpusen är ung. 220,000 ord per språk, se http://perdix.lingfil.uu.se/scania.html. Scania 98 finns på svenska, engelska och tyska. Omfånget på resp. språk är ung. 1,2 miljoner ord.
Användning: Båda korpusarna har använts för att definiera den kontrollerade vokabulären Scania Swedish och för att bygga motsvarande ord- och termdatabas. Scania 98 har använts för träning och utvärdering av maskinöversättning, för en- och tvåspråkig

termextraktion, för byggandet av ett sv-en maskinöversättningslexikon, för utveckling och utvärdering av metoder för ordlänkning (word alignment) samt för forskning om terminologi.

Scarrie Swedish newspaper corpus
Korpusen består av svensk tidningstext från Upsala Nya Tidning och från Svenska Dagbladet från 1995/1996. Omfånget på korpusen är 67 miljoner ord.
Användning: Utveckling av ett svenskt lexikon för stavnings- och grammatikkontroll, insamling av felskrivningar, samt utprovning av metoder för ord- och grammatikkontroll.

Scarrie Swedish Error Corpus Database
Korpusen omfattar c:a 9000 meningsfragment med fel och korrektioner. De har hämtats från Scarrie-korpusen och kan sökas via http://www.lingfil.uu.se/ling/ecd/.
Användning: Kategorisering av stavnings- och grammatikfel i svensk tidningstext samt fastställande av en uppsättning feltyper att inkludera i granskningsverktyget (Scarrie checker for Swedish).

Swedish political texts
Korpusen innehåller svenska regeringsförklaringar. De utkommer vanligen på fem språk: svenska, engelska, franska, tyska och spanska (sedan 1996). Korpusen omfattar 11,000 ord.
Användning: Forskning om ordlänkning och undervisning.

A Swedish dictionary of general language
Lexikonet innehåller 60,000 lemman och flera tusen fraser.
Användning: Grammatikkontroll, stavningskontroll, textpredicering, maskinöversättning, grammatikextraktion, textgradering, korpusuppbyggnad.

A multilingual term bank for automotive maintenance
Termbanken består av c:a 4000 svenska bilunderhållstermer med översättning till engelska.
Användning: maskinöversättning

Swedish Immigrant Newspaper corpus
Korpusen omfattar texter från Invandrartidningen, som utgivits på 8 språk: svenska, arabiska, engelska, finska, persisisk, polska, serbiska/kroatiska/bosniska och spanska. Swedish, Arabic, English, Finnish, Persian, Polish, Serbo-Croatian and Spanish. Korpusen har begränsad omfattning.
Användning: experiment med och forskning om ordlänkning och maskinöversättning

UPlug – en verktygslåda för bearbetning av parallella korpusar. Den utvecklades i en första version inom ramen för PLUG-projektet av dåvarande doktoranden Jörg Tiedemann och han har därefter kontinuerligt vidareutvecklat den. UPlug finns tillgänglig som open source (glp). Den innehåller moduler för grundläggande strukturell uppmärkning (text → xml), tokenisering, meningsuppdelning, ordlänkning, taggning mm.
Användning: Verktygslådan har fått stor spridning internationellt och nationellt.

BaseVocabulary: A package with scripts for creating a base vocabulary pool from a lemmatised and categorised corpus, including a Swedish and an English pool. Utvecklad och publicerad (gpl) av Eva Forsbom.

BaseModel: A package with a wordform-baseform mapper, sample server and client scripts, and a set of Swedish models. (Try a demo.) Utvecklad och publicerad (gpl) av Eva Forsbom.

MT Quality Evaluation Toolbox: A Java program for evaluation of translation quality, and meta-evaluation of evaluation measures. Utvecklad och publicerad (gpl) av Eva Forsbom.

parole2xml.pl: Perl script for converting SUC2.0-PAROLE file(s) into XML compliant file(s)). Short usage information is included in the head of the file, or printed via with `parole2xml.pl -h`. Utvecklad och publicerad (gpl) av Eva Forsbom.

3) Framtida behov

Databaser för maskinöversättning av nya språk

För att kunna utveckla vår maskinöversättning till att omfatta de orientaliska språk som undervisas och utforskas vid institutionen och viss europeiska språk, behöver vi kunna bygga upp omfattande tvåspråkiga databaser över svenska, engelska och de aktuella språken. I första hand kommer vi att satsa på utveckling av vår pågående forskning om statistisk maskinöversättning, vilken för träning kräver tillgång till språkdatabaser omfattande minimum 300.000 meningspar, dvs. mellan 2 och 3 miljoner löpord per språk.

Förutsatt att engelska inkluderas, kommer språkdatabaserna att ha en omfattande nyttjandefrekvens över hela världen. Tonvikten på svenska kommer givetvis att i första hand gynna det svenska och ev. också nordiska forskarsamhället. Samarbete/samordning inom Sverige kan i första hand ske via det väl etablerade nätverk, som den nationella forskarskolan. GSLT, representerar. Forskning om maskinöversättning i Sverige bedrivs vid UU, SU, LiU och Högskolan i Skövde.

Som utgångspunkt för kostnadsberäkningen tas upprättandet och hittills genomförd annotering av den turkisk-svenska parallellkorpusen och initieringen av den hindi-svenska. Den svensk-turkiska korpusen omfattar c:a 120.000 löpord på turkiska och 140.000 på svenska. Materialen är länkade till varandra, ordparsvis och meningsvis. Vidare är de ordklasstaggade. Visst manuellt arbete återstår. För att uppnå en korpusstorlek som är tillräcklig för statistisk maskinöversättning krävs en arbetsinsats om ytterligare 2 årsarbetstider, dvs. runt 2 miljoner kronor.

I snitt räknar vi sålunda med 2 miljoner kronor för varje nytt språkpar. Den samlade kostnaden för 5 språkpar (svenska-turkiska, svenska-hindi, svenska-kinesiska, svenska-persiska, svenska-arabiska) beräknas till 10 miljoner under en 5-årsperiod. Om engelska

ska tillföras, vilket är ett önskemål, bör man räkna med en halv miljon extra per språkpar, dvs. totalt 12,5 miljoner under perioden.

Infrastruktur för språkvetenskaplig forskning och utbildning

Vid Språkvetenskapliga fakulteten bedrivs forskning och undervisning i närmare 30 språk. Merparten av denna forskning sker med utnyttjande av språkdatabaser, och tillgängliga databaser kommer att utvecklas och anpassas även till undervisningsändamål. Verksamheten kommer att samlas inom EPARIT. I samband med planeringen av centret gjordes en inventering av fakultetens språkdatabaser. Det visade sig att databaser förelåg för så gott som alla språk men i varierande format, omfång och annoteringsstandard. Vidare fanns ett inte obetydligt antal lexikala databaser. Resultatet av korpusinventeringen bilägges (bilaga 1). Nu planeras en homogenisering av databaserna i olika avseenden, så att de blir sökbara med gemensamma verktyg. Att göra dem effektivt sökbara är ett stort åtagande, då det också rymmer lingvistisk märkning av databaserna samt utvidgning av dem till erforderligt omfång för att de på ett optimalt vis ska kunna tjäna sina ändamål som datakällor för forskning och undervisning. Även om generella, maskininlärningsmetoder kommer att användas för den lingvistiska annoteringen, måste de kompletteras med språkberoende manuella insatser. Vi räknar med en samlad insats motsvarande 3 miljoner årligen under en 5-årsperiod. Samma standarder kommer att användas för dessa språkdatabaser som för dem som utvecklas i och för den språkteknologiska forskningen. Härigenom blir databaserna ömsesidigt användbara för såväl språkvetenskapliga som språkteknologiska forskningsuppgifter och tillämpningar.

Med fokus på svenska och ett nationellt perspektiv

Nedan upptas behov av språkdatabaser med verktyg för svenska, som bedöms vara av ett generellt nationellt intresse.

Svensk BLARK – Basic Language Resource Kit:
* Med omfattande metadata om befintliga svenska korpusar, språkteknologiska verktyg och utvärderingsresurser.

Svensk nationell korpus:
* Ca 100 miljoner ord.
* Med talat och skrivet material från olika genrer som möjliggör genrestudier.
* Med omfattande metadata som möjliggör uppdelning i delkorpusar.
* Med bevarad logisk struktur (stycken, listor, rubriker m.m.) som möjliggör studier på textnivå.
* Med annotering av åtminstone morfologisk information och grundform för ord, förhoppningsvis även lexikal-semantisk information.
* Med eventuell annotering av namngivna enheter och fraser.
* Med eventuell annotering av syntaktisk information.

Parallella/Jämförbara korpusar för svenska:
* Med infödd svenska och inlärarsvenska, vilket möjliggör studier av andraspråksinlärning.
* Med svenska och förenklad svenska, vilket möjliggör studier av lättlästhet.
* Med artiklar om samma händelse skrivna ur olika perspektiv, vilket möjliggör studier om urval och komposition.
* Med fullständiga dokument och sammanfattningar, vilket möjliggör studier av sammanfattningsteknik.
* Med svenska och andra språk, vilket möjliggör studier av översättningsteknik och utvinning av översättningsekvivalenter.

Grundläggande allmänspråkliga lexikon för svenska:
* Med enspråkig information.
* Med flerspråkig information.
* Med information om ordböjning och ordbildning.
* Med information om lexikal-semantisk släktskap.

Svenska utvärderingsdata och -program:
* För morfologisk analys.
* För ordbetydelsedisambiguering.
* För uppmärkning av namngivna enheter.
* För informationssökning

Det handlar om satsningar på många miljoner för att tillgodose de språkteknologiska behoven för svenska. Se vidare den ansökan om planeringsbidrag för en infrastruktur för svenska som ingivits tidigare med Lars Borin, Göteborg, som koordinator.

Inventering av språkliga korpusar och datoriserade lexikon vid Språkvetenskapliga fakulteten

Nedanstående är en lista på korpora som etablerats av forskare vid Språkvetenskapliga fakulteten. Listan är inte fullständig, men ger en överskådlig bild av den korpusbaserade forskningen som bedrivs.

Asiatiska språk

Kinnauri Corpus

Etablerad av: Anju Saxena
Språk: Kinnauri-Engelska
Omfattning: c:a 30 000 ord
Språktyp: muntlig sakprosa, berättelser, sagor, sånger, konversationer
Period: 1989 och framåt
Egenskaper: Det är en muntlig korpus, fonetiskt transkriberad med åtföljande annotering: (1) morfemsegmentering; (2) morfem-för-morfem-översättning av hela korpusen; (3) en del är också syntaktiskt annoterad; (4) fri översättning till engelska.
Användning: forskning, forskarutbildning, undervisning inom kurserna: Funktionell grammatik/Grammatik 2, Lingvistik 2, Lingvistik 1, Syntax, Grammatik

Tinan corpus

Etablerad av: Anju Saxena, delvis i samarbete med CIIL, Indien inom projeket ”Digital documentation of Indian minority languages”
Språk: Tinan-Engelska
Omfattning: c:a 20 000 ord
Språktyp: muntlig sakprosa, berättelser, sagor, sånger
Period: 1989-90, 2003-2005
Egenskaper: Det är en muntlig korpus, fonetiskt transkriberad med åtföljande annotering: (1) morfemsegmentering; (2) morfem-för-morfem-översättning av hela korpusen; (3) fri översättning till engelska.
Användning: forskning, forskarutbildning och för dokumentation av hotade språk. Korpusen kan också användas i undervisningen på samma sätt som kinnaurikorpusen.

Engelska

CENE - The Corpus of English Newspaper Editorials (1900–1993)

Etablerad av: Ingrid Westin
Språk: Brittisk engelska
Omfattning: 502,834 ord (864 tidningsledare)
Språktyper: tidningsledare från *Daily Telegraph, Guardian* och *The Times*
Period: 1900–1993 (10 perioder: 1900, 1910, 1920, 1930 osv.)
Användning: avhandling

UPC - The Uppsala Press Corpus

Etablerad av: Margareta Westergren Axelsson
Språk: Brittisk engelska
Omfattning: 354,084 ord (88 texter)
Språktyper: tidningstexter som motsvarar tidningstexter i LOB (Lancaster/Oslo-Bergen Corpus med texter från 1961): reportage, tidningsledare och recensioner
Period: 1994
Användning: forskarutbildning

UCTT - The Uppsala Corpus of Travel Texts

Etablerad av: Tore Nilsson
Språk: Brittisk engelska
Omfattning: 497,091 ord (139 texter)
Språktyper: turistbrochyrer, researtiklar, resehandböcker
Period: 1994–1999
Användning: forskarutbildning

The Victorian Corpus and the Contemporary Corpus of British Children's Literature

Etablerad av: Hanna Andersdotter Sveen
Språk: Modern brittisk engelska
Omfattning: 2 x 95,000 ord (2 x 10 texter); Contemporary Corpus är baserad på ett urval av texter i BNC
Språktyper: barnlitteratur för pojkar och flickor
Period: 1851–1858 och 1981–1993
Användning: forskarutbildning

Finsk-ugriska språk

Estnisk morfologi

Etablerad av: Virve Raag
Språk: Estniska
Omfattning: pluralformer: 30 000, illativ sg 9000, superlativ 2700
Texttyp: tidningstext
Period: år 1887, 1907, 1923, 1938, 1961, 1991
Användning: forskarutbildning, forskning, uppsatskrivande

Estniska telefonsamtal

Etablerad av: Leelo Keevallik
Språk: Estniska
Omfattning: 103 000 ord
Språktyp: samtal
Period: 1998/99
Användning: forskning, forskarutbildning, undervisning

Sverigeestniskt talspråk

Etablerad av: Raimo Raag, Virve Raag, Leelo Keevallik
Språk: Estniska
Omfattning: 253 timmar

Språktyp: radioprogram, forskningsintervjuer, sociala tillställningar
Period: 1979-2006
Användning: forskning, undervisning

Vrajavali ordlistor

Språk: Vrajavali & English
Omfattning 2-300 sidor
texttyp: drama, lyrik (skönlitteratur)
Period: 15-1600 talet
användning: forskning

Klassiska språk

Bysantinsk hagiografi

Etablerad av: Jan-Olov Rosenqvist
Språk: bysantinsk grekiska
Omfattning: 18,3 miljoner tecken
Texttyp: hagiografi
Period: 300-1300-talen
Användning: forskning

CLEDBase

Etablerad av: Christer Henriksén, Henrik Vitalis, Patrik Granholm
Språk: latin, viss grekiska
Omfattning: f.n. ca 4500 ord
Texttyp: romerska, icke-kristna gravinskrifter på daktylisk vers
Period: ca 50 f.Kr – 350 e.Kr.
Användning: forskning

Romanska språk

CCUCFI - Le Corpus Contrastif d'Uppsala du Conditionnel français et italien

Etablerad av: Hans Kronning
Språk: franska och italienska
Omfattning: Motsvarande 8 böcker, 1000 konditionalisförekomster
Språktyp: romaner
Period: 1900-tal
Användning: forskning

Corpus de journaux francophones

Etablerat under ledning av Mats Forsgren
Språk: franska från Schweiz och Belgien
Omfattning: 49 nr
Språktyp: dagstidningar
Period: 80-90-talet
Användning: forskarutbildning, forskning, uppsatsskrivande

Corpus des periodicas hispanos

Etablerat under ledning av Mats Forsgren
Språk: spanska
Omfattning: 116 nr
Språktyp: dagstidningar
Period: 80-90-talet
Användning: forskarutbildning, forskning, uppsatsskrivande

CUCF - Le Corpus d'Uppsala du Conditionnel français

Etablerad av: Hans Kronning och Coco Norén
Språk: franska
Omfattning: Motsvarande 30 böcker, 5000 konditionalisförekomster
Språktyp: romaner, teaterpjäser, dagstidningar, vetenskaplig prosa
Period: 1700-tal, 1800-tal, 1900-tal, 2000-tal

Dialogues entres jeunes Français

Etablerat av: Coco Norén
Språk: franska
Omfattning: 5½ timmar, transkriptioner
Period: 1990-tal
Språktyp: informella tvåpartssamtal
Användning: forskarutbildning, forskning, undervisning

Fransk-svenskt Parallellkorpus

Etablerat av: Carina Andersson
Språk: franska-svenska
Omfattning: 2 miljoner ord
Period: 1900-tal
Språktyp: romaner
Användning: forskarutbildning, forskning

ISU - Interaktionsstudier Stockholm - Uppsala

Etablerad av: Cilla Häggkvist, Johan Gille
Språk: spanska
Omfattning: 7 samtal à 40 minuter
Språktyp: vardagliga samtal mellan universitetsstuderande
Period: 2000-tal
Användning: forskning, C- och D-uppsatser, inspelat under kurserna Samtalsanalys och Diskursanalys

The Uppsala French TAP-Corpus ("Think-Aloud-Protocols")

Etablerad av: Kerstin Jonasson
Språk: Franska-Svenska
Omfattning: ca 50 sidor
Texttyp: "tänka-högt-protokoll"
Ålder: etablerat 1993-1995
Användning: forskning

Semitiska och turkiska språk

Muntliga berättelser från Östra Turkiet, Norra Syrien och Jemen

Etablerad av: Bo Isaksson, Ablahad Lahdo
Språk: arabisk dialekt och arameisk dialekt
Omfattning: 20 timmars ljudfiler varav ca 10 % transkriberat
Språktyp: muntliga berättelser
Period: 2000-2005
Användning: forskning, avhandling, uppsatskrivande på Orientalistikprogrammet och Arabiska C-D (kandidat- och magisteruppsatser)

Turkisk-Svensk Korpus / Turkish-Swedish Corpus

Etablerad av: Prof. Anna Sågvall Hein, Prof. Éva Á Csató Johanson, dr. Beáta Bandmann Megyesi, dr. Bengt Dahlqvist
Språk: Turkiska-svenska / Turkish-Swedish
Omfattning: 118 000 token på turkiska och 144 000 token på svenska.
Språktyp: skönlitterär text och facktext
Period: 1900-tal och 2000-tal
Användning: forskning, uppsatskrivande, kurser i ämnet turkiska språk samt orientalistikprogrammet med turkisk inriktning

Hebreiska texten i UUB O Heb 32

Etablerad av: Mats Eskhult
Språk: hebreiska
Omfattning: ca 100 trycksidor från handskriftens 200 foliantsidor.
Språktyp: bibeltext
Period: 1700-talet
Användning: forskning

Vrajavali ordlistor

Etablerad av: William Smith
Språk: Vrajavali & English
Omfattning 200-300 sidor
Språktyp: drama, lyrik (skönlitteratur)
Period: 1500-1600 talet
Användning: forskning

Slaviska språk

The Uppsala Russian Corpus -- Upsal'skij korpus russkich tekstov

Etablerad under ledning av prof. Lennart Lönngren
Språk: ryska
Omfattning: 1 miljon löpord
Språktyp: skönlitteratur sakprosa, stort urval
Period: 1900-talet, 1960-1988, 1985-1989
Användning: forskning, grundutbildning

Svenska

The Swedish Map Task Corpus

Etablerad av: Pétur Helgason
Språk: svenska
Omfattning: 70 minuter, ca 8 000 ord
Språktyp: samtal
Period: samtida
Användning: 3 forskarutbildningar, C- och D-uppstater

SCARRIE SvD9596

Etablerad av: Anna Sågvall Hein
Språk: Svenska
Omfattning: 47.4 miljoner ord
Texttyp: dagstidningstext
Period: 1995-96
Användning: forskning, forskarutbildning, uppsatskrivande, inom kursen/programmet: STP

SCARRIE UNT9596

Etablerad av: Anna Sågvall Hein
Språk: Svenska
Omfattning: 22.8 miljoner ord
Texttyp: dagstidningstext
Period: 1995-96
Användning: forskning, forskarutbildning, uppsatskrivande, inom kursen/programmet: STP

UNT92

Etablerad av: Anna Sågvall Hein
Omfattning: 6.5 miljoner ord
Texttyp: dagstidningstext
Period: 1992
Användning: forskning, forskarutbildning, uppsatskrivande, inom kursen/programmet: STP

ECD - Error Corpus Database (http://www.lingfil.uu.se/ling/ecd/)

Etablerad av: Anna Sågvall Hein
Språk: Svenska
Omfattning: ca 9 000 meningsfragment
Texttyp: dagstidningstext
Period: 1995-96
Användning: forskning, forskarutbildning, uppsatskrivande, inom kursen/programmet: STP

SCANIA (http://perdix.lingfil.uu.se/scania.html)

Etablerad av: Anna Sågvall Hein
Språk: Svenska
Omfattning: ca 200 000 ordformer
Texttyp: facktext
Period: 1995-2005
Användning: forskning, forskarutbildning, uppsatskrivande, inom kursen/programmet: STP

Lärobokstexter från projektet "Elevers möte med skolans textvärldar"

Etablerad av: Caroline Liberg, Agnes Edling, Jenny Folkeryd, Åsa af Geijerstam
Språk: Svenska
Omfattning: 27 953 ord
Språktyp: skönlitterär text, diskursiv lärobokstext
Period: insamlat 1999-2003 (tryckår ca 1980-2001)
Användning: forskning, forskarutbildning

Elevskrivna texter från projektet "Elevers möte med skolans textvärldar"

Etablerad av: Caroline Liberg, Åsa af Geijerstam, Agnes Edling, Jenny Folkeryd
Språk: Svenska
Omfattning: ca 400 kortare texter
Språktyp: skönlitterär text, sakprosa
Period: insamlat 1999-2003
Användning: forskning, forskarutbildning

Samtal, åldrande och identitet 1: identitetsskapande strategier i äldresamtal

Etablerad under ledning av Bengt Nordberg
Språk: svenska
Omfattning: dialoger 27 timmar, gruppsamtal 14 timmar, delvis transkriberat
Språktyp: arrangerade dialoger med medelålders och pensionerade kvinnor,
dels av informella gruppsamtal (kafferep, kortspel o.l.) mellan äldre kvinnor.
Egenskaper: I dialogerna samtalar sammanlagt 40 kvinnor som inte är bekanta med varandra i syfte att lära känna varandra, dels över generationerna, dels inom dem. Deltagarna i gruppsamtalen känner i de flesta fall varandra sedan tidigare.
Period: 1997–2000
Användning: forskning

ÅbEsk - Kontinuitet och förändring i nutida talspråk. Återbesök i Eskilstuna

Etablerad under ledning av Bengt Nordberg
Språk: svenska
Omfattning: 78 timmar, delvis transkriberat
Språktyp:
Egenskaper: Korpusen är insamlad för ett projekt som följer upp Eskilstunaundersökningen från 1967. De infödda informanterna har samma sociala fördelning och inspelningarna samma stilistiska kvaliteter som i den tidigare undersökningen (se Svenska stadsmål 1). 13 talare ingick också i 60-talsundersökningen, medan övriga 72 är nya.
Användning: forskning

GIC - Samtal i nödsituation. Telekommunikationen vid en giftinformationscentral

Etablerad under ledning av Bengt Nordberg
Språk: svenska
Omfattning: 17 timmar, transkriptioner
Språktyp: Samlingen består av 377 autentiska telefonsamtal från privatpersoner över hela landet till giftinformationscentralen gällande akuta eller befarade förgiftningar.
Egenskaper: Den rådgivande personalen består av 13 apotekare/informatörer.

Texter i europeiska skrivsamhällen

Etablerad under ledning av Britt-Louise Gunnarsson
Språk: svenska
Omfattning: 70 intervjuer
Språktyp: 70 intervjuer med personer ansvariga för transkriptioner skrivande av olika texter inom en bank, en ingenjörsbyrå, en historisk institution samt en yrkesmedicinsk institution i tre länder: Sverige, Tyskland och Storbritannien.
Period: 1994–96

Interaktionen vid seminarier

Etablerad under ledning av Britt-Louise Gunnarsson
Språk: svenska
Omfattning: 20 doktorandseminarier, transkriptioner
Språktyp: doktorandseminarier från tre institutioner vid Uppsala universitet: en humanistisk, en samhällsvetenskaplig, en naturvetenskaplig.
Period: 1992–95

FORTIS - Sverigefinnars två språk – språkbruk och attityder hos två generationer

Etablerad under ledning av Bengt Nordberg
Språk: finska-svenska
Omfattning: 54 timmar, transkriptioner
Språktyp: formella och informella intervjuer, gruppsamtal med självrekryterade deltagare
Egenskaper: Situationellt varierade inspelningar med två generationerSverigefinnar (16–19 och 35–55 år, barn och föräldrar, båda könen) boende i en Stockholmsförort. Ämnesvalet i gruppsamtalen är fritt, de informella intervjuerna tar upp vardagsliv, informanternas migrationshistoria och situation som sverigefinne, de formella intervjuerna är autenticitetstrogna anställningsintervjuer och radiointervjuer.

Talutveckling

Etablerad av: Birgitta Garme
Språk: svenska
Omfattning: Ca 170 samtal
Språktyp: samtal med elever i olika åldrar, från skolår 2 till gymnasiet, som arbetar med olika typer av samtalsuppgifter. Oftast samtalar eleverna i små grupper utan vuxen ledare, men korpusen innehåller även lärarledda diskussioner.
Period: 1988–92

Kommunikationen vid en larmcentral

Etablerad av: Bengt Nordberg
Språk: svenska
Omfattning: 8 timmar, transkriptioner
Språktyp: Autentiska inspelningar av telefon- och radiokommunikationen vid en länsalarmeringscentral mellan uppringare (hjälpsökande), operatör och utförare. Samtalen rör akuta sjukdomsfall, olyckor, bränder, ambulansbeställningar, driftstopp etc.
Period: 1986

Läkarens och lekmannens begreppsvärldar

Etablerad av: Ulla Melander Marttala
Språk: svenska

Omfattning: 95 intervjuer, 15 läkar-patientsamtal
Språktyp: En semantisk och samtalsanalytisk undersökning av föreställningar och begrepp kring reumatiska sjukdomar. Två delar: 95 semantiska djupintervjuer, audioinspelade; 15 läkare–patientsamtal med intervjuer och videovisningar, audio- och videoinspelningar.
Period: 1986-89

Ungdomars samtalsstil

Etablerad av: Bengt Nordberg
Språk: svenska
Omfattning: 2 timmar, transkriptioner
Språktyp: Fria samtal mellan nära vänner
Egenskaper: Flickor och pojkar i åldern 12–16 år i enkönade grupper. Samtalen kretsar kring personliga intressen, vardagliga händelser, kamratskvaller. Deltagarna uppväxta i Uppsala.
Period: 1984–88

Språkanvändning och språkmiljöer i staden och på landet

Etablerad av: Bengt Nordberg
Språk: svenska
Omfattning: 29 timmar
Språktyp: Informella intervjuer
Egenskaper: 20 informanter i åldern 30–50 år, hälften boende i urban, hälften i rural miljö (Mellansverige, södra Norrland), hälften män, hälften kvinnor, hälften arbetare, hälften tjänstemän. Intervjuerna rör deltagarnas kommunikationsmiljö, kontaktnät och språkaktiviteter, upplevda förändringar i dessa och attityder till dessa.

Stad och omland: Urbaniseringen speglad i språket

Etablerad av: Mats Thelander
Språk: svenska
Omfattning: 52 timmar
Språktyp: Halvstrukturerade telefonintervjuer
Egenskaper: 120 informanter från norra Västerbotten resp. Eskilstuna kommun. De intervjuade personerna (i åldern 20–80 år) är slumpmässigt valda ur sju geografiskt definierade populationer: födda och boende på landet i Västerbotten (12 informanter), födda och boende i Skellefteå stad (12), födda på landet i Västerbotten men inflyttade till Skellefteå stad (24), födda och boende på landet i Eskilstuna kommun (12), födda och boende i Eskilstuna stad (12), födda på landet utanför Eskilstuna men inflyttade till staden (24) och födda på landet i Västerbotten men vid inspelningen bosatta i Eskilstuna stad (24). Varje intervju är ca 30 min. lång och i mer eller mindre ledig sam talston. Mest diskuteras flyttning och språk. Intervjuaren är en rikssprakstalande man i 35-årsåldern.
Period: 1978-79

Barnets språkliga identifikation (BSI)

Etablerad av: Bengt Nordberg
Språk: svenska
Omfattning: 78 timmar, delvis transkriberat
Språktyp: Inspelade situationer: högläsning, kommunikationsspel, kortspel och diskussion.
Egenskaper: Situationellt varierade inspelningar av 85 infödda grundskoleelever i Eskilstuna under tre på varandra följande år, årskurs 1–3, 4–6, 7–9. Eleverna är jämnt könsfördelade men socialt differentierade.
Period: 1977–79

Språk, roll och sociala relationer (Burträskundersökningen)

Etablerad av: Mats Thelander
Språk: svenska
Omfattning: 29 timmar gruppsamtal + 3,5 timmar intervjuer.
Språktyp: gruppsamtal
Egenskaper: 14 situationellt varierade gruppsamtal med fyra deltagare i varje grupp – alla 56 talarna (i åldern 14–64 år) från dåvarande Burträsk kommun. Varje inspelning varar ca 2 timmar och samtalsstilen är för det mesta otvungen. Under andra hälften av samtalet medverkar också en utomstående forskare. Ämnet för samtalen är Bygd i förvandling (tankar om Burträsk inför den då aktuella kommunsammanslagningen). Nio av deltagarna i gruppsamtalen har senare spelats in i mer formella intervjuer.
Period: 1973-75

Svenska stadsmål 1. Eskilstuna

Etablerad av: Bengt Nordberg
Språk: svenska
Omfattning: 53 timmar, transkriptioner
Språktyp: Samtalsliknande intervjuer
Egenskaper: 83 infödda, socialt, åldersmässigt och könsmässigt differentierade Eskilstunabor om vardagliga ämnen, personliga minnen och lokalhistoriska notiser. Några dialoger utan intervjuare ingår.
Period: 1967–68

Facktexter under 1900–talet

Etablerad av: Britt-Louise Gunnarsson
Språk: svenska
Omfattning: Totalt omfattar den inlästa 1900-talskorpusen 1340 normalsidor om 3000 tecken; de vetenskapliga texterna 650 normalsidor och de populärvetenskapliga 690.
Språktyp: 180 inskannade vetenskapliga och populärvetenskapliga texter (90 inom varje genre) inom områdena (national)ekonomi, (lung)medicin och (elektro)teknik (60 inom varje område) från perioderna 1895–1905, 1935–1945 och 1975–1985.
Egenskaper: Från varje period är 60 texter inskannade knutna till två ämnen: bank- och kreditväsen samt skatter inom ekonomi, lungsjukdomar samt hud- och könssjukdomar inom medicin, elteknik samt telekommunikationer inom teknik. Varje område är således representerat med 60 texter, fem från vardera ämne, från respektive genre och period (se nedan).

Fackspråkens framväxt

Etablerad av: Britt-Louise Gunnarsson
Språk: svenska
Omfattning: Totalt omfattar 1700- och 1800-talskorpusen 1590 normalsidor om 3000 tecken; de vetenskapliga 940 normalsidor och de populärvetenskapliga 650 normalsidor.
Språktyp: 180 inskannade vetenskapliga och populärvetenskapliga texter inom områdena (national)ekonomi, medicin och teknik från 1700-talet samt perioderna 1800 1849 och 1850 1880. Från varje period är 60 texter inskannade knutna till två ämnen: bank- och kreditväsen samt handel inom ekonomi, lungsjukdomar samt hud- och könssjukdomar inom medicin, elteknik samt telekommunikation och mekanik inom teknik. Varje ämne är således

representerat med 60 texter, fem från vardera ämne, från respektive genre och period (se nedan).

Texter i europeiska skrivsamhällen

Etablerad av: Britt-Louise Gunnarsson
Språk: svenska
Omfattning:
Språktyp: Insamlade och systematiserade svenska, engelska och tyska texter tillkomna inom miljöerna bank, ingenjörsbyrå, yrkesmedicinsk institution och historisk institution i Sverige, Storbritannien och Tyskland.
Egenskaper: Texterna representerar fyra olika genrer: 1. riktade till personalen (personaltidning); 2. riktade till ägare, kunder och allmänhet (pressmeddelande, årsberättelse); 3. riktade till (presumtiva) kunder (anbud, brev, broschyr, expertutlåtande, forskningsansökan, läromedel, populärvetenskaplig artikel, projektbeskrivning, protokoll, vetenskaplig artikel); 4. avsedda för imagearbete (allmän presentation, annonser). Antal genrer från varje närmiljö och område varierar (se nedan).

Samnordisk runtextdatabas

Etablerad under ledning av: Lennart Elmevik och Lena Peterson
Språk: runskrift
Omfattning: ca 6 000 inskrifter.
Språktyp: nordiska runtexter, inklusive sådana funna utanför Norden,
Egenskaper: Texterna återges i translittererad och normaliserad form, dessutom i översättning till engelska. Till texterna är knutna uppgifter om tidsperiod, fyndområde, excerperad källa, typ av föremål m.m. Genom speciellt utformade program kan sökningar av olika slag göras i textfilerna.
Användning: forskning inom ett flertal discipliner

2006-11-21

Språkdatabaser inom ämnet lingvistik vid Uppsala universitet

Åke Viberg

Inom ämnet lingvistik vid Uppsala Universitet (LingU) bedrivs arbete inom tre områden av relevans för denna enkät: lexikala databaser, flerspråkiga parallellkorpusar samt inlärar-korpusar.

Lexikala databaser

Åke Viberg arbetar med lexikala databaser inom ramen för ett teoretiskt studium av lexikonets organisation i tvärspråkligt perspektiv. Internationellt finns (utvecklas) för närvarande olika former av omfattande betydelsebaserade lexikon på dator som kan kopplas till ett antal parallella lexikon för andra språk. Vid LingU bedrivs arbete med (1) Ordnät och (2) Ramnät.

Svenskt OrdNät. Ordnät bygger på semantiska relationer som synonymi, antonymi, hyponymi och meronymi (helhet/del). En grundversion föreligger av ***Svenskt OrdNät*** (SWN1.1) som är uppbyggt enligt samma principer som de lexikala databaser som utarbetats inom EuroWordNet (EWN). För närvarande omfattar databasen 25 000 begrepp (synsets) och 34 000 ord fördelade på 28 000 substantiv och 6 000 verb hämtade från allmänspråket (validerat gentemot frekvenserna i Stockholm-Umeåkorpusen, SUC). Storleken torde svara mot den genomsnittliga storleken för motsvarande lexikon i EWN-format men är betydligt mindre än det ursprungliga (Princeton) WordNet (WN) som (nov. 2006) omfattar närmare 130 000 ord som ett resultat av en verksamhet som mest intensivt sträckte sig över hela 1990-talet. Det vore önskvärt att väsentligt öka täckningsgraden även hos SWN men detta är ett resurskrävande arbete, även om utbyggnaden kan ske i betydligt snabbare takt än vid uppbyggnaden av den första versionen. Alla grundläggande begrepp är redan inkodade då det gäller substantiv och verb. Det rör sig främst om att fylla på med ord som ligger relativt lågt i de hierarkier (hyponymi, partonymi,) som redan kodats. De grundläggande begreppen är dessutom redan kopplade till det mellanspråkliga index som utvecklats inom EWN (ILI-kopplingar). Det är mindre intressant att upprätta sådana kopplingar för specifika och lågfrekventa ord.

Ordnät har främst uppfattats som en språkteknologisk resurs. För sådana ändamål krävs i många fall en högre täckningsgrad än vad den nuvarande versionen av Svenskt OrdNät har, vilket skulle kunna åtgärdas genom en utbyggnad. Samtidigt är ordnät av stort intresse för sådant som enl. rundskrivelsen inte ”är att betrakta som strikt språkteknologi”. De är en basresurs även för språkundervisning och utvecklande av språkläromedel. George A Miller som ursprungligen utvecklade WordNet tänkte sig detta som en psykolingvistisk modell vilket fortfarande återspeglas på projektets hemsida: ”WordNet® is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory.” Även för kontrastiva och typologiska studier på den lexikala nivån är ordnät en resurs genom att tillhandahålla en semantisk klassifikation av ett omfattande basordförråd.

Internationella kontakter: SWN är knutet till Global WordNet Organization och finns upptaget bland projekten på dess hemsida.

Svenskt RamNät. Ordnät relaterar isolerade ord på grundval av betydelsen. Inom FramNet utvecklat vid Berkeley under ledning av Charles Fillmore har byggts upp en lexikal databas på grundval av Fillmores ramsemantik. En ram länkar semantik (begreppsstruktur) och syntax och utgör en schematiserad representation av situationer med deltagare, rekvisita och andra begreppsliga roller, vilka var och en utgör ett ramelement. Ramarna bildar ett omfattande system och uppgår för närvarande till ett antal överstigande 600 ramar. Ord grupperas med utgångspunkt från de begreppsliga scheman (ramar) de bygger på och deras semantiska och syntaktiska kombinatoriska egenskaper undersöks med hjälp av omfattande korpusdata. (FrameNet har främst utnyttjat British National Corpus, BNC). En tydlig koppling finns mellan ramar och argumentstruktur.

Vid LingU har genomförts pilotstudier för att utveckla en svensk motsvarighet. Åke Viberg är medlem i Global FrameNet Organization och deltog med en presentation av arbete kring Svenskt RamNät i organisationens första symposium i Berkeley med Charles Fillmore som värd i okt. 2004. Fillmore har också lovat att stödja projektet och kontakt knöts vid detta möte med Manfred Pinkal (Saarbrücken) Ulrich Heid (Stuttgart) och Hans Boas (Austin/Texas) som är intresserade av ett samarbete inom ramen för German WordNet samt med Carlos Subirats som inbjudit till samarbete med Spanish FrameNet. Samarbetet har dock inte utvecklats vidare i avvaktan på finansiering för Svenskt RamNät men skulle snabbt kunna vidareutvecklas.

Inlärarkorpusar

Åke Viberg har sedan tidigare en *Datoriserad korpus av inlärardata och tvåspråkiga data* från talare i skilda åldrar. På grundval av de transkriptioner av inspelat tal som gjorts med inlärare av svenska, tvåspråkiga elever samt infödda kontroller i skilda åldrar har en omfattande korpus upprättats där orden i inspelningarna är taggade med ordklassbeteckningar, vilket möjliggör ett stort antal skilda analyser. Totalt täcker materialet utvecklingen från fem års ålder till vuxen ålder. De flesta inspelningarna gjordes under 80-talet och början av 90-talet.

Flerspråkig parallellkorpus

Åke Viberg arbetar inom sin egen forskning inriktad mot korpusbaserad kontrastiv analys med en pilotkorpus bestående av utdrag ur 10 svenska originalromaner (totalt drygt 200 000 ord) och deras översättningar till engelska, tyska, franska och finska. I viss utsträckning finns också originalromaner på de andra språken med deras svenska översättningar. Planer finns att bygga ut med andra genrer än romaner och med andra språk, bl a flera av dem som finns vid Institutionen för lingvistik och filologi/UU där lingvistik ingår som ämne.

Anju Saxena som är lektor i lingvistik arbetar också med flerspråkiga korpusar men är bortrest på konferens till 28/11, varför vi ber att få inkomma med en komplettering.

REFERENSER

Svenskt OrdNät
http://www.lingfil.uu.se/ling/swn.html

Internationella lexikala databaser
http://wordnet.princeton.edu
http://www.globalwordnet.org/
http://framenet.icsi.berkeley.edu/
http://www.ub.es/gilcub/SIMPLE/simple.html

Växjö 2006-10-31

Vetenskapsrådet
Database Infra-Structure Committee
SE-10378 Stockholm

Kartläggning av databasresurser inom språkteknologi vid Växjö universitet

Som svar på skrivelsen angående kartläggning av databasresurser inom språkteknologi kommer här en beskrivning av dagsläget och framtida behov för språkteknologigruppen vid Matematiska och systemtekniska institutionen, Växjö universitet.

Med vänlig hälsning,

Joakim Nivre

Professor i datalingvistik
Matematiska och systemtekniska institutionen
Växjö universitet

1. Språkteknologisk forskning som kräver tillgång till stora databaser och databasverktyg

Det VR-finansierade projektet *Inductive Dependency Parsing* studerar maskininlärningsmetoder för automatisk syntaktisk analys av naturligt språk och är helt och hållet beroende av storskaliga textdatabaser, uppmärkta med syntaktisk information, s.k. *trädbanker*. Dessa databaser används för att genom induktiv inlärning skapa modeller för syntaktisk analys samt för att utvärdera kvaliteten på de inducerade modellerna. Förutom själva databaserna krävs verktyg för annotering, visualisering, induktiv inlärning samt utvärdering. Projektgruppen består av professor Joakim Nivre samt tre doktorander (Susanne Ekeklint, Johan Hall och Jens Nilsson).

2. Befintliga forskningsdatabaser vid institutionen

För den pågående forskningen används ett femtontal trädbanker, vilka samtliga skapats för att bedriva såväl korpusbaserad lingvistisk forskning som språkteknologiskt utvecklings- och utvärderingsarbete. Bland dessa databaser, som används dagligen i vår forskning, kan nämnas:

Penn Treebank (engelska)
Penn Chinese Treebank
Prague Dependency Treebank (tjeckiska)
Prague Arabic Dependency Treebank
Slovene Dependency Treebank
Danish Dependency Treebank
Alpino Treebank (holländska)
Tiger Treebank (tyska)
Cast3LB Treebank (spanska)
Turin University Treebank (italienska)
Sinica Treebank (kinesiska)
TüBa-J Treebank (japanska)
Floresta sintactica (portugisiska)
BulTreeBank (bulgariska)
Talbanken (svenska)

Med undantag för den svenska databasen Talbanken, som ursprungligen utvecklades vid Lunds universitet under 1970-talet men som vi konverterat till ett modernt och för våra ändamål mera lämpligt format, har vi inte själva deltagit i utvecklingen av databaserna, utan har endast förvärvat dem för användning i forskningen. Flertalet av dessa resurser är tillgängliga till ett relativt lågt pris, men enstaka databaser kan kosta upp till USD 2500. Dock är de flesta trädbanker än så länge små, och det finns ett stort behov av att utveckla större och bättre annoterade databaser, liksom bättre verktyg för att utnyttja dem (se nedan).

Vi samarbetar med ett flertal institutioner, både nationellt och internationellt, kring uppbyggnaden och utnyttjandet av trädbanker. Vi har även koordinerat ansökningar till både Riksbankens jubileumsfond och Vetenskapsrådet för att få anslag till storskalig trädbanksuppbyggnad framför allt för svenska (än så länge utan framgång). Det mindre arbete vi gjort för att göra Talbanken tillgänglig för modern, språkteknologisk forskning har delvis möjliggjorts av ett tidigare VR-finansierat projekt (*Stochastic Dependency Grammars for Natural Language Parsing*). Vi har också koordinerat ett nordiskt nätverk, *Nordic Treebank*

Network, med stöd från Nordiska ministerrådets språkteknologiprogram (2003–2005), där vi bland annat arbetat med att ta fram riktlinjer för en standardisering av trädbanksdatabaser.

3. Framtida behov av resurser för för utveckling m.m. av databaser

Som redan nämnts finns ett stort behov av större och bättre annoterade språkdatabaser med syntaktisk information, s.k. trädbanker. Eftersom ett viktigt mål med den språkteknologiskt inriktade forskningen är att ta fram språkoberoende metoder är det viktigt att dessa resurser täcker så många språk som möjligt, men i Sverige är den främsta prioriteten att få fram bättre resurser för svenska, där de enda befintliga trädbankerna dels är små, dels är föråldrade.

För att få fram en trädbank för svenska som ligger i nivå med de bästa tillgängliga resurserna för andra språk (främst engelska men även t.ex. tyska och tjeckiska) krävs ett storskaligt projekt som inrymmer datainsamling och semi-automatisk annotering av textmaterial motsvarande minst en miljon löpord. Kostnaden för ett sådant projekt beräknas till mellan 10 och 30 MSEK, beroende på databasens storlek samt noggrannheten i annoteringen.

UTVECKLINGEN AV HSFR:s STÖD TILL INFORMATIONSTEKNOLOGIRELATERAD FORSKNING[1]

* December 1984	Regeringen uppdrar åt HSFR att redovisa åtgärder för att höja den allmänna kompetensen inom rådets område
* Bå 1985/86	HSFR anvisas särskilda medel för ”forskning om datateknikens användning”
* Februari 1986	HSFR:s svar på regeringsuppdraget, *Kulturvetenskaperna i framtiden*: 1) förslag om skapandet av ett storprogram inom området språk, kommunikation och teknologi; 2) tanken på samverkan mellan HSFR och STU
* Mars 1987	SKOT-rapporten[2]
* Bå 1987/88	HSFR tillförs ytterligare medel för ”forskning om datateknikens användning” och samtidigt utvidgas området till att omfatta ”forskning om förutsättningar för och konsekvenser av informationsteknologins användning”
	HSFR avsätter planeringsmedel för forskning inom SKOT-området
* Bå 1988/89	*SKOT-programmet* startas
* Bå 1989/90	Fyra professurer med anknytning till området utlyses och tillsätts av regeringen: Logik, Psykologi, Lingvistik, och Kommunikation
	SKOT-programmet döps om till *Informationsteknologi och SKOT-programmet*
* Bå 1990/91	Språkdelen ”lyfts av” Informationsteknologiprogrammet och blir separat program med samfinansiering mellan HSFR och STU/NUTEK: *Språkteknologi* (Fas 1: 90/91-92/93 inleds)
	HSFR fortsätter sitt programstöd under rubriken ”Forskning om förutsättningar för och konsekvenser av informationsteknologins användning”
* Bå 1993/94	*Språkteknologi* (Fas 2: 93/94-95/96 inleds)
* Bå 1994/95	Sista år för HSFR:s IT-program
* Bå 1996/97	*Språkteknologi* (Fas 3: 96/97-98/99 inleds)

[1] Sammanställning baserad på uppgifter från Barbro Hänström, VR, tidigare handläggare av språkteknologiprogrammen inom HSFR:s beredningsgrupp för språkvetenskap
[2] SKOT = Språk, kommunikation och teknologi

Graduate School of Language Technology (GSLT)

The Graduate School of Language Technology (GSLT) is a national graduate school for which Göteborg University (Faculty of Arts) is the coordinating host. General information about the school and its original programme statement can be found on its web page `http://www.gslt.hum.gu.se`.

Students may currently be registered at any of the following academic institutions:

University College of Borås
Chalmers University of Technology
Göteborg University
KTH (Royal Institute of Technology)
Linköping University
Lund University
University of Skövde
Stockholm University
Uppsala University
Växjö University

Supervision is also available from SICS (Swedish Institute of Computer Science). We are still open for the addition of further institutions in the future if this should become appropriate.

The school currently has over forty graduate students.

The school also maintains a collection of databases and also provides computational resources on its server nationally for database processing and computationally intensive tasks. This is used on a daily basis by the school's graduates and other researchers associated with the school.

Current projects

The following is a selection of current PhD thesis projects which require access to large databases and database tools:

Atelach Alemu Argaw, Department of Computer and Systems Sciences, Stockholm University

Previous degree: M.Sc. in Information Science
Thesis topic: Cross Language Information Retrieval

Karin Cavallin, Department of Linguistics, Göteborg University
Previous degree: M.A. in Computational Linguistics, Göteborg University, 2003
Thesis topic: Aspects of the existential constructions in the Nordic languages

Loredana Cerrato, Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm
Previous degree: Foreign language, literatures and linguistics at University of Naples, 1992.
Thesis topic: Verbal and non-verbal dialogue behaviour across modalities and channels

Susanne Ekeklint, School of Mathematics and Systems Engineering, Växjö University
Previous degree: M.A. in Computational Linguistics, Göteborg University, 2001
Thesis topic: Automatic labeling of semantic entities in natural language based on dependency relations

Eva Forsbom, Department of Linguistics and Philology, Uppsala University
Previous degree: B.A. in English and Swedish, Stockholm University, 1990; University Diploma in Language Consultancy, Stockholm University, 1996
Thesis topic: Textlinguistic methods in summarisation and information access

Mikael Gunnarsson, Swedish School of Library and Information Sciences, University College of Borås
Previous degree: Librarianship diploma
Thesis topic: Genre Identification

Ebba Gustavii, Department of Linguistics and Philology, Uppsala University
Previous degree: Master of Philosphy in Language Engineering, 2003
Thesis topic: Automatic translation of productively formed lexical units

Harald Hammarström, Language Technology Group, Deptartment of Computing Science,

Chalmers University of Technology
Previous degree: MA in Computer Science, Uppsala University, 2003
Thesis topic: Unsupervised Induction of Morphology: Input a corpus of an arbitrary natural language and output a description of its morphology

Cecilia Hemming, Department of language, University College of Skövde (Department of Linguistics, Göteborg University).
Previous degree: B.A. French and Linguistics, Skövde
Thesis topic: Morphology and semantics of compounds and multi-word terms used to designate technical items

Hans Hjelm CL-Group, Deptartment of Linguistics,
Stockholm university
Previous degree: M.S. in Computational Linguistics, Gothenburg university, 2001
Thesis topic: Ontology learning from parallel texts

Per-Anders Jande, Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm
Previous degrees: B.A. Phonetics, B.A. General Linguistics, B.A. Computational Linguistics, M.A. Computational Linguistics
Thesis topic: Pronunciation variation modelling for Swedish

Rebecca Jonson, Department of Linguistics, Göteborg University
Previous degree: M.A. in Computational Linguistics, Göteborg University, 2000
Thesis topic: Using dialogue management to improve speech recognition

Monica Lassi, Swedish School of Library and Information Sciences University College of Borås
Previous degree: M.Sc. Library and Information Science, 2001
Thesis topic: Automatic thesaurus construction for interdisciplinary research and communication

Svetoslav Marinov, School of Humanities and Informatics, University College Skövde
Previous degree: M.A in Linguistics, University of Tromsø, Norway
Thesis topic: Dependency parsing for Bulgarian

Kristina Nilsson Computational Linguistics Group, Department of Linguistics, Stockholm University
Previous degree: 2003 M.A. in Computational Linguistics, Uppsala University
Thesis topic: Coreference Resolution for Information Access

Fredrik Olsson, Department of Swedish Language, Göteborg University and SICS.
Previous degree: Ph.lic. in Computational Linguistics, Department of Linguistics, Uppsala University.
Thesis topic: Portability Issues in Information Refinement

Susanne Schötz, Linguistics and Phonetics, Centre for Languages and Literature, Lund University
Previous degree: M.A. in Phonetics, 2001
Thesis topic: Perception, analysis and synthesis of Speaker Age

Per Weijnitz, Department of Linguistics, Uppsala University
Previous degree: M.A. of Philosophy in Language Engineering
Thesis topic: Hybrid methods in machine translation

Marcus Uneson, Linguistics and Phonetics, Centre for Languages and Literature, Lund University
Previous degree: B.A. in Musicology, M.A. in Phonetics
Thesis topic: Data-driven induction of phonological rules

Kenneth Wilhelmsson, Department of Linguistics, Göteborg University
Previous degree: M.A. in Computational Linguistics, Göteborg University, 2002
Thesis topic: Identification of Functional Roles in the Main Clause

Lilja Øvrelid, NLP-Unit, Deptartment of Swedish,
Göteborg University
Previous degree: B.A. in Language, Logic and Information (major), Linguistics (minor) and English (minor)
M.A. in Language, Logic and Information, Oslo University, Norway, 2003; *Thesis topic:* Disambiguating animacy: Acquisition of animacy information for syntactic disambiguation

Future plans and needs

The school plans to maintain a collection of databases containing language data available under license and make these available to students and associates of the school nationally. Examples of such databases are those available from the Linguistic Data Consortium (`http://www.ldc.upenn.edu`) at the University of Pennsylvania and the European Language Resources Association (`http://www.elra.info/`).

Databases would be used nationally on a daily basis by our doctoral students and associated researchers.

Approximate costs for the purchase of licenses and maintenance of computer equipment is 80kSEK/year, i.e. 4mSEK over a five year period.

The school has an income of 12mSEK/year up to 2012. However, these funds are intended for the support of doctoral students, teaching and administration and we are not able to support a full programme of data distribution and collection with these funds. We currently only have a small collection of databases from LDC and none from ELRA.

There is a great need to create Swedish databases corresponding to those existing for other languages. This need becomes particularly apparent in the context of the graduate school where one sees the needs of virtually all current graduate students in language technology in Sweden. While it is not the task of the graduate school to create such databases but rather of the various research centres which contribute to the school, it is clear that the creation and maintenance of these databases is a prerequisite for maintaining an internationally competetive graduate programme in this field.

Standarder och pågående standardiseringsarbeten: text- och taldatabaser*

Uppgifter från Lars Borin, Språkdata/Språkbanken:

Egentligen håller standardiseringsarbetet på det här området på att komma igång på allvar först nu. Det finns nu en ISO-kommitté som arbetar med standarder för språkresurser; mer information här:
<http://www.tc37sc4.org/>.

Den har förstås inte uppkommit ur intet, utan är resultatet av flera tidigare initiativ där man har definierat format för språkresurser:

Lagringsformat för texter
(1) Text Encoding Initiative <www.tei-c.org>
(2) EAGLES Corpus Encoding Standard (CES) i det nya XML-baserade formatet: XCES <www.xces.org>

Format för uppmärkning och metadata
(3) EAGLES taggformat för morfosyntaktisk uppmärkning
<http://www.ilc.cnr.it/EAGLES96/browse.html>
(4) ISLE/IMDI metadata för språkresurser <http://www.mpi.nl/IMDI/>
(5) OLAC (Open Language Archives Community) metadata för språkresurser
<http://www.language-archives.org/>

Format för lexikonresurser
(6) EAGLES lexikonformat <http://www.ilc.cnr.it/EAGLES96/browse.html>
(7) ISLE/MILE-formatet för flerspråkiga lexikon

Det finns en stark rörelse för att åstadkomma (större) standardisering av resurserna inom språkteknologiområdet. Samtidigt så kommer det faktiskt till många resurser som inte är standardiserade. Det beror nog till en del på att arbetet med att definiera standarder pågår som bäst och att medvetenheten om standarder inte har trängt in överallt, men det finns säkert också andra faktorer som spelar in. I ett läge där standarder inte har satt sig ordentligt (så att det t.ex. inte är så att relevant mjukvara i allmänhet automatiskt hanterar standardformat) innebär det ofta mer arbete att följa standarder än att inte göra det, vilket kan ha stor betydelse givet en ordinär forskningsprojektbudget.

Arbetet med standarder är en mycket viktig del av byggandet av en svensk infrastruktur av språkresurser för språkteknologi och språkvetenskap.

* Lars Borin, Språkdata/Språkbanken och Kjell Elenius, CTT, KTH, har var för sig ombetts lämna uppgifter om standardisering. Viss överlappning finns därför mellan de två uppgiftslämnarna.

Uppgifter från Kjell Elenius, CTT, KTH:

Här beskrivs en del av de standarder som finns för taldatabaser. De har alla tagits fram under de senaste 20 åren. Listan är långtifrån heltäckande, men torde ändå ge en bra belysning av området.

I EU-projektet SAM, Speech Assessment Methods, http://www.phon.ucl.ac.uk/resource/eurom1/, som löpte under slutet av 1980-talet tog man fram en standard för taldatabaser som använts i flera olika sammanhang. De båda stora EU-projekten SpeechDat och SpeeCon använde sig båda av SAM-formatet. En huvudprincip är att man skiljer på själva talsignalen och informationen om densamma. Man har alltså en talfil med enbart taldata (rådata) och en separat, associerad fil med information om talfilen, till exempel: talare, ålder, dialekt, transkription med mera. För uppmärkning, eller annotering, av själva talet på fonemnivå använder man sig av SAMPA, Speech Assessment Methods Phonetic Alphabet, som är en maskinläsbar version av det internationella fonetiska alfabetet IPA. Även SAMPA-standarden togs fram i SAM-projektet.

I USA använder sig NIST, The National Institute of Standards and Technology, av ett standardiserat talfilsformat NIST SPHERE, där filhuvudet innehåller en beskrivning av filen i vanligt textformat, normalt 1024 byte lång. Därefter följer talsamplen i binär form (rådata).

I mitten på 90-talet arbetade man i EU-projektet EAGLES, http://www.ilc.cnr.it/EAGLES/home.html, med att öka användbarheten av text och taldatabaser genom att försöka introducera allmänna standarder. Man försökte även standardisera utvärderingen av talteknologier som talsyntes, taligenkänning och talarverifiering. En bok som kan ses som slutresultatet av projektet är "Handbook of Standards and Resources for Spoken Language Systems" med redaktörerna Gibbon, Moore, och Winski.

Det finns även en internationell organisation som har fokus på taldata, COCOSDA, The International Committee for the Co-ordination and Standardisation of Speech Databases and Assesment Techniques for Speech Input/Output, http://www.cocosda.org/.

På senare tid har XML, Extensible Markup Language, alltmer utnyttjats för att strukturera och organisera information/data, till skillnad från till exempel HTML som används till att åskådliggöra och visa data. För text och tal finns till exempel XCES, Corpus Encoding Standard for XML, http://www.xces.org/.

En alternativ beskrivning av tal fås med, AGTK, Annotation Graph Toolkit. Annotationsgrafer kan här ses som ett formellt ramverk för att beskriva lingvistisk annotering av seriella data som talsignaler, http://agtk.sourceforge.net/. Styrkan är man kan beskriva samband över tid på ett mycket flexibelt sätt.

Ett närliggande område är olika initiativ för att standardisera användningen av tal på webben. För talsyntes har man tagit fram Speech Synthesis Markup Language (SSML), http://www.w3.org/TR/speech-synthesis/. För allmän talteknologi finns VoiceXML, Voice Extensible Markup Language, http://www.voicexml.org/, och Speech Application Language Tags, SALT, http://www.saltforum.org/saltforum/downloads/SALT1.0.pdf.

GÖTEBORGS UNIVERSITET
GSLT

GÖTEBORG UNIVERSITY Graduate School of Language Technology

28 januari 2004

Kulturminister Marita Ulvskog,
Näringsminister Leif Pagrotsky,
Utbildningsminister Thomas Östros

103 33 STOCKHOLM

Behovet av en sammanhållen strategi för svensk språkteknologi och ökade forsknings- och utvecklingsinsatser har påtalats i flera sammanhang både av det språkteknologiska forskarsamhället och i de utredningar som berört språkteknologiområdet, såsom Nuteks (1999) om maskinöversättning och Mål i mun-utredningen (SOU 2002: 27). Av skäl som sammanfattas i bifogade skrivelse är det viktigt att en sådan nu etableras och att Sverige förvaltar och förstärker de framgångsrika resultat som uppnåtts. Vi anhåller med detta brev också om att få träffa representanter för berörda departement och diskutera behov och åtgärder för att utveckla svensk språkteknologi.

Lars Ahrenberg
Professor, Linköpings universitet
Biträdande föreståndare, GSLT

Robin Cooper,
Professor, Föreståndare,
Sveriges nationella forskarskola
i språkteknologi GSLT

Olle Josephson
Föreståndare, Svenska Språknämnden

Anna Sågvall Hein,
Professor, Ordförande i det tidigare
programrådet för VINNOVAs program
i språkteknologi

Bengt Waernulf
Styrelseordförande, Centrum för talteknologi, CTT

Sverige behöver en strategi för språkteknologi

Språkteknologi är ett tvärvetenskapligt forskningsområde av avgörande betydelse för utvecklingen av de informations- och kommunikationstjänster som utgör grunden för visionerna om 24-timmarsmyndigheten, ett informationssamhälle för alla, och den semantiska webben. Exempel på språkteknologiska insatsområden är interaktion mellan människan och datorn i tal och skrift, informationssökning och informationsfiltrering, hjälpmedel för kommunikationshandikappade, språkundervisning samt maskinöversättning, Det är också, som utredningen SOU 2002:27 Mål i mun visar, ett område med betydande kulturella och språkpolitiska implikationer eftersom de tjänster som kan erbjudas står i proportion till de resurser i form av kunskap, utbildade människor, databaser och programvara som en språkgemenskap förfogar över. Av dessa skäl är det oerhört väsentligt att Sverige har en långsiktig strategi för språkteknologi med stöd i Sveriges riksdag.

Behovet av en sammanhållen strategi för svensk språkteknologi har påtalats i flera sammanhang både av det språkteknologiska forskarsamhället och i de utredningar som berört språkteknologiområdet, såsom Nuteks (1999) om maskinöversättning och Mål i mun-utredningen. Den senare anvisar också förslag på hur en sådan strategi skulle kunna realiseras, bland annat genom inrättandet av ett statligt finansierat språkteknologiskt sekretariat med samordnande funktion. De senaste årens utveckling inom svensk forskningsfinansiering och internationell forskning och utveckling inom språkteknologin gör behovet än mer angeläget.

Svenska forskningsfinansiärer har tagit flera initiativ som främjat utvecklingen av svensk språkteknologi. Vi kan här nämna de olika forskningsprogram som genomfördes under 90-talet av HSFR och Nutek i samarbete. 1996 startades Centrum för talteknologi (CTT) som en av VINNOVAs 10-åriga satsningar på kompetenscentra. CTT, med KTH som värd utgör ett framgångsrikt initiativ för samarbete mellan akademisk forskning och industriell utveckling. Sveriges nationella forskarskola i språkteknologi, GSLT, startades 2001 som en av sexton nya forskarskolor med statlig finansiering. GSLT med Göteborgs universitet som värdhögskola har samlat i stort sett alla lärosäten i Sverige som har forskarutbildning inom språkteknologi, både humanistiska och tekniska, i en gemensam utbildningsinsats och därtill Växjö universitet och högskolorna i Skövde och Borås som tidigare inte haft forskarutbildning inom området.

Flera av de framgångsrika satsningarna på slutet av 90-talet är snart avslutade. VINNOVAs kompetenscentrum CTT avslutas 2006. VINNOVAs pågående språkteknologiska program upphör med utgången av 2004 och fortsättningen är högst oviss. VINNOVA har för femårsperioden 03-07 definierat ett antal tillväxtområden och kunskapsplattformar inom vilka satsningar skall ske. Språkteknologi omnämns som en del av tillväxtområdet Programvaruprodukter men väntas inte bli prioriterat. Vetenskapsrådet finansierar för närvarande ett mindre antal begränsade projekt. Vad gäller GSLT finns inga beslut för tiden efter 2007. Den fortsatta forskningsfinansieringen inger således oro och en strategi behövs för att fullfölja de viktiga satsningar som hittills gjorts.

Dagens språkteknologi kräver också stora satsningar, inte minst i uppbyggnaden av den nödvändiga infrastrukturen: text och tal som är uppmärkt med nödvändig språklig information, från det fonetiska till det semantiska, analysatorer för att åstadkomma detta och utvärderingsmetoder och verktyg för att bedöma systems prestanda ur olika perspektiv. Ett stort kliv i utvecklingen tas för de stora språken när man nu bygger upp sådana resurser. Den franska satsningen Technolangue omfattar exempelvis drygt 20 miljoner euro per år i tre år. Även Norge genomför en stor satsning inom språkteknologi med ett program som omfattar 66 miljoner NOK under fem år.

Sverige har hittills saknat ett permanent sammanhållande organ för språkteknologi. Ett språkteknologiskt sekretariat kan öka samarbetet i Sverige mellan näringsliv, forskning, högre utbildning och myndigheter, agera för svensk språkteknologi och ge Sverige en högre profil på

området internationellt. Sverige har i EU-sammanhang varit framgångsrikt när det gäller deltagande i språkteknologiska forskningsprojekt men haft svårare att hävda sig i de större samverkansprogrammen som exempelvis MLIS-programmet (Multi-Lingual Information Society). Vi har alltså inte kunnat påverka programmen särskilt mycket eller rätt kunnat utnyttja de resurser som ställts till förfogande. Sverige riskerar i dag att hamna utanför pågående europeiska initiativ som bland annat gäller etablerandet av ett språkteknologiskt nätverk som ett ERA-nät (European Research Area net) för att stärka samarbete och samverkan mellan nationella europeiska forskningsfinansieringsorgan.

Element i en strategi för svensk språkteknologi

Enligt vår mening måste en språkteknologisk strategi eller plattform vila på fem hörnpelare:

Forskning och utveckling. Flera tidsbegränsade svenska satsningar har gjorts. Nu behövs en rejäl och långsiktig satsning för att ta vara på och fullt ut utnyttja den kompetens som byggts upp i Sverige och som ytterligare kommer att förstärkas med de doktorer som blir färdiga under de närmaste åren. Detta gäller såväl grundforskning som behovsmotiverad forskning.

Grund- och forskarutbildning. Sverige är i dag förhållandevis väl försett både vad gäller grundutbildning och forskarutbildning på området. Utan ett långsiktigt perspektiv riskerar dock det nationella engagemang och den dynamik som skapats i forskarutbildningen genom tillkomsten av GSLT att gå förlorad.

Kommersialisering och kunskapsöverföring. Kommersialisering av språkteknologiska forskningsresultat hör samman med utvecklingen av å ena sidan generella IKT-tjänster och å andra sidan förändring av språkligt inriktade arbeten som traditionellt utförs av personer med en humanistisk utbildning såsom översättning, informationssökning, terminologi, lexikografi och språkutbildning. Forskningen inom CTT har skapat ett produktivt kunskapsflöde mellan akademi och industri. Enligt en nyutkommen studie: ”Benchmarking HLT progress in Europe” (HLT = Human Language Technologies) framtagen i ett projekt inom EU:s IST-program, har Sverige ”the highest HLT Opportunity score in the EU, due to its highly competitive knowledge society infrastructure”. Att utnyttja denna möjlighet kräver en fortsatt industriell och akademisk satsning.

Språkbanker, i första hand för svenska, men även för andra för Sverige viktiga språk, som är allmänt tillgängliga. En språkbank är en stor och växande samling av språkliga material, både text och tal, i ren och i bearbetad form som exempelvis lexikon, termbanker, trädbanker, språkligt uppmärkta tal-, dialog- och textdata.

Ett språkteknologiskt sekretariat med en samordnande funktion som också kan agera för svensk språkteknologi i en nordisk och europeisk kontext.

Utformningen av en långsiktig svensk strategi för språkteknologi bör, menar vi, uppmärksammas i den forskningspolitiska propositionen liksom i den proposition som ska följa på utredningen Mål i mun. De åtgärder som nu framstår som de mest angelägna är följande:

- Väsentliga långsiktiga förstärkningar till forskning, forskarutbildning och utveckling inom språkteknologi, häri inräknat fortsatta och ökade infrastrukturella satsningar på språkbanker.
- Skapandet av ett nationellt samordnande organ i form av ett språkteknologiskt sekretariat.