

Intresseanmälan om samverkan för ny europeisk forskningsinfrastruktur

CLARIN, ett ESFRI-infrastrukturprojekt för EU:s sjunde ramprogram

CLARIN <<http://www.mpi.nl/clarin/>> är ett europeiskt initiativ för att skapa en integrerad och standardiserad forskningsinfrastruktur för språkliga resurser, vilket inbegriper dels dataresurser (korpora, taldata, lexikon, grammatiker, etc.), dels de teknologier och verktyg som behövs för att lagra, distribuera och arbeta med dataresurserna (både primärdata och härledda data). Sådana språkliga resurser är en omistlig komponent i språkteknologisk forskning och utveckling samt inom den allmänna språkvetenskapens olika områden. Dessutom har de språkliga resurserna potentiellt en framträdande roll att spela inom alla humanistiska och samhällsvetenskapliga discipliner där text och tal är viktiga studieobjekt.

I CLARIN-konsortiet ingår över 70 medlemsinstitutioner från 32 länder (se <<http://www.mpi.nl/clarin/fullMembers.htm>>). Sverige representeras av tre (inom kort fyra) medlemmar: Göteborgs universitet/Språkbanken, KTH/Institutionen för tal, musik och hörsel, Lunds universitet/Humanistlaboratoriet och snart också Uppsala universitet/Institutionen för lingvistik och filologi. CLARIN samordnas av Steven Krauwer (Universitetet i Utrecht), Tamás Varadi (Ungerska vetenskapsakademien), Martin Wynne (Oxford Text Archive) och Peter Wittenburg (Max-Planck-sällskapet).

Förutom dataresurser samt teknologier och verktyg i snävare bemärkelse, kommer CLARIN att arbeta med två områden som utgör grundförutsättningar för distribution och delande av språkliga resurser, nämligen standardiserade metadata och hantering av immaterialrättsliga frågor.

Tidsplanen för CLARIN spänner över 6 år (2007–2012), inklusive en inledande förberedelsefas om ett år, med en uppskattad total budget om c:a 108 miljoner euro.

Inom CLARIN finns följande arbetsgrupper:

<i>Arbetsgrupp</i>	<i>Koordinatorinstitution(er)</i>
Språkliga resurser	CST, Köpenhamn och IIPAN, Warszawa
Språkteknologi (för skrivet språk)	Ungerska vetenskapsakademien och ILSP, Aten
Talteknologi	Universitetet i Nijmegen
Standarder	ILC, Pisa och Universitetet i Wien
Grid-teknologi och distribuerade tjänster	MPI Nijmegen
Immaterialrättsliga frågor	ELDA, Paris
Utbildning och informationsspridning	Universitetet i Utrecht
Avancerade användarfall	Lunds universitet

SNK/BLARK-konsortiet: För svenska språkliga resurser

Intresset i Sverige för CLARIN är mycket stort. Som redan nämnts har CLARIN fyra svenska medlemmar. Dessa är alla även medlemmar av ett svenskt konsortium för skapandet av svenska språkliga resurser av den typ som CLARIN-arbetet syftar till. Ett förberedelsearbete för detta har kommit igång i Sverige i form av ett VR-stött planeringsprojekt ("En infrastruktur för svensk språkteknologi"; VR:s dnr 2006-6763), samordnat av Lars Borin vid Göteborgs universitet..

Planeringsprojektet syftar till skapandet av två överlappande resurser till stöd för svensk språkteknologiforskning. Språkteknologi är ett samlingsnamn för sådan informations- och kommunikationsteknologi (IKT) som låter datorer hantera mänskligt språk i alla dess former – tal, skrift och teckenspråk. Språkteknologi är ett starkt tvärvetenskapligt forskningsområde

som är relevant överallt där människor interagerar med datorer och faktiskt även vid interaktion människor emellan, i form av olika sorters kommunikationshjälpmedel. I skönlitteratur och film pratar människor med intelligenta datorer som naturligtvis även förstår gester och kroppsspråk. Det är helt klart att språkanvändande datorer kommer att förändra vår vardag enormt, av den enkla anledningen att datorer och människor är bra på olika saker. Datorer byggs in i fler och fler apparater, och språkteknologi behövs för att möjliggöra interaktion med dessa allt mer avancerade tekniska produkter till exempel i våra hem, på vägarna och på arbetet. Den mesta informationen i våra IT-system uttrycks dessutom fortfarande i något mänskligt språk (och dessutom på allt fler språk).

Den vanligaste användningen för datorer idag är förmodligen för informationshantering: att skapa, läsa, ordna eller söka information i form av text, ljud eller video. Vi behöver språkteknologi för att vi inte ska drunkna i all denna information och som hjälp för att skapa den, översätta den, läsa den och alltmer för att hämta relevanta delar av dokument (gärna i sammanfattad form) utan att läsa dem. Särskilt behöver vi hjälp för att hantera språk som vi inte kan så bra eller som är besvärliga att använda på grund av funktionshinder eller den aktuella situationen (synskadade eller bilförare, t.ex.).

I den statliga utredningen om svenska språkets ställning (*Mål i mun*) understryks att språkteknologi har vittgående betydelse för svenskans framtid som fullödigt språk. Informations-samhället avancerar på bred front, och utan språkteknologi för ett språk kan man inte räkna med att upprätthålla önskvärd tillgång till digital information eller digitala tjänster på det språket. En satsning som 24-timmarsmyndigheten kan knappast förverkligas utan språkteknologi. Här blir flerspråkiga lösningar viktiga eftersom man vill kunna hantera så många som möjligt av Sveriges språk.

Språkteknologi har både språkoberoende och språkberoende aspekter. Detta betyder att resultat som kommer ur språkteknologisk forskning om svenska och andra språk i Sverige är högst relevanta för den internationella forskargemenskapen, men också att språkteknologi för svenska (eller svenska i kontrast mot andra språk, t.ex. för översättningstillämpningar) inte kommer till utan vidare; den måste skapas i Sverige.

Den språkteknologiska forskningen och utvecklingen av språkteknologisystem behöver som redan nämnts ovan en infrastruktur av allmänt tillgängliga och standardiserade basresurser, både data och program för att arbeta med dessa data (en grunduppsättning sådana resurser kallas med en engelsk förkortning för BLARK – Basic LAnguage Resource Kit). Sådana resurser måste skapas för varje språk för sig, men detta görs med mycket stor fördel i internationellt samarbete som det i CLARIN-projektet, just därför att det finns betydande språkoberoende aspekter av sådana resurser och deras användning (format-, gränssnitts- och verktygsstandarder, metadata, immaterialrättsliga och etiska frågor, etc.). För svenskans del finns en del resurser redan (i stor utsträckning skapade av de forskargrupper som ingår i SNK/BLARK-konsortiet), men det är oklart hur mycket och hur tillgängliga de är; en noggrann inventering behövs (en utmärkt bas är VR/DISC:s rapport; se nedan). Klart är däremot att flera grundläggande resurser inte finns allmänt tillgängliga för svenska och än mindre för många av Sveriges andra språk, t.ex. lexikonresurser med information om ords böjning och betydelser och god täckning (åtminstone 50.000 uppslagsord), stora databaser med talspråk och stora textdatabaser – tal- och textkorpusar – med en sammansättning som motsvarar det talade eller skrivna språkets genrefördelning och variation och som är försedda med rik språklig information om t.ex. ordklasser och satsanalyser.

Vi behöver både enspråkiga och flerspråkiga textkorpusar som avspeglar det faktum att svenskan är ett standardspråk med en lång skrift- och språkvårdstradition, men som lever och verkar i en vardag med dubbelriktad flerspråkighet: utåt mot de nordiska språken samt engelska och andra världsspråk, och inåt mot landets minoritets- och invandrarspråk. En

modern svensk språkteknologisk infrastruktur har i vår vision två överlappande och samverkande komponenter, en BLARK och en svensk nationell korpus (SNK). Arbetet med att ta fram dessa planeras ske i tre steg – (1) definition av resursbehov (BLARK- och SNK-komponenter) givet nuvarande och förutsedd svensk språkteknologiforskning; (2) inventering av befintliga resurser i relation till behoven samt kostnadsberäkning; (3) skapande av resurserna – varav de första två stegen hör till det beviljade planeringsprojektet (här kan projektet dra stor nytta av att VR/DISC redan har tagit initiativet till ett inventeringsarbete som resulterat i professor Eva Strangerts omfattande och värdefulla rapport *Svensk språkteknologi - existerande forskningsinfrastruktur och framtida behov* [februari 2007]), och det tredje med fördel kan vara en del av CLARIN.

Forskare/forskargrupper som står bakom intresseanmälan

Denna intresseanmälan har formulerats i samarbete mellan tre av medlemmarna i SNK/BLARK-konsortiet, nämligen:

- Göteborgs universitet/Chalmers, Centre for Language Technology (samarbete mellan GU, Hum. fak. [Lingvistik, Svenska språket, Filosofi] och GU/Chalmers, [IT-universitetet/Datavetenskap]), professor Lars Borin
- KTH, Skolan för datavetenskap och kommunikation, Institutionen för tal, musik och hörsel, professor Rolf Carlson
- Uppsala universitet, Institutionen för lingvistik och filologi, avdelningen för datorlingvistik, professor Joakim Nivre

I konsortiet ingår dessutom Lunds universitet, Språk- och litteraturcentrum, och GSLT (Graduate School of Language Technology), den svenska nationella forskarskolan i språkteknologi, som är ett samarbete mellan Chalmers, Göteborgs universitet, Högskolan i Borås, Högskolan i Skövde, KTH, Linköpings universitet, Lunds universitet, Stockholms universitet, Uppsala universitet och Växjö universitet, med medverkan av SICS AB. Genom GSLT representeras hela den svenska språkteknologiforskningsgemenskapen i konsortiet.

Namn och kontaktuppgifter till kontaktperson i Sverige

Lars Borin
Göteborgs universitet
Institutionen för svenska språket
Box 200
405 30 Göteborg

tel. 031 786 4544, 070 747 8386
fax 031 786 4455
e-post <lars.borin@svenska.gu.se>