Språkteknologisk infrastruktur i Norden

Rapport med artikelbidrag

Seminarium om språkteknologisk infrastruktur i Norden
Konferenscentrum Wallenberg, Göteborg
26 oktober 2006

Arbetsgruppen för språkvård och språkteknologi i Norden
Nordens språkråd
16 februari 2007

Rapporten är sammanställd av Rickard Domeij
Arbetsgruppen för språkvård och språkteknologi i Norden
Nordens språkråd
16 februari 2007
E-post: rickard.domeij@sprakradet.se

INNEHÅLL

Om seminariet:

Artiklar:

Bilder från presentationerna (i särskild bilaga):

Deltagarlista

Namn	**E-post**	**Organisation**
Ahrenberg, Lars	lah@ida.liu.se	Linköpings univ
Borin, Lars	lars.borin@svenska.gu.se	Språkdata, Göteborgs univ
Braasch, Anna	anna@cst.dk	CST
Brøndsted, Tom	tb@kom.auc.dk	Ålborg univ
Carlson, Rolf	rolf@speech.kth.se	KTH
Cooper, Robin	cooper@ling.gu.se	Göteborgs univ
Dalianis, Hercules	hercules@kth.se	Stockholms univ/KTH
de Smedt, Koenraad	desmedt@uib.no	Univ i Bergen
Domeij, Rickard	Rickard.domeij@spraknamnden.se	Språkrådet
Dura, Elzbieta	elzbieta@lexwarelabs.com	Högsk. i Skövde; Lexware labs
Forsbom, Eva	evafo@stp.lingfil.uu.se	Uppsala univ
Gíslason, Hjálmar	hjalmarg@siminn.is	Já
Helgadóttir, Sigrún	sigrunh@lexis.hi.is	Árni Magnússon instituttet
Henrichsen, Peter Juel	pjh.id@cbs.dk	Copenhagen Business School
Horne, Merle	Merle.Horne@ling.lu.se	Lunds univ; Vetenskapsrådet KFI
Karlsson, Ola	Ola.karlsson@spraknamnden.se	Språkrådet
Kirchmeier-Andersen, Sabine	sabine@dsn.dk	Dansk sprognævn
Koskenniemi, Kimmo	kimmo.koskenniemi@helsinki.fi	Helsingfors univ
Larsson, Lars-Erik	lars-erik.larsson@acapela-group.com	Acapela group
Lindén, Krister	krister.linden@helsinki.fi	Helsingfors univ
Loftsson, Hrafn	HRAFN@ru.is	Reykjavík univ
Lounela, Mikko	mikko.lounela@kotus.fi	Forskningscentralen
Maegaard, Bente	bente@cst.dk	CST, Köpenhamns univ
Moshagen, Sjur	sjur.moshagen@samediggi.no	Sametinget i Norge
Nordgård, Torbjörn	torbjorn@hf.ntnu.no	NTU
Nordström, Bengt	bengt@cs.chalmers.se	Chalmers tekniska högsk
Paile, Alexander	alexander.paile@lingsoft.fi	Forskningscentralen
Petersen, Peter Erik	peter.erik.petersen@maxmanus.no	Max Manus
Rasmusen, Jens Erik	jer@mikrov.dk	Mikroverkstædet
Reuter, Mikael	reuter@focis.fi	Forskningscentralen
Ronkainen, Otto-Ville	ronkaine@stoker.lingsoft.fi	Lingsoft
Rögnvaldsson, Eiríkur	eirikur@hi.is	Islands universitet
Strangert, Eva	Eva.Strangert@ling.umu.se	Vetenskapsrådet, Disc
Svavarsdóttir, Ásta	asta@lexis.hi.is	Leksikografisk institut
Sågvall-Hein, Anna	Anna.sagvall_hein@lingfil.uu.se	Uppsala univ
Volk, Martin	volk@ling.su.se	Stockholms univ

SEMINARIUM OM SPRÅKTEKNOLOGISK INFRASTRUKTUR I NORDEN

I oktober 2006 hölls ett seminarium i Göteborg med temat Språkteknologisk infrastruktur i Norden. Seminariet behandlade möjligheterna att samarbeta för att ta fram nödvändiga resurser för utvecklingen av språkteknologi i Norden. Innehållet dokumenteras i den här rapporten.

Ämnet är aktuellt. Både EU och de nordiska länderna lägger upp planer för att bygga ut infrastrukturen för framtida forskning och utveckling i Europa. Nordens språkråd har nyligen låtit framställa en s.k. vismansrapport som föreslår stora satsningar på nordisk språkteknologi under de kommande 10 åren. Den viktigaste förutsättningen för en god utveckling av språkteknologisk forskning och utveckling är tillgången på kvalitetssäkrade språk- och teknikresurser, t.ex. i form av uppmärkta text- och taldatabaser med verktyg för att hantera dem. Problemet är att kostnaderna för att utveckla och underhålla sådana resurser är stora. Därför finns mycket att vinna på ett samarbete kring språkteknologisk infrastruktur i Norden. Men det finns också många frågor att besvara innan vi är där: Varför är språkteknologin viktig för det nordiska språkområdet? Vilka resurser finns i de nordiska länderna idag? Vilka resurser behövs i framtiden? Hur ska de tas fram och göras tillgängliga? Hur kan de nordiska länderna samarbeta om detta? Vad bör göras nationellt och vad bör göras gemensamt?

Tanken med seminariet var att samla ledande aktörer på området och diskutera sådana frågor mot bakgrund av vismansrapporten och det som händer på området i de nordiska länderna och inom EU. Forskare, utvecklare och andra nyckelpersoner från de nordiska länderna inbjöds att tala och delta i seminariet. Som ett resultat av detta hoppas vi kunna driva vismansrapportens förslag vidare och arbeta för att ytterligare ett steg på vägen tas mot ett nordiskt samarbete om språkteknologisk infrastruktur. I diskussionerna framkom följande åtgärder som särskilt angelägna att börja med:

- Att mer systematiskt undersöka behovet av språkteknologiska resurser i de nordiska länderna utifrån en inventering av befintliga resurser (en s.k. blark-undersökning), vilket ska resultera i en konkret plan (med prioriteringar och kostnadsberäkningar) för hur de nordiska länderna gemensamt kan ta fram nödvändiga resurser och göra dem tillgängliga för nordisk språkteknologi
- Att särskilt arbeta med lösningar på de upphovsrättsliga problem som försvårar förverkligandet av planen, bl.a. genom att ta upp diskussioner med författar- och upphovsrättsorganisationerna.

Seminariet arrangerades av Nordens språkråd genom Arbetsgruppen för språkvård och språkteknologi i Norden. Seminariet hölls den 26 okt 2006, kl 9.30-18 på konferens-centrum Wallenberg i Göteborg i anslutning till den första svenska språkteknologi-konferensen SLTC 2006.

LÄNKAR:

Vismansrapporten
http://www.ling.helsinki.fi/~klinden/pubs/Spr%E5kVisFullReport.pdf

Arbetsgruppen för språkvård och språkteknologi i Norden
http://www.sprakradet.se/asp

Nordens språkråd
http://www.norden.org/sprak/nordenssprakrad/sk/index.asp

SLTC 2006
http://www.ling.su.se/DaLi/SLTC06/index.htm

BAKGRUND OCH TEMAN

Arbetsgruppen för språkvård och språkteknologi i Norden är ett samarbete mellan språknämnderna i Norden med syfte att främja nordisk språkteknologi. Arbetet stöds av Nordens språkråd.

Bakgrunden är att språknämnderna i Norden de senaste åren har fått ett bredare verksamhetsområde och större språkpolitisk betydelse. T.ex. har Svenska språknämnden, numera Språkrådet, ombildats till myndighet med uttalad uppgift att ”främja språkteknologisk utveckling”. I propositionen *Bästa språket* står att myndigheten ”långsiktigt [ska] verka för att uppmärkta och representativa text- och taldatabaser utvecklas”. Där står också: ”Den nya språkvårdsorganisationen bör aktivt delta i det nordiska samarbetet och verka för att den nordiska språkgemenskapen stärks”. Liknande språkpolitiska dokument har tagits fram för andra språknämnder i Norden. Likaså är betydelsen av språkteknologisk utveckling och nordiskt språksamarbete något som uppmärksammats i *Deklaration om nordisk språkpolitik*, 2006[1].

Våren 2005 arrangerades ett nordiskt seminarium i Pargas om språkkontroll. Det fick bl.a. resultatet att arbetsgruppen bildades och att Nordens språkråd lät ta fram en vismansrapport med en tioårsplan för att utveckla språkteknologin i Norden. Seminariet ledde också till ett ökat samarbete mellan språkvårdare och språkteknologiföretag.

Det aktuella infrastrukturseminariet var det andra i ordningen. Syftet var att diskutera hur vi kan samarbeta för att ta fram nödvändiga språk- och teknikresurser och ställa dem till förfogande för språkteknologisk forskning och utveckling i de nordiska länderna. I vismansrapporten, som nyligen presenterats för Nordens språkråd, finns flera förslag att ta ställning till, varför det var naturligt att utgå från den. En stor del av dagen ägnades därför åt den. Seminariet avslutades med diskussioner kring rapportens förslag.

Programmet var indelat i fyra delteman med presentationer som gav bakgrund till diskussionen:

- Tema 1: Vad finns och vad händer?
- Tema 2: Vad saknas? Språkteknologins behov av resurser
- Tema 3: Vilka är problemen? Vilka hinder måste vi ta oss över?
- Tema 4: Hur ska vi gå vidare och hur samarbeta?

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

PRESENTATIONERNA

Här beskrivs kort de presentationer som hölls vid seminariet. Här nämns också de artiklar som ingår i rapporten. De fyra artiklarna, som finns att läsa på följande sidor, behandlar olika ämnen med anknytning till seminariets teman: 1. språkteknologins språkpolitiska betydelse 2. behovet av flerspråkiga resurser 3. upphovsrätt och 4. vismansrapportens förslag. För de presentationer som inte beskrivs i artiklarna finns utskrifter av Powerpoint-bilder i en särskild bilaga.

1. Vad finns och vad händer?

Rickard Domeij inledde kort med att berätta om språkteknologins betydelse i språkpolitiskt perspektiv. Artikeln *Språkpolitik och språkteknologi i Sverige och Norden* på nästa sida beskriver situationen i Sverige och i Norden[2].

Bente Maegaard från Köpenhamns universitet gav en översikt över den språkteknologiska infrastrukturen i Norden och Europa.

Lars Borin från Universitetet i Göteborg berättade om planerna på en svensk nationell korpus och en uppsättning grundläggande språkteknologiska verktyg och resurser, en s.k. blark (basic language resource kit).

2. Vad saknas? Språkteknologins behov av resurser

Peter Erik Petersen från företaget Max Manus i Norge berättade om talteknologins behov av resurser.

Koenraad de Smedt från Universitetet i Bergen redogjorde för behovet av flerspråkiga resurser i Norden. Det beskrivs i den andra artikeln på följande sidor: *Flerspråklige ressurser – Utfordringer for Norden*.

Hjálmar Gíslason från företaget Já på Island visade vilka behov informationssökningen har av språkteknologiska resurser.

3. Vilka är problemen? Vilka hinder måste vi ta oss över?

Eva Strangert från Vetenskapsrådet i Sverige delade med sig av sina erfarenheter från en undersökning av infrastrukturbehovet för humanvetenskaplig och språkteknologisk forskning i Sverige.

Hercules Dalianis från Stockholms universitet berättade om sina erfarenheter från arbetet med en nordisk nätordbok och flerspråkig sökning.

Sigrún Helgadóttir från Árni Magnússon instituttet gjorde en genomgång av de upphovsrättsliga problem som kan uppstå vid insamling och tillgängliggörande av språkresurser. Det finns utförligt beskrivet i artikel 3: *Spørgsmål om ophavsret – den islandske erfaring*.

4. Hur ska vi gå vidare och samarbeta?

Kimmo Koskenniemi från Helsingfors universitet och en av vismännen presenterade vismansrapportens förslag på en nordisk satsning för att göra Norden till en ledande region inom språkteknologi. Presentationen följdes av diskussioner. En sammanfattning av vismansrapporten finns i artikel 4 som avslutar rapporten: *Språkvis – en språkteknologisk vismansrapport. Utvidgad sammanfattning.*

[2] Artikeln publiceras också i en rapport till Vetenskapsrådet i Sverige: *Svensk språkteknologi – existerande forskningsinfrastruktur och framtida behov.* Vetenskapsrådet, Disc. November 2007.

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>

Flerspråklige ressurser – Utfordringer for Norden

Koenraad de Smedt
Universitetet i Bergen, Norge

Språkteknologisk infrastruktur i Norden
Göteborg, 26. oktober 2006

Sammendrag

Norden er et høyteknologisk informasjonssamfunn med store kommunikasjonsbehov i en globalisert verden. De nordiske landene ønsker å ta vare på sine egne språk og kulturer, men trenger samtidig god kommunikasjon mellom de nordiske landene, med Europa, og med resten av verdenen. Innenfor hvert nordisk land er det også minoritetsspråk og innvandrerspråk. Skal man fortsatt ha aktivt bruk av de nordiske språkene i den sosiale, politiske og økonomiske sfæren, så trenger man ikke bare datastøttet språkprossessering for de enkelte språkene, men også på tvers av språk.

Den nordiske FoU-innsatsen for utvikling av store flerspråklige ressurser er lav, selv om kompetansen er høy. Dette sammendraget har ikke som mål å gi et fullstendig bilde av behovene, aktivitetene og forskningsresultatene. Noen eksempler på anvendelsesområder som krever flerspråklige ressurser er følgende:

- Andrespråkslæring: bedre læresystemer tilpasset språklig bakgrunn
- Tverrspråklig informasjonssøking: gi et søkeord eller dokument på ett språk, finn informasjon på et annet språk
- Webbaserte og mobile tjenester: flerspråklige grensesnitt og hjelp for fremmedspråklige
- Oversettelse: helautomatisk maskinoversettelse eller hjelpemidler til oversettere.

Anvendelser innen disse områdene forutsetter bruk av visse språkressurser. Énspråklige ressurser dekker kun en del av behovet. Behovet for spesifikke flerspråklige ressurser omfatter blant annet følgende:

- Oversettelsesordbøker og transferleksika
- Flerspråklige termlister, ontologier og ordnett
- Parallelle korpus inkl. parallelle trebanker og andrespråkskorpus

- Parallelle grammatikker.

De fleste flerspråklige forsknings- og utviklingsprosjekter trenger å utvikle sine egne ressurser. LEXIN (Ordbøker for minoritetsspråklige innvandrere) er et eksempel på et prosjekt som utvikler flerspråklige ressurser til et bestemt formål. I noen tilfeller kan ressurser likevel gjenbrukes til flere formål. Prosjektene TVÄRSØK og TVÄRSLÅ, som utvikler flerspråklige nettordbøker og tverrspråklig søking, gjenbruker ordlister fra LEXIN og andre kilder. Det norske prosjektet *Fra parallelle korpus til ordnett* bruker et eksisterende norsk-engelsk korpus. Det norsk-engelske maskinoversettelsesprosjektet LOGON gjenbruker eksisterende énspråklige grammatikker og leksikale ressurser, men måtte selv utvikle transferressurser.

Foruten anvendelsorienterte prosjekter finnes det prosjekter i Norden som utvikler flerspråklig basisteknologi. I Stockholm forskes det på parallelle trebanker, det vil si, parallellstilling på frasenivå. Slike parallelle trebanker kan være en viktig ressurs for maskinoversettelse og andre formål. Mens man i LOGON-prosjektet ekstraherer informasjon fra énspråklige trebanker for å optimalisere genereringsdelen, hadde en parallell trebank vært enda mer nyttig, både for maskinoversettelse og for kontrastive studier, fordi den synliggjør oversettelseskorrespondanser på alle syntaktiske nivåer.

Flerspråklige ressurser, for eksempel parallelle korpus, trebanker, ordnett, osv. som skal være brukbare til flere forsknings- og utviklingsformål bør tilfredsstille visse kvalitetskrav. Ideelt må ressursene være mangfoldige med hensyn til språk, materialutvalg og lingvistisk annotasjon. Ressursene bør være representative, ha en høy dekningsgrad, være kvalitetssikret og bruke internasjonale standarder for koding og annotasjon. Til slutt bør ressursene være tilstrekkelig dokumentert og rettighetene bør være såpass avklart at de ikke er til hinder for forskning og utvikling. Man bør også ta i betraktning CLARINs målkrav for distribusjon av språkressurser: *integrated, interoperable, stable, persistent, accessible, extendable.*

Flerspråklige ressurser og verktøy er i større grad enn énspråklige avhengig av internasjonal koordinering. Det nordiske forskningsprogrammet for språkteknologi (2000-2004) har hatt en positiv effekt på forskningssamarbeidet, men har hatt en begrenset omfang og varighet. Selv om dette forskningsprogrammet ikke har fokusert spesielt på flerspråklige ressurser, har det stimulert oppbygging av parallelle eller kompatible ressurser og verktøy for flere nordiske språk. Denne forskningen har vært noenlunde begrenset til de nordiske språkene, en satsing som har vært nødvendig, men som bør komplementeres med koblinger mellom de nordiske språkene og andre språk i verdenen, blant annet europeiske språk, minoritetspråk og store verdensspråk.

De nordiske landene har en høy kompetanse innen språkvitenskap og språkteknologi, men denne kunnskapen er spredt relativt tynt, slik at fortsatt samarbeid er viktig. Dessuten bør kunnskapen overføres og få et bredere grunnlag gjennom forskerutdanning og forskerutveksling. Gjennom Marie-Curieprosjektene BATMULT og MULTILINGUA har unge forskere fra Polen, Frankrike, Italia, Spania, Nederland, Tyskland, Finland, Slovenia og Romania de siste årene

kommet til Norge for forskningssamarbeid innen flerspråklig teknologi. I det 7. rammeprogrammet er det mulighet for de nordiske landene å utvide slikt samarbeid.

I konklusjon mener jeg at kommunasjonsbehovene i Norden på tvers av språk, både innenfor de nordisk landene, mellom de nordiske landene, og med Europa og verdenen, er betydelige og trenger datamaskinell støtte. Det har vært flere kvalitativt gode forskningsinitiativer om flerspråklig forskning i Norden, men dimensjoneringen av prosjektene har vært for liten i forhold til behovene. Språkressurser er svært viktige såkorn for språkvitenskapelig forskning og utvikling. Verdien til forskningen og produkter generert av tilgjengelige språkressurser er større enn investeringen til oppbygging av ressursene. En slik investering kan likevel ikke foretas av ett enkelt forskningsmiljø. Flerspråklige ressurser bør betraktes som en internasjonal infrastruktur for forskning og utvikling og oppbyggingen av denne infrastrukturen trenger internasjonal koordinering.

Spåkteknologisk infrastruktur
Nordiskt seminarium vid Wallenberg center i Göteborg
26. oktober 2006. Nordens språkråd

Sigrún Helgadóttir
Árni Magnússon instituttet for islandske studier – Leksikografisk afdeling[1]

Spørgsmål om ophavsret – den islandske erfaring

1. Introduktion

I dette foredrag vil jeg kort gennemgå lov om ophavsret, beskrive hvordan man har båret sig ad ved at oprette 5 forskellige korpusser og give en oversigt over det som allerede er blevet gjort i instituttet med hensyn til at sikre licens til brug af tekster i det islandske korpus.

Vi har allerede hørt fra Lars Borin at en del af de nødvendige ressourcer for sprogteknologien for et sprog er tekstkorpusser. Et tekstkorpus er en struktureret samling af tekstprøver i elektronisk form der dokumenterer et sprog sådan som det skrives i en bestemt tid. Tekstkorpusser må gøres tilgængelige for almenheden, især for dem som driver sprogforskning og fremstiller sprogteknologiske produkter. For at være i stand til samle sådanne tekstprøver er det nødvendigt at få adgang til tekster i elektonisk form og at fremskaffe licens fra indehavere af ophavsret til teksterne for at bruge dem på en bestemt måde. Udgangspunktet er selvfølgelig at man ikke kan lægge ophavsrettsbeskyttet materiale ud på internettet uden ophavsmandens samtykke.

Ophavsretslov i Island, Danmark, Sverige og Norge synes at være i overensstemmelse med hinanden, i det mindste med hensyn til det problem som bliver diskuteret i dette foredrag. Siden dette foredrag tænkes at være fremført på dansk vil jeg bruge den danske lovtekst til at vise det som man må tage hensyn til vedrørende tekstkorpusser.

I den første paragraf i loven om ophavsret siges der bl. andet:

> **§ 1.** Den, som frembringer et litterært eller kunstnerisk værk, *har ophavsret til værket*, hvad enten dette fremtræder som en i skrift eller tale udtrykt *skønlitterær eller faglitterær fremstilling*, som musikværk eller sceneværk, som filmværk eller fotografisk værk, som værk af billedkunst, bygningskunst eller brugskunst, eller det er kommet til udtryk på anden måde.

I lovens anden paragraf står der endvidere:

> **§ 2.** Ophavsretten medfører, med de i denne lov angivne indskrænkninger, *eneret* til at råde over værket ved at fremstille eksemplarer af det og ved at gøre det tilgængeligt for almenheden i oprindelig eller ændret skikkelse, i oversættelse, omarbejdelse i anden litteratur- eller kunstart eller i anden teknik.

Dette betyder at tekster, både skønlitterære og faglitterære, er beskyttet af ophavsret og man skal fremskaffe tilladelse fra ophavsmanden til at bruge teksterne i digitaliseret form i et tekstkorpus.

[1] *Árni Magnússon instituttet for islandske studier* blev oprettet 1. spetember 2006. 5 institutter blev lagt ind i det nye institut, et af dem er *Orðabók Háskólan*s (Leksikografisk institut) som nu er en leksikografisk afdeling i det nye institut. I det følgende vil jeg dog bruge „Ordbogen“ når jeg henviser til det „gamle“ institut.

I lovens syvende paragraf er der en formodning om ophavsrettens indehaver:

> **§ 7.** Som *ophavsmand* anses, når ikke andet er oplyst, den, hvis navn eller alment kendte pseudonym eller mærke på sædvanlig måde er påført eksemplarer af værket eller opgives, når det gøres tilgængeligt for almenheden.
>
> *Stk. 2.* Er et værk udgivet, uden at ophavsmanden er angivet i overensstemmelse med stk. 1, kan udgiveren, hvis denne er nævnt, og ellers forlæggeren handle på ophavsmandens vegne, indtil denne bliver angivet på et nyt oplag.

For at være i stand til að søge licens til at bruge tekst er det nødvendigt at vide hvem ophavsmanden er. Det er let når det drejer sig om bøger eller artikler i magasiner. Men det kan være kompliceret at finde ud af hvem ophavsmanden er og det kommer vi ind på senere.

Andre paragraffer som har betydning er:

> **§ 4.** Den, som *oversætter*, omarbejder eller på anden måde bearbejder et værk, herunder overfører det til en anden litteratur- eller kunstart, har *ophavsret* til værket i denne skikkelse, men kan ikke råde over det på en måde, som strider mod ophavsretten til det oprindelige værk.

Dette betyder at oversætteren har ophavsret til oversættelsen. Det er særlig vigtigt i islandsk sammenhæng siden 20-30% af teksterne i det islandske korpus skal være oversættelser.

I den danske lovs niende paragraf (samme paragraf i den islandske, svenske og norske lovtekst) er der bestemmelser om offentlige aktstykker:

> **§ 9.** Love, administrative forskrifter, retsafgørelser og lignende offentlige aktstykker er ikke genstand for ophavsret.

Dette betyder forhåbentlig at offentlige aktstykker frit kan inkluderes i et tekstkorpus.

I de nordiske lande findes der rettighedshaverorganisationer som tager sig af aftalelicenser om fotokopiering. Disse rettighedshaverorganisationer er: Bonus Presskopia, COPY-DAN, Fjölís, Fjølrit, Kopinor, Kopiosto og Samikopiija. I den danske lov handler paragraf 50 om disse organisationer:

Fælles bestemmelser om aftalelicens

> **§ 50.** Aftalelicens efter §§ 13, 14 og § 16 b, § 17, stk. 4, § 23, stk. 2, og §§ 30, 30 a og 35 kan påberåbes af brugere, der har indgået en aftale om den pågældende værksudnyttelse med en organisation, som omfatter en væsentlig del af ophavsmænd til en bestemt art af værker, der anvendes i Danmark. *Aftalelicensen giver brugeren ret til at udnytte andre værker af samme art, selv om ophavsmændene til disse værker ikke repræsenteres af organisationen.*
>
> *Stk. 2.* Aftalelicensen giver kun brugeren ret til at benytte de ikke-repræsenterede ophavsmænds værker på den måde og på de vilkår, som følger af den indgåede aftale med organisationen og af de i stk. 1 nævnte bestemmelser.
>
> *Stk. 3.* Rettighedshaverorganisationer, som indgår aftaler af den i stk. 1 nævnte karakter, skal godkendes af kulturministeren. Der kan kun godkendes *én* organisation inden for hver værksart. Ministeren kan bestemme, at en godkendt organisation på nærmere angivne områder skal være en fællesorganisation, som omfatter flere organisationer, der opfylder kravene efter stk. 1.

Det ville være nemt hvis man kunne lave en aftale med disse organisationer om brug af tekster i tekstkorpusser. Desværre lader det sig ikke gøre. Organisationerne har kun lov til at forvalte licens om fotokopiering og de er ikke i stand til at give licens til brug uden betaling. Men det er en vigtig forudsætning for tekstkorpusser at man ikke behøver at betale for brug af teksterne.

2. Korpusser i forskellige lande – hvordan man har fået licens til at bruge tekster i korpusser fra ophavsrettens indehavere

Vi har nu fastslået at i det mindste mange tekster er værnet af ophavsret. Det er måske oplysende at undersøge hvordan man bærer sig ad med at fremskaffe licens til at bruge tekster i et tekstkorpus i forskellige lande. Vi skal her se på hvordan det bliver udført for BNC (British National Corpus, engelsk), ANC (American National Corpus, amerikansk engelsk) , Korpus 2000 (dansk), Oslo korpuset av taggede norske tekster (norsk) og SUC (Stockhom Umeå Corpus, svensk).

Anglo-amerikansk ophavsret er sandsynligvis ganske forskellig fra den kontinental-europæiske som de nordiske lande tilhører. Det kan alligevel være oplysende at se på hvordan man i England og Amerika bærer sig ad.

BNC er et tekstkorpus med lidt over 3000 tekster, ialt 100 millioner ord. For alle tekster som er værnet af ophavsret har man skrevet til ophavsrettens indehaver og bedt om licens til at bruge teksten. Men det ser ud til at i mange tilfælde er ophavsrettens indehaver udgiveren. Man har passet på at specificere nøjagtigt hvordan teksten skulle bruges. Man tager aldrig med hele tekster som er værnet af ophavsret. Korpusset er søgbart i konkordansform på projektets hjemmeside. Man kan også købe brugerlicens og få hele korpusset på disketter eller købe „subscription service" og få adgang gennem internettet. Man betaler kun behandlingsomkostninger.

Det amerikanske korpus **ANC** er under opbygning og anden udgave har nu 22 millioner ord. Gennem projektets webside kan man lægge ind tekster (upload), d.v.s. man vælger ikke tekster som skal inkluderes, ophavsmændene selv byder dem for inkludering. Ophavsmanden skal sende e-post til projektlederen med standardiseret tekst som giver projektet licens til at bruge teksten „for the purposes of linguistic education, research, and development".

Korpus 2000 blev opbygget i Danmark omkring år 2000. Korpusset består af 28 millioner ord fra ca. 110.000 forskellige tekster skrevet i perioden 1998-2002. I projektbesrivelse til værket siges der:

„Det er vigtigt at fastslå, at søgning efter sproglige fænomener ikke er det samme som almindelig informationssøgning, og offentliggørelsen af Korpus 2000 vil derfor ikke automatisk give umiddelbar adgang til de involverede teksters fulde indhold. Denne adgang må af ophavsretslige grunde blive indskrænket: kun en mindre kontekst, der må antages omfattet af *citatretten*, vil kunne vises i de tilfælde, hvor tilladelse til udvidet fremvisning ikke har kunnet indhentes hos ophavsrettens indehaver. Det er sprogbrugen der skal kunne aflæses af en snæver kontekst - teksten i sin helhed er ikke relevant i denne sammenhæng."

Men i 22. paragraf i den danske lovtekst siger:

> **§ 22.** Af et offentliggjort værk er det tilladt at citere i overensstemmelse med god skik og i det omfang, som betinges af formålet.

Korpus 2000 er derfor et citatkorpus som defineres på websiden således.

„Et **citatkorpus** er et tekstkorpus, som først er splittet op i enkelte sætninger, som herefter er blandet i tilfældig rækkefølge. Det indeholder altså præcist det samme

sproglige materiale, som det oprindelige tekstkorpus, men sætningerne kommer blot i vilkårlig rækkefølge, så det ikke længere er muligt at rekonstruere de oprindelige tekster."

Hjemmesiden konstaterer at denne fremgangsmåde er nødvendig af ophavsretlige grunde.

Oslo-korpuset av taggede norske tekster, bokmålsdelen, indeholder omtrent 18,5 millioner ord og nynorskdelen omtrent 3,8 millioner ord. Jeg har fået information fra Anders Nøklestad om at teksten til korpusset blev hentet fra tekster som allerede var tilgængelige for internal brug inden for universitetet i Oslo. Når korpusset skulle blive åbnet for søgning på webben har man skrevet til alle institutter som havde bidraget med tekster. I brevet har man sagt at hvis man ikke fik svar skulle det betragtes som samtykke. Men adgang er kun med brugernavn og password som jeg forstår det.

Stockholm Umeå Corpus (SUC) blev udviklet 1990–1996. Den første version indeholder 500 filer med omtrent 2065 ord hver fil. Version 2 er søgbar på internettet gratis i konkordansform. Der findes ikke ret meget skrevet om hvordan licens for brug af tekster i SUC er blivet fremskaffet undtagen at „legal agreements" har været optegnet og at man måtte udelade nogle tekster på grund af at man ikke fik licens til at bruge dem.

3. Det islandske korpus

På Árni Magnússon instituttet findes der et korpus med omtrent 500.000 ord. Det består af 100 tekster som hver indeholder omtrent 5000 ord. Hvert ord er forsynet med morfosyntaktisk tag og lemma. Korpusset blev opbygget i anledning af arbejde med Den islandske frekvensordbog som blev udgivet i 1991 af det Leksikografiske institut.

Ministeriet for undervisning, forskning og kultur støttede sprogteknologiske projekter i årene 2000–2004 med 133 millioner islandske kroner. Et af projekterne som blev påbegyndt senest i denne periode var etablering af et større islandsk korpus. Korpusset skulle indeholde omtrent 25 millioner ord som er skrevet i år 2000 og senere. Hvert ord i korpusteksterne vil blive forsynet med oplysninger om ordklasse og bøjning, d.v.s. morfosyntaktisk tag, og lemma. Hver tekst får også tilføjet oplysninger om selve teksten.

Tekster til korpusset bliver hentet fra Ordbogens tekstsamling. En væsentlig del af projeket vil derfor dreje sig om at komplettere Ordbogens tekstsamling med tekster fra flere forskellige genrer, sikre licens til at bruge tekster som er værnet af ophavsret i tekstsamlingen og at tage tekstprøver derfra til korpusset. Det er endvidere nødvendigt at søge efter licens for tekster som allerede er i tekstsamlingen således at de kan bruges i korpusset.

I det følgende vil jeg give et status rapport om arbejde som foregår ved at sikre licens fra rettighedshavere til de tekster som vi gerne vil inkludere i korpusset.

Vi har selvfølgelig først prøvet at finde ud af hvordan andre har båret sig ad. I det foregående har jeg givet et oversigt over 5 korpusser.

Fremgangsmåden er afhængig af hvilken type tekst man vil få fat i og fra hvilket medie. Sidste sommer arbejdede vi med en student som var delvis finansieret af De Islandske studenters innovationsfond (*is*: Nýsköpunarsjóður námsmanna; *en*: The Icelandic Student Innovation Fund) for at skaffe tekster fra internettet til Ordbogens tekstsamling. Vi diskuterede livligt sagen om ophavsret i forbindelse med hendes

arbejde. Vi fik i første omgang vejledning fra en jurist i ministeriet for undervisning, forskning og kultur. Hans første vejledning var: „alle tekster er værnet af opahvsret uanset man ved hvem rettighedshaveren er“ (undtagen selvfølgelig offentlige aktstykker som vi allerede har set).

Dette betyder at man ikke kan inkludere en tekst i en tekstsamling som skal være offentlig tilgængelig uden at få tilladelse fra tekstens rettighedshaver. Men hvem er tekstens ophavsmand? Det er i mange tilfælde ikke så let at konstatere. Jeg skal først redegøre hvordan vi har båret os ad med forskellige typer af internettekster.

3.1 Internettet

3.1.1 Blogs

Vi begyndte med blogtekster. Først lavede vi en oversigt over forskellige weblogs. I mange tilfælde er bloggeren anonym eller bruger et pseudonym. For weblogs hvor der findes en blogmaster har han i sit register hvem bloggerne er men med hensyntagen til personværn har han ikke lov til at opgive deres identitet. Vi var derfor nødt til at begrænse valget til blogs fra bloggere som vi kunne kontakte. Bloggerne fik tilsendt e-post med projektbeskrivelse og en anmodning om at give licens til at bruge deres blog i Ordbogens tekstsamling og i korpusset. En erklæring blev sendt som attachment. Man kunne underskrive erklæringen og sende per post eller fax eller kopiere eklæringens tekst og klæbe den i en e-post og sende tilbage. Juristen i kulturministeriet havde konstateret at licens som blev sendt per e-post var gyldig. Det burde være muligt at finde hvem senderen er hvis der opstår noget tvivl. De fleste bloggere sendte svar per e-post. Nogle sendte svar per fax men ingen sendte brev. Med denne metode fik vi ialt blogs med omtrent 2,3 millioner ord fra 67 bloggere. Blogteksterne er nu tilgængelige i den åbne del af Ordbogens tekstsamling på www.lexis.hi.is , splittede i tre filer: præsteblog, blog fra politikere og anden blog. Bloggere fik lejlighed til at undtage enkelte dele af deres blogs.

Juristen i kulturministeriet havde endvidere foreslået at man skulle prøve at indgå samarbejde med dem som driver weblogs. De fleste weblogs drives på den måde at brugeren skal registreres og samtidig undergå betingelser om brug af servicen. Som led i betingelserne kunne man inkludere en paragraf om at Ordbogen fik lov til at bruge alle blogtekster på en speciel weblog service i sin tekstsamling og sine korpusser. Som følge heraf henvendte vi os til Morgunblaðið som er en af dem som driver weblog service. De behandlede sagen meget grundigt og fik en advokat til at bearbejde en rapport. Hendes konklusion var at af ophavsretlige grunde var der ikke noget imod at bloggere kunne acceptere at Ordbogen skulle bruge deres blogs i sin tekstsamling. Men hun bekymrede sig lidt om overtrædelse af loven om personværn og af den almene straffelov. Det er muligt at blogs indeholder oplysinger om andre personer som overtræder loven om personværn og kunne også indeholde injurierende udtalelser. Når man registrer på Morgunblaðiðs blogweb undergår man betingelser om at hvis man skriver noget som er injurierende og andre klager over bliver det fjernet med det samme.

3.1.2 Postlister

Vi har fået tekster fra e-postlister som blev fremskaffet ved at studenten har subskriberet på nogle af disse med samtykke fra postlistens webmaster. De fleste postlister indeholder annoncer af forskellig art og synes ikke at indeholde materiale

som er værnet af ophavsret. I et tilfælde var det nødvendigt at søge samtykke fra hver ophavsmand. Det var i tilfælde af en liste på Islands universitet hvor debatten var meget livlig. Problemet med postlisterne har været at i mange tilfælde er det svært at få fat i webmasteren som ikke synes at læse sin e-post eller at webmasteren er ukendt.

3.1.3 Taler

Det har været muligt at hente fra internettet adskillige taler som er blevet holdt i forskellige anledninger som prædikener, mindeord, korte taler på møder o.l. Vi har brugt e-post til at sende projektbeskrivelse og erklæring. De fleste sender erklæringen tilbage per e-post. Det synes ikke at være et problem for rettighedshavere at give sin tilladelse for brug af teksterne.

3.1.3 Webpladser

Vi har hentet en hel del af tekster fra forskellige webpladser som bliver opereret af offentlige institutter såvel som private firmaer. Forfatteren til tekst på websiderne er i de fleste tilfælde ukendt. Vi har nu fået besked om at man skal henvende sig til instituttets eller firmaets direktør. I de fleste tilfælde har offentlige institutter og firmaer sikret at de har ophavsret til en tekst som deres medarbejdere skriver og offentliggøres anonymt på deres webplads. Vi er nu i gang med at lave en standardiseret kontrakt som vi vil bede disse om at underskrive.

3.1.4 Den islandske videnskabsweb

Islands universitet har drevet en videnskabelig web i næsten 6 år. Man kan sende dertil spørgsmål om alt „mellem himmel og jord“ og få svar fra en specialist. Svaret bliver publiceret på webben. I tidens løb har man der samlet tekster som er skrevet for almene brugere om forskellige ting. Islands indbyggere er kun omtrent 300.000 således at det lønner sig ikke at publicere bøger om meget specielle emner. Videnskabswebben er derfor ofte den eneste kilde for tekster om forskellige specielle emner. Hver forfatter har ophavsret til sin tekst. Vi behøver derfor at lave en kontrakt med hver rettighedshaver og webbens hovedredaktør.

3.2 Nyheder i tv og radio

Vi har snakket med chefredaktøren for den statslige radio- og tv-stations nyhedsbureau om at få adgang til nyhedstekster. Der synes ikke at være problemer med at aflevere teksterne til brug i Ordbogens tekstsamling. Chefredaktøren har fået tilladelse fra reporterene til at aflevere teksterne til denne brug og han kan underskrive en kontrakt om brugen. I øjeblikket er problemet at det er ikke så let at eksportere teksten ud af radioens datasystem. Sandsynligvis får vi teksten fra et firma, Fjölmiðlavaktin, som indtaster alle nyheder fra alle nyhedsbureauer i islandske radio- og tv-stationer. Men forresten indtaster de kun indenlandske nyheder. Som ekstra bonus følger der transkriberet tekst fra interviews. Vi har fået besked om at vi kan bruge disse interviews men sandsynligvis behøver man at anonymisere dem som er blevet interviewet.

3.3 Aviser

Ordbogen råder allerede over store mængder af tekster fra aviser, især fra dagbladet Morgunblaðið. En del af disse er tilgængelige for søgning på ordbogens webside med

mundtligt samtykke fra Morgunblaðiðs redaktion. Vi tror at Morgunblaðiðs redaktion har sikret ophavsret til alle tekster i bladets database. Ordbogen can frit vælge tekster herfra. Men vi vil alligevel lave en kontrakt med Morgunblaðið um brug af deres tekster.

3.4 Andre medier

Flertallet af tekster i tekstsamlinger stammer fra bøger og tidsskrifter. For de fleste bøger er det helt klart at det er forfatteren som er rettighedshaveren. For artikler i tidsskrifter har forfatteren ophavsretten undtagen i de tilfælde hvor udgiveren har sikret ophavsret til artikler som udgives i tidsskriftet.

For alle tekster af denne type er det nødvendigt at vende sig til ophavsmanden og få licens til at bruge teksten. Vi har allerede informeret formændene til den islandske forfatterforening, faglitterær forfatterforening og foreningen af udgivere om projektet. De fleste ser positivt på projektet, men er desværre ikke i stand til at give licens på deres medlemmers vegne.

Vi har nu søgt juridisk assistance fra en advokat om hvordan kontrakter med rettighedshavere skulle se ud. Advokaten har konstateret at man godt kan bede enkelte forfattere om lov til at bruge alle deres tekster. Den „engelske“ metode går ud på at man først bestemmer hvilke tekster man vil bruge og derefter søger licens til at bruge teksten. Men det ville være nemt at kunne få en „global“ licens fra enkelte forfattere. Det ville måske løse vores problem med tekster som allerede er til stede i ordbogens tekstsamling og vi ikke kan tilbyde for søgning på ordbogens webplads af ophavsretlige grunde.

4. Konklusion

For en uge siden havde jeg skrevet et udkast til dette foredrag. På det tidspunkt var vi af den mening at rettighedshavere i det hele taget var positive over for projektet. Men der havde været debat om ophavsret med hensyn til digital kopiering af bøger og andet materiale som er beskyttet af ophavsret i aviser og andre steder i nogen tid. Den 8. oktober blev der publiceret en artikel i Morgunblaðið skrevet af formanden for den islandske forlæggerforening som han kaldte „Stafrænn óréttur“ eller „Digital uret“. Han havde været til et møde i Oslo i september måned med udgivere fra de andre nordiske lande. Formænd for alle de nordiske forlæggerforeninger havde vedtaget en erklæring om digital formidling. Formændene erklærer deres bekymringer over øget misbrug af ophavsretsbeskyttet materiale især indskanning af bøger. I artiklen er der fremført et eksempel fra Danmark. Man siger i erklæringen at indskanning af bøger som bliver formidlet over internettet er ganske hyppig i Norden. Der må man kæmpe imod den udbredte holdning at „alting bør være frit på internettet“. Mediet har selvfølgelig mange fordele, men der bliver ulemper for indehavere af ophavsret.

I de seneste par uger har man også opdaget et eksempel om ulovlig indskanning af lærebøger i en islandsk læreanstalt. Vi er derfor bange for at de foreninger som vi allerede havde kontaktet og syntes at være positive over for projektet nu har ændret deres holdning. Man er positiv over for selve projektet. Men man er imod den idé at give offentlige institutter licens til at bruge tekst uden betaling uanset værdien af de enkelte projekter imens det offentlige ikke vil lave en aftale om generel brug af digitaliserede tekster.

Det er helt sikkert at situationen er eksplosiv og at vi som beskæftiger os med korpusser må være meget forsigtige.

Den islandske lov om ophavsret blev senest ændret i februar 2006. De ændringer som blev vedtaget i den omgang var ændringer som man måtte vedtage på grund af Directive 2001/92/EC, The EU Copyright Directive. Dette direktiv indeholder en række valgfrie bestemmelser og nu drøfter man i Island hvilke af disse skal tages med i den islandske lov om ophavsret.

Man ved ikke på dette tidspunkt hvordan sagen udvikler sig. Det er klart at der vil være en kamp mellem udgivere og rettighedshavere på den ene side og det offentlige på den anden side om retten til at indskanne tekster, især lærebøger og gøre dem tilgængelige på læreanstalters intranetter.

Men vi håber at man i de næste uger vil finde ud af hvordan vi på vores institut kan komme til en aftale med rettighedshaverne om brug af deres tekster i korpusset. Vi kan i det mindste fortsætte med at hente materiale fra webben, men som vi allerede har sagt er det vores erfaring at man gerne giver lov til at teksten bliver brugt hvis det ikke er for kompliceret og tidskrævende at sende sin erklæring om samtykke. Man vil helst have mulighed for at sende den per e-post.

Man må passe på at alle rettighedshavere får detaljerede oplysninger om projektet, især om hvordan teksten skal bruges og hvordan man giver adgang til teksterne i tekstsamlingen og korpusset. Det er også nødvendigt at specificere at man aldrig inkluderer hele ophavsretligt værnede tekster i korpusset. Man må overbevise rettighedshaverne om at deres tekster ikke kan rekonstrueres fra korpusset.

Processen som man bruger til at fremskaffe licens til brug af tekster i tekstsamlinger og korpusser må være meget specifik. Man skal oplyse tekstleverandørerne om hvad vi gør med deres tekster, og herunder oplyse dem om, hvor tilgængelige teksterne bliver for tredjepart. Det er nødvendigt at henvende sig til dem som har e-post per e-post og give dem lejlighed til at svare per e-post. Andre bør man sende en adresseret og frankeret konvolut.

Måske er det ikke så svært som man tror at sikre licens fra rettighedshaveren til en tekstbrug af teksten i et korpus. Måske er det kun advokaterne som ser „djævelen i hvert hjørne“ som man siger på islandsk og er altid bange for at man bliver sagsøgt.

I den Språkteknologiske vismansrapport som bliver fremført senere i dag på seminariet diskuteres ophavsretsproblemet ved udvikling af korpusser. Det panel af eksperter som har skrevet rapporten har to forslag vedrørende ophavsretsproblemet. Det første drejer sig om at skabe fælles modelkontrakter for de nordiske lande for indsamling af ophavsrettsbeskyttet materiale. Det andet er om lovgivning. I rapporten siges:

> „Gemensamma modellkontrakt för att samla in copyright-skyddade korpusdata som garanterar möjligheterna att använda materialet på lämpligt sätt, borde skapas för alla de nordiska länderna, vilket kunde reducera utvecklingskostnaderna för språkmoduler betydligt.

Og endvidere siges der:

> „Lagstiftningen borde ändras så att det blir möjligt att samla in text- och talkorpus som används för forskning och utveckling av språkteknologiredskap. Att använda dylika korpus bör anses vara förenligt med principerna om kopieringsskydd när återpublicering av korpusen utesluts.“

Hvis disse to forslag bliver realiseret bliver livet meget lettere for fremtidens korpusudviklere.

SpråkVis - Språkteknologisk vismansrapport

Krister Lindén, Kimmo Koskenniemi och Torbjørn Nordgård

Utvidgad sammanfattning

Mandat

Nordiska Ministerrådet och Nordens Språkråd beställde en tioårsplan i form av en vismansrapport av prof. Kimmo Koskenniemi och prof. Torbjørn Nordgård över hur de nordiska (och baltiska) länderna kan göras till en ledande region i språkteknologi.

Med språkteknologi avses sådan teknologi som används av datorer för att bearbeta och stöda användningen av mänskligt språk. Traditionell språkteknologi är stavnings- och grammatikkontroll, maskinell översättning och taligenkänning. Tillämpningar för slutanvändare är många och skiftande, t.ex. skrivstöd i textbehandling, informationssökning i myndighetsportaler, dialoger i datorspel och hemelektronik, datorstödd språkinlärning, etc.

Avsikten med rapporten är att identifiera gemensamma nyckelområden för olika former av språkteknologi, storleken på nödvändiga investeringar, samarbetspartners och samarbetsformer som skapar förutsättningar för att göra Norden till en ledande region.

Arbetsform

Vi samlade in finansiell bakgrundsinformation om tidigare projekt i Norden och i de enskilda nordiska länderna (Danmark, Finland, Island, Norge, Sverige) för att få en överblick över tidigare investeringar. Informationen hämtades från offentliga databaser i de nordiska länderna och verifierades av inbjudna experter. Vi samlade även in policydokument och rapporter.

Vi sammanställde ett frågeformulär där vi bad experter kommentera och formulera en vision för 2016, identifiera hinder och trender. Vi bad även experterna ange storleken på de nödvändiga åtgärderna och investeringarna. Vi bjöd in 70 experter, varav 30 svarade. På basen av dessa svar identifierade vi olika nyckelområden.

Vi identifierade sex nyckelområden: policy, resurser, forskning och utveckling, utbildning och undervisning, lagstiftning och företagsaspekter, för vilka vi lägger fram rekommendationer i vismansrapporten. Avslutningsvis föreslår vi även en följd av åtgärder.

Bakgrund

Nordiska rådet har just avslutat ett forskningsprogram ”Nordisk Sprogteknologisk Forskningsprogram 2000-2004” med avsikt att höja profilen för det nordiska språksamfundet och säkerställa god nordisk språkteknologi för användarna. Mera specifikt innebar det tre mål för att stöda forskning och forskningsbaserad undervisning:

- förbättra kommunikationen mellan de nordiska forskarna i språkteknologi,
- förbättra samarbetet inom forskarutbildningen,
- etablera dokumentationscenter för att garantera tillgången till och spridningen av forskningsresultat, insamlade data och utvecklade redskap.

För att nå dessa mål valdes tre specifika prioritetsområden:

- CALL (Computer-Aided Language Learning) - datorstödd språkundervisning
- CLIM (Cross-Lingual Information Management) - tvärspråklig informationshantering
- NLHCI (Natural Language Human Computer Interaction) - kommunikation med datorer på naturligt språk

För att uppnå detta mål avsatte Nordiska rådet ca. 5 miljoner DKK årligen (23 278 500 DKK) dvs. Norden 0,6 M€/år (tot. 3,1 M€) under 2001-2004.

Satsningar i de nordiska länderna

För att jämföra forskningsfinansieringen i de enskilda nordiska länderna, sökte vi i de nordiska ländernas offentliga databaser och valde att titta på den statliga finansieringen av universitetsledda projekt, eftersom den fanns tillgänglig för alla de nordiska länderna under perioden 2003-2005. Siffrorna verifierades genom att cirkulera dem bland de berörda experterna i rapporten. Generellt kan sägas att grundsatsningarna i Sverige, Norge och Danmark har varit på samma nivå räknat per capita. I Norge och Island har man dock gjort strategiska tilläggssatsningar på språkteknologi under perioden. I jämförelse med de nationella satsningarna har den nordiska satsningen bidragit med ungefär en tiondel per capita.

Land	Årligen	Per invånare
Danmark	0,9 M€	0,2 €
Finland	2,1 M€	0,4 €
Island	0,2 M€	0,7 €
Norge	3,1 M€	0,7 € (0,2 € utan strategisk tilläggssatsning)
Sverige	1,6 M€	0,2 €
Norden	0,6 M€	0,02 €

I dessa siffror ingår inte statliga bidrag till kommersiellt ledd forskning. Inte heller EU-finansierad forskning ingår. Totalt har de enskilda Nordiska länderna finansierat universitetsledda forskningsprojekt för ca 24 M€ under 2003-2005.

Vad gjordes för pengarna?

De olika länderna har dock betonat olika typer av språkteknologi. En grov bild av satsningarna kan man få genom att dela in dem i t.ex. textbaserade och talbaserade teknologier. Alla länder har gjort något i båda kategorierna men endast Norge har satsat ungefär lika mycket på båda.

Land	Text	Tal
Danmark	x	(x)
Finland	(x)	x
Island	x	(x)
Norge	x	x
Sverige	x	(x)
Norden	x	(x)

Danmark

I Danmark finansierar Videnskabsministeriet forskning i språkteknologi under byrån för Forskning, teknologi och innovation, som sköter sekretariatuppgifter för ett antal självständiga råd. De två råden som sköter språkteknologi är det danska rådet för fri forskning (Danish Council for Independent Research) and det danska rådet för strategisk forskning (Danish Council for Strategic Research). Under 2003-2005 har Danmark spenderat ungefär 2,6 M€ huvudsakligen på textbaserad språkteknologisk forskning.

Finland

I Finland är de två statliga huvudfinansiärerna av forskning Finlands Vetenskapsakademi och TEKES (the Finnish Funding Agency for Technology and Innovation). Vetenskapsakademin finansieras av Undervisningsministeriet and TEKES finansieras av Handels- och industriministeriet. Under 2003-2005 har Finland spenderat ungefär 6,3 M€ med betoning på talteknologisk forskning.

Island

På Island har under 2003-2005 investerats ungefär 0,7 M€ med betoning på grundläggande textbaserade redskap och resurser.

Norge

I Norge är den huvudsakliga finansiären av universitetsledd forskning Norges forskningsråd (the Norwegian Research Council). Under 2003-2005 har Norge haft ett strategiskt forskningsprogram för språkteknologi "Kunnskapsutvikling for norsk språkteknologi (KUNSTI, 2001-2006)", vilket svarar för 70 % av finansieringen under perioden. Dessutom har Norge ett antal fristående projekt. Under 2003-2005 har Norge spenderat ungefär 9,2 M€ med en tämligen jämbördig täckning av text- och talbaserad språkteknologisk forskning.

Sverige

I Sverige sköts finansieringen av flera olika instanser, av vilka de huvudsakliga instanserna är Sveriges forskningsråd (The Swedish Research Council), VINNOVA (The Swedish Governmental Agency for Innovation Systems) och i lite mindre utsträckning Kunskapsstiftelsen (the Knowledge Foundation). En strategisk investering i språkteknologi avslutades före den valda jämförelseperioden. Under 2003-2005, har Sverige spenderat ungefär 4,8 M€ huvudsakligen på textbaserad språkteknologisk forskning.

Vad borde göras?

Man kan kanske begrunda huruvida det är lämpligt att på nordisk nivå göra precis som i de enskilda nordiska länderna? Kan man fördela arbetet mellan länderna? Det finns ju gott om uppgifter. Finns det en specifikt nordiska och mellanstatliga uppgifter? Vad bör och kan man göra med offentliga medel på nordisk nivå som gynnar alla parter och samtidigt gynnar en marknad för språkteknologi i Norden?

Vi har identifierat vissa gemensamma nyckelområden på mellanstatlig nivå, som skapar förutsättningar för att göra Norden till en ledande region för olika former av språkteknologi. Dessa nyckelområden är:

- policy
- resurser
- forskning och utveckling
- utbildning och undervisning
- lagstiftning och
- affärsverksamhet

Policy

Vi måste sprida insikten att språkteknologi har en nyckelposition för att bevara och upprätthålla våra språk och vår kultur. Språkteknologi behövs t.ex. i den digitala infrastrukturen för den humanvetenskapliga och den socialvetenskapliga forskningen. Det är ingen skillnad om språkteknologin har utvecklats akademiskt, med öppen källkod eller kommersiellt, så länge den finns och språkteknologimodulerna är kompatibla och tillgängliga för att bygga stora system och tillämpningar. Vi behöver en språkteknologisk infrastruktur.

Små språksamfund kommer inte att få språkteknologi på kommersiella grunder, så de flesta (eller alla) språk i regionen behöver åtminstone en viss mängd offentligt stöd och somliga kommer kanske att vara helt beroende av det.

På nordisk nivå behöver vi komma överens om rekommendationer för hur vi skall agera på det nationella planet. För att utvärdera situationen för språkspecifika och språkoberoende resurser för språken i regionen, borde en BLARK-rapport utarbetas (Basic Language Resource Kit), där de grundläggande språkresurserna i Norden kartläggs (10-25 k€/språk). Norden behöver hålla sig ajour med utvecklingen inom EU för att inte upprepa redan gjorda insatser och för att fokusera på det specifikt

nordiska. På nordisk nivå kan vi stöda sådant som alla har nytta av, dvs. metoder, standarder, avtalsmodeller, medan korpus och data bör samlas in på nationell nivå.

Deltagarna i NODALIDA 2005 beslöt grunda en förening för tal- och språkteknologi, som skall kallas NEALT (Northern European Association for Language Technology). En sådan organisation vore idealisk för att koordinera olika initiativ och nätverk (50 k€). Av specifikt nordiskt intresse är:

- att starta upp och etablera NEALT och en elektronisk publikation under dess ledning,
- någon form av fortsättning för NorDocNet[?] centren (jfr. Utbildning och undervisning),
- någon form av fortsättning för NGSLT via NordForsk[?] (jfr. Utbildning och undervisning), och
- individuella småprojekt (koordinerade och möjligen utförda av NEALT), t.ex. för att förbereda mera detaljerade rekommendationer för att
    - ändra lagstiftningen för immateriella rättigheter (IPR, jfr. Lagstiftning),
    - rekommendationer för finansierande institutioner för att garantera tillgång och återanvändning av språkteknologiska resurser skapade med offentliga medel (jfr. Forskning och utveckling), och
    - rekommendationer för forskning och/eller kommersiell användning av ordböcker och ordlistor skapade som en del offentligt finansierad kompilering av ordböcker (jfr. Resurser).

Resurser

Den mest uppenbara och viktigaste investeringen vore att skapa en lämplig infrastruktur som har tillräckligt med språkteknologiska resurser för relevanta språk i regionen. Resurserna bör kunna användas fritt för såväl forskning och undervisning som för kommersiell produktutveckling. På basen av den utvärdering av situationen som framkommer av BLARK-rapporten bör de viktigaste korpusarna skapas på nationell nivå med samarbete på nordisk nivå kring utveckling och utbyte av viktiga språkoberoende redskap och metoder.

Resurser för språkteknologisk infrastruktur:

- färdig uppsättning moduler såsom morfologiska och syntaktiska analysatorer och generatorer (2-5 M€),
- redskap för att bygga moduler (2-5 M€).
- korpus annoterade och oannoterade (10-15 M€ per språk),
- lexikon för tal och skriftspråk (10 M€ per språk).

OBS! Vi måste göra något för att få ner utvecklingskostnaderna på korpus och lexikon för språkteknologisk forskning och produktutveckling t.ex. genom lagstiftning och avtal.

Moduler

Både kommersiellt och akademiskt skapade språkteknologiska moduler behöver kompatibilitet och gemensamma gränssnitt för att kunna återanvända fristående

moduler och resurser. Språkoberoende redskap kan användas för att skapa både moduler och resurser. Gemensamma programvarugränssnitt gör det möjligt att använda modulkombinationer som befrämjar samkörbara och mångspråkiga produkter och system.

Redskap

Fritt användbara och uppdaterbara språkoberoende redskap behövs för att investeringarna i språkteknologi inte skall gå förlorade på långsikt. Samkörbara komponenter och mångspråkiga produkter kan åstadkommas med sådana redskap. T.ex. teorin och teknologin kring ändliga finita automater ger förutsättningar för mycket effektiva och modulära implementationer för ett antal olika uppgifter.

Korpus

Tal- och textkorpus och deras kombinationer är nödvändiga som utgångspunkt för många typer av språkteknologiska moduler och tillämpningar. Den nödvändiga kvantiteten av bearbetade korpusdatasamlingar har växt med flera magnituder på senare år, när man skapat metoder där datorer automatiskt kan lära sig från data. Olika typer av annotering av korpusdata är nödvändiga för olika metoder och forskningsändamål. Ofta utesluter tillgången till korpusmaterial kommersiell användning av slutresultatet, vilket omöjliggör utvecklandet av återanvändbara språkmoduler. Gemensamma modellkontrakt för att samla in copyright-skyddade korpusdata som garanterar möjligheterna att använda materialet på lämpligt sätt, borde skapas för alla de nordiska länderna, vilket kunde reducera utvecklingskostnaderna för språkmoduler betydligt.

Lexikon

Ordböcker och ordboksmaterial som har utvecklats med offentliga medel borde publiceras som öppen källkod så att de kan användas för att skapa språkteknologiska moduler så som morfologiska och syntaktiska analysatorer. Mer specifikt borde ordlistor med ord- och böjningsklass göras användbara så fritt som möjligt både för akademiskt och kommersiellt bruk. Hela texten i publicerade ordböcker kan reserveras för akademiskt bruk, men det får inte finnas begränsningar på metoder, regler och program, som har utvecklats på basen av dylikt material, om de inte innehåller bitar som är skyddade av copyright av original.

Forskning och utveckling

Finansiärer av akademisk forskning bör anamma rekommendationer och regler för språkresurser som skapas (eller har skapats) med allmänna medel. Det borde vara normal praxis att forskare gör språkresurserna tillgängliga för övriga forskare med så fria villkor och licenser som möjligt, vilket kan stödas med modellavtal (50 k€).

Dessutom bör vi överväga att öppna upp språkteknologiska resurser som utvecklats med offentliga medel för att bygga en nordisk språkteknologisk infrastruktur. Detta kan jämföras med att vi inte heller bygger offentligt finansierade vägar enbart för privat bruk!

Gemensamma gränssnitt och redskap bör skapas i samarbete med både kommersiella och akademiska parter. Vi bör utveckla API-standarder, kvalitetsstandarder och testmetoder för kvalitetsgranskning av färdiga moduler (15 M€).

På nationell nivå bör det även satsas på tillämpningar och vidareutveckling för olika specialområden där de olika länderna har kärnkompetens fördelat både på grundforskning (15 M€) och tillämpad forskning (50-80 M€).

Utbildning och undervisning

Mera samarbete behövs kring akademisk utbildning mellan universiteten i den nordiska och baltiska regionen. Som en del av det nordiska språkteknologiska forskningsprogrammet startades NorDocNet[2] i de fem nordiska länderna, vilket bör få en fortsättning och en utvidgning till en mera internationell dimension så som http://www.lt-world.org/ eller som en baltisk eller en gemensam nordisk-baltisk insats.

En tillräcklig mängd specialister med doktors- och kandidatexamen bör behärska de mest avancerade färdigheterna och alla regionens länder och språkgrupper bör delta inklusive minoriteter och små språkgrupper.

För att stöda utbildning och undervisning bör vi:

- dokumentera existerande resurser (1 M€),
- utveckla material för undervisning av formell språkkunskap i skolorna (1 M€),
- producera introduktionsmaterial för att distansutbilda personalen inom IT-industrin i språkteknologi (50 k€),
- publicera en vetenskaplig tidskrift på internet för NEALT (50 k€),
- diversifiera och specialisera Master's utbildningen genom distansundervisning, utbytesprogram, och gemensamma utbildningsprogram (2 M€),
- koordinera doktorsutbildningen: NGSLT (1 M€).

Lagstiftning

Nuvarande lagstiftning om kopieringsskydd gör det onödigt svårt och dyrt att samla in och annotera text- och talkorpus. Vissa privilegier ges för tillfället åt några nationella bibliotek för att arkivera elektroniska kopior av böcker, tidningar, osv. och ett liknande privilegium behövs för att skapa språkteknologiresurser. Lagstiftningen borde ändras så att det blir möjligt att samla in text- och talkorpus som används för forskning och utveckling av språkteknologiredskap. Att använda dylika korpus bör anses vara förenligt med principerna om kopieringsskydd när återpublicering av korpusen utesluts. En arbetsgrupp för att driva saken borde upprättas (10 k€). Detta kunde göra det mera produktivt att samla tal- och textkorpus genom att garantera bredare spridning och bättre användningsmöjligheter för forskningsmaterial som samlats in av olika centra (t.ex. nationella språkbanker) eller genom att låta enskilda forskare utbyta material.

Dessutom måste vi på olika sätt motarbeta tendensen att det utfärdas programvarupatent på uppenbara eller publicerade lösningar och idéer.

Affärsverksamhet

Licensvillkoren för språkteknologiresurser måste tillåta och uppmuntra både kommersiell och akademisk användning. Tillämpad forskning på medellång sikt i samarbete mellan universitet och industri bör uppmuntras nationellt för att skapa tillämpningar som utnyttjar språkteknologi (5 M€).

Man kunde stimulera marknaden för mera ambitiösa språkteknologiska tillämpningar genom att anslå medel för den offentliga sektorn att utveckla service med språkteknologiska hjälpmedelmedel för eget bruk (5 M€).

Åtgärdsplan

Målet med rapporten var att identifiera nyckelområden, storleken på finansieringen, berörda parter och former för samarbete. För att förverkliga målen och för att utarbeta mer detaljerade planer och tidsramar för områdena i 10-årsplanen, föreslår vi att resurser allokeras för:

1. etablering av NEALT och dess arbetsutskott,
2. mandat för att utarbeta BLARK-rapporter för de nordiska språken, som inventerar existerande språkresurser och resursbehov,
3. nordisk finansiering av samarbete inom språkteknologisk utbildning och undervisning,
4. nationell finansiering av tillämpad forskning på medellång sikt i samarbete mellan universitet och industri.

När BLARK-rapporterna har färdigställts, bör resurser under NEALTs koordinering allokeras för:

1. nordisk finansiering av språkteknologiska redskap baserade på BLARK-rapporternas rekommendationer,
2. nordisk och nationell finansiering av korpus, trädbanker, och lexikon i enlighet med BLARK-rapporterna.