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 Akademi och TEKES (Finnish Funding Agency for Technology and Innovation). Finlands Akademi 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 (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 (Swedish Research Council), VINNOVA (Swedish Governmental Agency for Innovation Systems) och i lite mindre utsträckning Kunskapsstiftelsen (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 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.