

# SPRÅKVIS - VISMANSRAPPORT - EXPERT PANEL REPORT

## The Nordic Countries - A Leading Region in Language Technology

Edited by [Krister Lindén](#), [Kimmo Koskeniemi](#) and Torbjørn Nordgård

The Nordic Council of Ministers has [commissioned a ten-year plan](#) in the form of an expert panel report (= vismansrapport) for making the Nordic Countries a leading region in language technology (LT). LT means a number of technologies used by computers for processing human language, e.g. spell-checking, machine translation and speech recognition to mention only the most well-known. Applications are diverse. The aim of the report is to identify the common key areas which need to be addressed when making the Nordic countries into a leading region. The report highlights [key areas, magnitudes of investments, suggested partners, modes of cooperation and some initial key actions](#).

The Nordic Council of Ministers has recently concluded a successful LT Research Programme, which is briefly outlined as [background](#) information. This investment should be seen in relation to the investments the Nordic Countries have made in university-lead LT development projects in [Denmark](#), [Finland](#), [Iceland](#), [Norway](#) and [Sweden](#). Information on these was collected from public databases in the Nordic Countries and the information was circulated for comments among the contributors to the report.

We sent out a questionnaire among [70 invited experts](#) from the Nordic Countries collecting comments on [an initial vision for LT in 2016](#) and its [prerequisites](#) as well as [current obstacles for LT development](#) and [general trends influencing LT](#) development and its applications. In the questionnaire we also asked for [recommendations](#) on the order of magnitude of investments and modes of cooperation needed. Of the invited experts, 30 contributed their comments, which we hereby gratefully [acknowledge](#). When analyzing the background and the comments on the questionnaire, we identified six key areas: [LT Policy](#), [LT Resources](#), [LT Research and Development](#), [LT Training and Education](#), [LT Legislation](#) and [LT Business Aspects](#) for which we present our recommendations and an action plan in this Expert Panel Report.

# Contents

Sammanfattning .....	3
Summary .....	5
Utvidgad sammanfattning.....	7
Extended Summary .....	15
Mandate .....	21
Key concepts .....	23
Background.....	26
LT Policy .....	28
LT resources.....	32
LT Research and Development.....	39
LT Training and Education.....	42
LT Legislation.....	45
LT Business Aspects .....	47
Initial Action plan.....	49
Acknowledgements .....	50
References.....	52

## APPENDIXES

Invited Experts for the Expert Panel Report.....	54
Danish LT projects .....	57
Finnish LT projects .....	58
Icelandic LT projects.....	60
Norwegian LT projects.....	61
Swedish LT projects .....	62
Weaknesses in or obstacles for LT development.....	64
Opportunities or Threats for current LT.....	67
Vision.....	70
Prerequisites for implementing the Vision .....	73
Key areas with magnitudes of investment .....	77

---

# SpråkVis - Språkteknologisk vismansrapport

av Krister Lindén, Kimmo Koskeniemi och Torbjørn Nordgård

## Sammanfattning

Nordiska Ministerrådet har beställt en tioårsplan i form av en vismansrapport för att göra de nordiska länderna till en ledande region i språkteknologi. Sex nyckelområden har identifierats: Policy, Resurser, Forskning och utveckling, Utbildning och undervisning, Lagstiftning och Affärsverksamhet, för vilka vi presenterar rekommendationer och en åtgärdsplan i den här vismansrapporten.

**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. 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 där de grundläggande språkresurserna i Norden kartläggs**. Norden behöver hålla sig ajour med utvecklingen inom EU för att inte upprepa redan gjorda insatser och för att fokusera på det specifikt nordiska. 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.

**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**.

**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.

Gemensamma gränssnitt och redskap bör skapas i samarbete med både kommersiella och akademiska parter.

**Utbildning och undervisning:** Mera samarbete behövs kring akademisk utbildning mellan universiteten i den nordiska och baltiska regionen. 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.

**Lagstiftning:** Nuvarande lagstiftning om kopieringsskydd gör det onödigt svårt och dyrt att samla korpus. 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.**

**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.

**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
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

---

# SpråkVis - Language Technology Expert Panel Report

by Krister Lindén, Kimmo Koskenniemi och Torbjørn Nordgård

## Summary

The Nordic Council of Ministers has commissioned a ten-year plan in the form of an expert panel report for making the Nordic Countries a leading region in language technology (LT). Six key areas were identified: [LT Policy](#), [LT Resources](#), [LT Research and Development](#), [LT Training and Education](#), [LT Legislation](#) and [LT Business Aspects](#), for which we present recommendations and an action plan in this Expert Panel Report.

**LT Policy:** We need to raise awareness that **LT has a key position for protecting and maintaining our languages and our culture**. LT is necessary e.g. for developing a digital infrastructure for research in the humanities and the social sciences. It does not matter whether LT is academic, open source or commercial, as long as it exists and its modules are compatible and available for building large systems and applications. Small language communities will not get LT on a commercial basis alone, so most (or all) languages in the area need at least some public support and some may be totally dependent on it. **At the Nordic level, we need to establish recommendations** for the actions on the national level. To assess the situation for language-specific and language-independent resources for the languages in the area, a **Basic Language Resource Kit (BLARK) report for the Nordic languages** should be prepared. The Nordic region needs to stay abreast with the development in the EU in order not to duplicate efforts and in order to focus on the aspects that are specifically Nordic. The participants of the NODALIDA 2005 decided to establish an association for speech and language technology which will be called [NEALT \(Northern European Association for Language Technology\)](#). Such an association would be ideal for coordinating various initiatives and networks.

**LT Resources:** The most obvious and substantial investment would be to create an appropriate infrastructure which has sufficient LT resources for relevant languages of the area. The resources belonging to the infrastructure should be freely available for research and training as well as for commercial product development. Based on the assessment of the situation in the BLARK report the most **urgent gaps in availability of corpora should be filled in using national funding with cooperation on the Nordic level for developing and exchanging language-independent tools and methods**.

**LT Research and Development:** The academic funding institutions ought to adopt recommendations or rules concerning linguistic resources which will be (or have been) developed using public funding. It ought to be a normal requirement that the researchers make the linguistic resources available for the rest of the research

community with as free conditions or licenses as possible. Common interfaces and tools must be created in cooperation between both commercial and academic parties.

**LT Training and Education:** More cooperation is needed in academic training among the universities in the Nordic/Baltic region. A sufficient number of highly skilled PhDs and Masters ought to be trained with the best possible LT skills and all countries and language groups should be participating, including minorities and small language communities.

**LT Legislation:** Current copyright legislation makes the collection of resources unnecessarily difficult and costly. Certain privileges are currently granted to a few national libraries for archiving electronic copies of books, journals etc. and similar privileges are needed for creating LT resources. The legislation should be changed so that the **collection of text and speech corpora for the purposes of research and development** is possible. The use of such corpora **should be deemed to conform to the principles of copyright when excluding republication.**

**LT Business Aspects:** The licensing conditions of LT resources must allow and encourage both their commercial and academic use. Medium term applied research projects involving university and industrial partners should be encouraged.

**Action Plan:** The aim of the report was to identify key areas, magnitude of funding, parties involved and modes of cooperation. To implement the goals and to further specify the areas and their time-frames in the 10-year plan, we suggest that resources are allocated for:

1. Establishing of NEALT and its working groups
2. Commissioning BLARK reports for the Nordic languages
3. Nordic funding for cooperation on LT training and education
4. National funding of medium-term applied research projects involving university and industrial partners

When the BLARK reports have been delivered, resources coordinated by NEALT should be allocated for

1. Nordic funding of LT tools according to the recommendations of the BLARK reports
2. Nordic and national funding of corpora, treebanks and lexicons based on the BLARK report recommendations

---

# 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 språkinläring, 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 prioriteringsområden:

- 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 (Finnish Funding Agency for Technology and Innovation). Vetenskapsakademien 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 modul kombinationer som befämjar 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 hjälpmedel 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.

---

# SpråkVis - Language Technology Expert Panel Report

by Krister Lindén, Kimmo Koskenniemi och Torbjørn Nordgård

## Extended Summary

The Nordic Council of Ministers has [commissioned a ten-year plan](#) in the form of an expert panel report for making the Nordic Countries a leading region in language technology (LT). Six key areas were identified: [LT Policy](#), [LT Resources](#), [LT Research and Development](#), [LT Training and Education](#), [LT Legislation](#) and [LT Business Aspects](#), for which we present recommendations in this Expert Panel Report. Finally, we also suggest an action plan.

### LT Policy

We need to raise awareness that **LT has a key position for protecting and maintaining our languages and our culture**. LT is necessary e.g. for developing a digital infrastructure for research in the humanities and the social sciences. It does not matter whether LT is academic, open source or commercial, as long as it exists and its modules are compatible and available for building large systems and applications. Small language communities will not get LT on a commercial basis alone, so most (or all) languages in the area need at least some public support and some may be totally dependent on it. **At the Nordic level, we need to establish recommendations** for the actions on the national level. To assess the situation for language-specific and language-independent resources for the languages in the area, a **Basic Language Resource Kit (BLARK) report for the Nordic languages** should be prepared. The Nordic region needs to stay abreast with the development in the EU in order not to duplicate efforts and in order to focus on the aspects that are specifically Nordic. The participants of the NODALIDA 2005 decided to establish an association for speech and language technology which will be called [NEALT \(Northern European Association for Language Technology\)](#). Such an association would be ideal for coordinating various initiatives and networks.

Action areas, where Nordic funding is needed instead of national funding, are:

- establishing and starting NEALT and establishing a scientific electronic journal by NEALT,
- some form of continuation for the Nordic LT documentation centers, see awareness under [LT Training and Education](#),
- some continuity for the [NGSLT](#), by NordForsk, see [LT Training and Education](#), and
- individual small-scale projects (possibly carried out and coordinated by NEALT) e.g. to prepare more detailed recommendations for

- altering the legislation of intellectual property rights (IPR, see [LT Legislation](#)),
- guidelines for funding agencies to guarantee access and reuse of LT resources created with public funding (see [LT Research and Development](#)), and
- guidelines for research and/or commercial use of dictionaries and word lists created as part of publicly funded dictionary compilation (see [LT Resources](#)).

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
NEALT start-up	50 kEUR	NMR for funding	association, working groups
BLARK Report	10-25 kEUR per language	NorDokNet, NEALT	national projects coordinated at the Nordic level

## LT Resources

The most obvious and substantial investment would be to create an appropriate infrastructure which has sufficient LT resources for relevant languages of the area. The resources belonging to the infrastructure should be freely available for research and training as well as for commercial product development. Based on the assessment of the situation in the BLARK report the most **urgent gaps in availability of corpora should be filled in using national funding with cooperation on the Nordic level for developing and exchanging language-independent tools and methods.**

### LT modules

Both commercially and academically created LT modules need compatibility and capabilities for reusing other modules and resources. Language-independent tools can be used for creating both kinds of modules, and common API interfaces make it possible to utilize module combinations in order to facilitate interoperable and multilingual products and systems.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Openly available LT modules and common APIs	2-5 MEUR	open source community, universities, public and private institutions, NEALT	Nordic LT network

### LT tools

Freely usable language-independent state of the art tools are needed so that investments in LT modules are not lost in the long term. Interoperable components and multilingual products and systems can be achieved through such tools. E.g. finite-state technology provides very efficient and modular implementations for a number of tasks.

Key Area	Magnitude of	Parties involved	Mode of
----------	--------------	------------------	---------

	funding needed		cooperation
Openly available LT tool	2-5 MEUR	open source community, universities, public and private institutions, NEALT	Nordic LT network

### LT corpora

Speech and text corpora and their combinations are necessary starting points for many types of LT modules and applications. The required quantities have grown in magnitude. Different levels of annotation are necessary for various methods and research topics. The availability of corpus material is often too restricted excluding all commercial use and, at the same time, any development of LT modules. **Model contracts for collections of copyright-protected corpora** should be created for all countries, and these model contracts should guarantee the necessary ways to use the materials.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Model contracts	50 kEUR	research organizations, lawyers, NEALT	networking across countries
Corpus collection, written text	10-15 MEUR pr language	universities, NEALT	networking across countries
Corpus collection, spoken data	10-20 MEUR pr language	universities, NEALT	networking across countries

### LT lexicons

Dictionary materials which have been developed with public funding ought to be published as open source material so that they can be used for creating LT modules such as parsers and analyzers. More specifically, **lists of headwords annotated with part of speech and inflectional class should be made available under very free conditions** permitting their use in both academic and commercial contexts. The full text of dictionaries published as books may be reserved for academic use, but there must not be limitations on further use of methods, rules or programs which have been developed using such material, provided that they do not contain parts infringing on the copyright of the original work.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Lexicon development	10 MEUR per language	universities, NEALT	networking across countries

## LT Research and Development

The academic funding institutions ought to adopt recommendations or rules concerning linguistic resources which will be (or have been) developed using public funding. It ought to be a normal requirement that the researchers make the linguistic resources available for the rest of the research community with as free conditions or licenses as possible. In addition we may need to open up language resources on all

levels (lexicons, grammars, written language corpora and speech corpora, etc.) which have been created through public funding. Common interfaces and tools must be created in cooperation between both commercial and academic parties.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Recommendations for research result materials	50 kEUR	funding organizations, universities, NEALT	working groups
Joint effort for standardization	15 MEUR	universities, industry, NEALT	academia/industry collaboration
Basic technology research	15 MEUR	universities	joint programme, researcher exchange, workshop, division of research tasks
R&D Funding	50-80 MEUR	universities, research institutes, industry	Nordic projects

The R&D funding can be further specified into various fields of services and applications for the society.

## LT Training and Education

As a part of the Nordic Language Technology Research Program 2000-2004, a [LT documentation centre](#) was established in each of the five Nordic countries. Some continuation for them is needed, either in conjunction with some world-wide effort such as the [LT world](#) or as a Nordic or Nordic-Baltic effort. More cooperation is needed in academic training among the universities in the Nordic/Baltic region. A sufficient number of highly skilled PhDs and Masters ought to be trained with the best possible LT skills and all countries and language groups should be participating, including minorities and small language communities.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Nordic LT documentation	1 MEUR	NMR	network of LT documentation centres
NEALT Journal start-up	50 kEUR	NEALT, Nordisk Publiceringsnämnd	scientific electronic journal
Coordinated PhD education	1 MEUR	Nordic/Baltic universities	NGSLT
Master's level education	2 MEUR	Nordic/Baltic universities	distance education, exchange programs for teachers and students, common curriculum
Distant learning	50 kEUR	Nordic/Baltic	production of the material

courses for  
commercial  
developers

universities

Popularization	1 MEUR	R&D, Government, Industry, Secondary Education	professional PR assignment
----------------	--------	--	-------------------------------

## LT Legislation

The development of LT tools depends on the availability of language resources such as corpora. Current copyright legislation makes the collection of resources unnecessarily difficult and costly. Certain privileges are currently granted to a few national libraries for archiving electronic copies of books, journals etc. and similar privileges are needed for creating LT resources. The legislation should be changed so that **collecting, annotating and sharing of text and speech corpora for the purposes of research and development** becomes easier. The use of such corpora **should be deemed to conform to the principles of copyright when excluding republication**. Changing the copyright legislation would make collecting corpora more productive by guaranteeing that corpora and annotated material are available for research and development purposes. Availability can be achieved either by allowing centres (such as national language banks) share materials with each other or by allowing individual researchers to share them.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Preparation of changes in the legislation	10 kEUR	relevant ministries, universities, NEALT	working groups

## LT Business Aspects

The licensing conditions of LT resources must allow and encourage both their commercial and academic use. Medium term applied research projects together with industrial partners should continue. Funding should be provided for creating and purchasing LT applications and services for the public sector. This funding is intended to stimulate the LT service and application market uptake. Such services could include more ambitious goals using LT-enhanced applications.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
LT module uptake	5 MEUR	industry and universities	action plan managed at Nordic level
Web services	5 MEUR	industry and universities	academia/industry collaboration

## Action plan

The aim of the report was to identify key areas, magnitude of funding, parties involved and modes of cooperation. However, we are still left with questions

regarding further specification of the plans as well as priorities and time-frames within the 10-year period. Some answers have been sketched for the organization of the work, but more detail is needed as well as some further consideration of the division of national and Nordic funding. To implement the goals and to further specify the areas and their time-frames in the 10-year plan, we suggest the following steps in allocating resources:

1. Establishing NEALT and its working groups
2. Commissioning BLARK reports for the Nordic languages
3. Nordic funding for cooperation on LT training and education
4. National funding of medium-term applied research projects involving university and industrial partners

When the BLARK reports have been delivered, resources coordinated by NEALT should be allocated for

1. Nordic funding of LT tools according to the recommendations of the BLARK reports
2. Nordic and national funding of corpora, treebanks and lexicons based on the BLARK report recommendations

---

# Mandate

## Background

The Language Technology Priority Area Project of The Nordic Council of Ministers has now completed its work. The Priority Area Project has been very effective in joining the Nordic Countries and in revealing the need and possibilities for a continued cooperation in order to maximize the gain within the Nordic region from national as well as Nordic investments in the Priority Area. The Project has also proved to be effective in involving the Baltic region in the Nordic community.

The investment has therefore created the prerequisites for attempting to fulfill the vision of the Nordic Countries (possibly including the Baltic region) as a leading region in language technology.

At the same time, the Nordic Council of Ministers, the Nordic Language Council, the Nordic cooperation between national Language Councils, and EK-IT have invested significantly in preparing common investments in areas like educational material, education, dictionary resources (the Web Dictionary, Scanlex, Tvärsök, etc.) and language control software (symposium in Pargas).

In several Nordic Countries, especially Norway (plans for a national language bank) and Denmark (plans for a strategic investment in language technology research), there are finally proposals for reasonably large investments in research and development on a national level that may potentially yield an enormous return, if firstly they are implemented and secondly they are co-planned on a Nordic level.

This is the background for the fact that it is realistic to ask an Expert Panel for a 10-year plan, which can highlight the magnitude of **investments in key areas** as well as the **modes of cooperation** between

1. publicly-funded basic and strategic research,
2. privately-funded research and development,
3. distribution to end-users via language councils, publishers and private entities, as well as
4. development of basic language technology resources

which are needed for realizing the vision of the Nordic Countries as a leading region in language technology.

## 1. Purpose

The Nordic Language Council starts an Expert Panel Report on Nordic Language Technology with regard to realizing within a 10-year period the vision of the Nordic Countries as a leading region in language technology.

## **2. Members**

The Nordic Language Council hereby appoints Professor Kimmo Koskenniemi, University of Helsinki, and Professor Torbjørn Nordgård, NTNU, Trondheim, to the Language Technology Expert Panel. The two language technology experts have the mandate:

- to create a plan for the Nordic Council of Ministers how, within a 10-year period, to realize the vision of the Nordic Countries as a leading region in language technology.

## **3. Contacts**

The Expert Panel - possibly in cooperation with the Secretariat of the Nordic Council - is required to involve national experts as well as experts from the on-going projects and from the Nordic cooperation between the Language Councils. It is especially emphasized that the Expert Panel should keep contact with the interested Ministries (in Denmark the Ministry of Science, Technology and Innovation, in the other Nordic Countries the Ministries of Education and Research) and research communities in all the Nordic Countries, possibly in the form of interviews, and that experience and results from the cooperation project in language technology can be utilized by involving its Coordinator Henrik Holmboe in the process.

## **4. Financing**

The Nordic Council of Ministers grants 100.000 DKK for the purpose.

## **5. Deadline**

End of June, 2006.

# **Signature of Senior Adviser**

Copenhagen, December 16, 2005  
Hulda Zober Holm

*(translated from the Danish original)*

-- [KristerLinden](#) - 01 Jun 2006

---

# Key concepts

## Language Technology (LT): digital language infrastructure vs. applications

**Language technologies** are *information technologies that are specialized in dealing with the most complex information medium in our world: human language*. Therefore these technologies are also often subsumed under the term *Human Language Technology (HLT)*. Human language occurs in spoken and written form. Whereas speech is the oldest and most natural mode of language communication, complex information and most of human knowledge is maintained and transmitted in written texts. Speech and text technologies process or produce language in these two modes of realization. But language also has aspects that are shared between speech and text such as dictionaries, most of grammar and the meaning of sentences. Thus large parts of language technology cannot be subsumed under speech and text technologies. Among those are technologies that link language to knowledge. We do not know how language, knowledge and thought are represented in the human brain. Nevertheless, language technology had to create formal representation systems that link language to concepts and tasks in the real world. This provides the interface to the fast growing area of knowledge technologies (<http://www.dfki.de/~hansu/LT.pdf>, and for a comprehensive survey of language technologies, see <http://www.dfki.de/~hansu/HLT-Survey.pdf>)

By **digital language infrastructure** we mean *all basic software tools, language and speech data, corpora and lexicons that are necessary for conducting research and developing applications in the field of HLT*. Since the costs of developing HLT resources are high, it is important that all parties involved, both in industry and academia, co-operate so as to maximize the outcome of efforts in the field of HLT. This particularly applies to languages that are commercially less interesting than English. (<http://www.cnts.ua.ac.be/Publications/2001/CDS01/20020103.6842.cds01.pdf>).

Although existing LT systems are far from achieving human ability, they have numerous possible **applications**. These applications are *software products or services that have some knowledge of human language*. Such products are going to change our lives. They are urgently needed for improving human-machine interaction since the main obstacle in the interaction between human and computer is merely a communication problem (<http://www.dfki.de/~hansu/LT.pdf>):

- Friendly technology should listen and speak
- Machines can also help people communicate with each other
- Language is the fabric of the web

## Public resources vs. commercial interest

**Resource-poor languages** are those languages for which the *digital language infrastructure is deficient* in some aspect as opposed to resource-rich languages with

no such handicaps. **Commercially interesting languages** are those languages for which it is profitable to produce *commercial LT applications* as opposed to commercially uninteresting languages, for which LT applications have to be produced using public funding. A language may be both commercially interesting in one aspect and resource-poor in another.

When a language community subsidizes or creates freely available infrastructure, it is generally advisable to invest in the precompetitive tools and resources of the commercially interesting aspects of a language, or in creating tools and resources that are considered vital for the survival of the language community but which are commercially uninteresting due to market size. What is commercially precompetitive may therefore vary with the size of the language community, e.g. a small language community like Sámi may find it vital to publicly fund the development of a grammar-checker as a precompetitive tool for word processing applications, whereas a large language community may find that even a morphological analyzer is a commercially interesting application because it can be developed with private funding and sold for a reasonable fee to enough customers for recovering the development costs and some profit. Similar reasoning may be applied to other parts of the digital language infrastructure.

### **Free and open source vs. non-free software**

**Free software** is a specific term referring to the ability of anybody to use, modify and develop the software. Free software is typically protected by copyright but distributed with a specific license permitting those freedoms. Best known of such licenses is the GNU General Public License or GPL which guarantees this kind of freedom to persist even after modifications or further developments, see <http://www.gnu.org/licenses/gpl.html>. In addition, the source code of free programs has to be available to anybody.

Free software is **not** the same thing as a program **not costing anything**. Shareware, evaluation copies of products and many proprietary products distributed free of charge are **not** considered free because one is typically not allowed to modify or develop them further, or there are other types of restrictions. E.g. the Sun corporation distributes Java software with a license (see <http://www.java.com/en/download/license.jsp>) where it is said among other things: *Unless enforcement is prohibited by applicable law, you may not modify, decompile, or reverse engineer Software.*

Thus, free software relates to the freedom of doing rather than to the absence of something. The opposite of free software is **proprietary** software. Proprietary software is owned by some company, who typically has an interest to keep the programs in its full control and prevent others from studying the internal methods and constructions of the software not to mention modifying or developing the program. Proprietary software is typically distributed only in binary forms (unreadable for humans).

There are several somewhat different licenses which are used for implementing this freedom for software, see e.g. <http://www.gnu.org/philosophy/license-list.html>. The term **open source software** refers to all these approaches which differ in several

respects but all include the free distribution of the source code of the software. One difference among these licenses is in the persistence of the license after modifications. Some, like the BSD license and the MIT license, allow open software to be turned into proprietary products.

## Copyright and creative commons

Copyright is one of the **intellectual property rights** and it protects the form of an **intellectual work** (such as a book, painting, composition, or a computer program) which consists of sufficient amount of nontrivial decisions made by a human. Copyright protects the **form** of such works, not their underlying ideas. Copyright as such restricts the making of copies (e.g. by printing or by producing CDs), and the publication of the work (e.g. performing in a concert, broadcasting in the radio, or making available at a web site). On the other hand, copyright as such does not restrict the further selling of legal copies of the work (but associated licenses may do so).

Copyright persists for some 70 years after the death of the author (or the last of the authors of a joint work). Whereas a few of our greatest authors, artists or composers have created works that are still of interest after such a period, it is likely that in most cases the length of the period is more than enough. In many countries, this protection will automatically be in force without any actions required from the creator of the work.

It has been claimed that for the vast majority of works much less protection would be sufficient. For a normal author, artist or composer, it has been very difficult to withdraw any of this protection without hiring an expensive lawyer.

For some areas of writing, such as scientific articles and research results, the author typically wishes to distribute the work as widely as possible in order to become better known and recognized, whereas commercial publishers have an interest to restrict dissemination to the paid copies. This has led to the emergence of **open access** publishing where the author only restricts the right to alter the text and its authorship, but allows free copying.

**Creative commons** is an effort to facilitate the authors, artists and composers to distribute materials with **some rights reserved**, see <http://creativecommons.org/>. The author may choose the level of protection needed, and include an icon on the web page which is a link to the corresponding summary of the license and its detailed paragraphs. This makes it easy even for a casual creator of works to reserve some rights and give a suitable level of freedom for others.

Along with the GPL license which was mentioned above, there are similar licenses which have been designed for user manuals and other technical texts for which it is essential that other people can go on improving the text. In the **Wiki environments**, it is common to oblige the authors to comply with the GNU Free Documentation License, see <http://www.gnu.org/licenses/fdl.html>. This is essential in Wiki environments where many people will contribute to the contents by altering and improving the text of others.

---

# Background

In "Nordisk Sprogteknologisk Forskningsprogram 2000-2004. Epilog", the aim of the previous Nordic language technology investment was to raise the profile of the Nordic language community and safeguard good Nordic language technology for the users. More specifically, this meant that three goals were given for supporting research and research-based education.

- Improved communication between the Nordic language technology researchers
- Improved cooperation on PhD education
- Establishing facilities or documentation centers to ensure the availability and reusability of research results, text collections and tools

In order to achieve these goals, three specific priority areas were selected:

- Computer-Aided Language Learning for Nordic Languages
- Cross-Lingual Information Management for the Nordic Languages
- Natural Language Human-Computer Interaction

For this purpose a budget of approximately 5 MDKK annually during 2000-2004 totaling 23.278.500 DKK was allocated, i.e. approximately 3.1 MEUR.

## Nordic Countries

In addition to efforts on the Nordic level, the investment in the specific Nordic Countries have also been significant but of varying magnitude. An estimate of the research investments by the Nordic Countries was collected from the public databases and records available on the internet. The projects were verified by the various persons contributing to this report. However, in order to make the figures comparable in the Nordic countries only external state-funding of university-lead research projects were included, i.e. business contributions or university budget contributions were not included. Nor did we include EU projects. All together, the Nordic Countries have financed research projects directly to the amount of 24 MEUR during 2003-2005. The period was chosen because public records in some form or another were available for all the Nordic Countries for this period.

### Denmark

In Denmark, the Ministry of Science, Technology and Innovation finances language technology research under The Danish Agency for Research, Technology and Innovation which performs secretariat functions for a number of independent councils. The two main councils for financing language technology research are the Danish Council for Independent Research and Danish Council for Strategic Research. During the period 2003-2005, Denmark spent

- approx. 2.6 MEUR mainly on text-based [Danish language technology research](#).

## **Finland**

In Finland, the two main Agencies for research funding are the Academy of Finland (Academy) and the Finnish Funding Agency for Technology and Innovation (TEKES). The Academy is financed by the Ministry of Education and TEKES is financed by the Ministry of Trade and Industry. During the period 2003-2005, Finland spent

- approx. 6.3 MEUR with the main emphasis on speech-based [Finnish language technology research](#).

## **Iceland**

In Iceland, the period 2003-2005 saw investments of

- approx. 0.7 MEUR emphasizing basic tools and resources for text-based [Icelandic language technology research](#).

## **Norway**

In Norway, the main financing body of university-lead research is the Norwegian Research Council. During the period 2003-2005 Norway had a ongoing strategic research program for language technology "Kunnskapsutvikling for norsk språkteknologi (KUNSTI, 2001-2006)", which accounts for 70 % of the funding under the chosen period of comparison. In addition, Norway also had independent projects. During the period 2003-2005, Norway spent

- approx. 9.2 MEUR covering both text-based and speech-based [Norwegian language technology research](#).

## **Sweden**

In Sweden, the funding is diverse with the main funding Agencies being The Swedish Research Council, The Swedish Governmental Agency for Innovation Systems (VINNOVA) and to some lesser extent the Knowledge Foundation. A strategic investment in LT was concluded in Sweden before the period which we are currently focusing on. During the period 2003-2005, Sweden spent

- approx. 4.8 MEUR mainly on text-based and to some lesser extent on speech-based [Swedish language technology research](#).

---

# LT Policy

## Current situation in 2006

There have been independent research programs in the Nordic countries dealing directly or partly with LT. They have mostly been modest in size and national in scope, and an approximation of their magnitudes is presented in the [Background](#) of this report. The most significant expression of a **Nordic LT policy** has been the **Nordic Language Technology 2000-2004** program of the Nordic Council of Ministers and the **Nordic Graduate School of Language Technology** (NGSLT). The most valuable result of these initiatives has been the creation of networks and contacts. Only a minor part of the funding of the Nordic activities was directed towards improving LT resources or promoting LT research. One can safely claim that there hardly exists a common Nordic LT policy at present and even rather modest ones at the national level. The reason might be the belief that LT will function on commercial terms alone after an initial public funding in each country. The following claims pertain to the lack of LT policies:

- Lack of LT threatens the survival of smaller language communities, because culture is transmitted via language. Some of the small communities, such as the Sámi people, have realized this and acquired LT funding but many others, including the Nordic main languages are more passive in this respect.
- Lack of large language resources is an obstacle for preserving our cultural diversity.
- Even when useful resources are collected or created, they remain inaccessible for the developers and researchers.
- Diversity in standards and incompatible technical methods scatter our efforts to create resources in a Nordic context.

There must be clear reasons for the decision makers to make **commitments** and take the necessary measures, i.e. understanding why and what has to be done and that the actions are worth the investment as was discussed earlier. Some possible **motivations** are:

- **Survival of our languages and cultural identity.** Cultural identity depends on language, and will be lost to a great extent if English conquers most of our daily life. In addition, local languages lose their prestige if they are useless in many situations. A language with no perceived prestige for its speakers erodes within a few generations. Governments may decide upon policies where local cultures and languages fade away, but they must do so openly and explicitly.
- Being **the first to master** and adopt multilingual LT technologies may open the path to success, not only within LT-related companies in the Nordic area, but also for a **wider spectrum of local software industry** which has a competitive advantage with multilingual LT technology readily available. Localization and internationalization are still difficult when more than canned translations are needed. With appropriate actions, Nordic Small and Medium

Sized Enterprises (SMEs) will have equal opportunities and a clear advantage world-wide.

In the past, especially before the Nordic research program for LT 2000-2004, the **coordination of the LT research, training and business community** was extremely informal materializing mostly **every second year at the NODALIDA conference** for two days. The participants of the NODALIDA 2005 decided to establish an association for speech and language technology which will be called **NEALT (Northern European Association for Language Technology)**. Such an association would be ideal for coordinating various initiatives and networking. Among other things, it intends to publish an electronic scientific journal.

## Vision for 2016

In 2016, **multilingualism is perceived as a strength** of the Nordic/Baltic region. Nordic and Baltic languages are small language communities, but we help **maintaining and protecting our local languages** by being in the forefront of LT. LT strengthens the Nordic/Baltic languages and the Nordic language community in a multilingual world **including not only the official Nordic/Baltic languages but also the minority languages, sign languages and immigrant languages**. The public administrative bodies of the Nordic countries take their information dissemination task seriously: **public information is freely and openly available** and disseminated in several languages **with the help of LT**. LT adds to the democratic participation in public life through eSociety, where the benefits of LT is for everyone regardless of language, gender, class, ethnic origin, cognitive or physical abilities, linguistic or technical competence, area of activity, etc.

The **smaller language communities in the region are able to participate in the LT development** with external support complementing the national funding for building the necessary resources. Tools and methods developed for the challenges of multilingualism beyond Nordic languages benefit the LT situation for the Nordic languages by making the Nordic languages part of the global body of languages with internationally compatible formal descriptions, i.e. the strength of multilingualism (including typologically diverse languages) can be harvested as a foundation for globally applicable LT through our long tradition of linguistic research. The **language care tradition of the Nordic countries have strong support** from language users and bodies, both publicly funded, e.g. the Research Institute for the Languages of Finland (Forskningscentralen för de inhemska språken, Kotus), Svenska akademien, Svenska språknämnden, and industrially funded bodies, e.g. TNC, with no legislative but with an established and accepted status on questions about language usage.

## Recommendations

We need to raise awareness of LT as a key factor for making languages survive and flourish. It does not matter whether the LT is academic, open source or commercial, as long as it exists and its resources are compatible and available for building large systems and applications. Small language communities will not get LT on a commercial basis alone, so most (or all) languages in the area need at least some public support and many will be totally dependent on it.

**At the Nordic level, we need to establish recommendations** for the actions on the national level, e.g. for national governments, funding organizations, research organizations, commercial players and individuals. A useful agent in the preparation and dissemination of such recommendations could be a Nordic (or Nordic/Baltic) association such as **NEALT**. Action areas, where Nordic funding is needed instead of national funding, are:

- establishing and starting NEALT and establishing a scientific electronic journal by NEALT,
- some form of continuation for the Nordic LT documentation centers, see awareness under [LT Training and Education](#),
- some continuity for the [NGSLT](#) by NordForsk, see [LT Training and Education](#), and
- individual small-scale projects (possibly carried out and coordinated by NEALT) e.g. to prepare more detailed recommendations for
  - altering the legislation of intellectual property rights (IPR, see [LT Legislation](#)),
  - guidelines for funding agencies to guarantee access and reuse of LT resources created with public funding (see [LT Research and Development](#)), and
  - guidelines for research and/or commercial use of dictionaries and word lists created as part of publicly funded dictionary compilation (see [LT Resources](#)).

Comment:

- *"Det behövs en nordisk samorganisation som arbetar med språkteknologisk infrastruktur och ser till att man inventerar, samlar, informerar om, tillgängliggör, utvecklar och tillhandahåller nödvändiga resurser både för språkteknologiska och språkvetenskapliga ändamål. Här ska finnas bred kompetens, också inom juridiska frågor om upphovsrätt, licensavtal m.m. Man ska ha tydliga angivelser av standarder för format och teknik. Man ska kunna bedöma och säkra kvaliteten på resurser och produkter. Organisationen ska ansvara för utvecklingen av en nordisk språkbank med gemensamma språkteknologiska resurser i samarbete med nationella centra för språkteknologisk och språkvetenskaplig infrastruktur. I Sverige har följande aktörer en viktig roll i ett sådant samarbete: GSLT, Svenska språknämnden (blivande Språkrådet), Språkbanken, SICS, dokumentationscentret Språkteknologi.se (med i Nordoknet), Vetenskapsrådet och Vinnova. Särskilt viktiga områden är informationssökning/-hantering (inte minst vid hanteringen av språkdata-baser), multimodala dialogsystem, översättning och språk- och skrivundervisning." -- Rickard Domeij*

To assess the situation for language-specific and language-independent resources for the languages in the area, a **Basic Language Resource Kit (BLARK) report for the Nordic languages** should be prepared and the most urgent gaps in availability of corpora should be filled in using national funding with **cooperation on the Nordic level for exchanging best practices**, whereas **gaps in tools and methods could be filled in using funding on a Nordic level** (see [LT Resources](#)). There are plenty of gaps and they must be filled with public funding in most cases. Some languages exist

in several countries and it is especially important that the allocated resources be coordinated on a Nordic level for these languages.

The Nordic region needs to stay abreast with the development in the EU in order not to duplicate efforts and focus on the aspects that are specifically Nordic. For this purpose it is important to **keep contact with organizations like CLARIN**, whose aim is to establish an integrated and interoperable research infrastructure of language resources and its technology by lifting the current fragmentation, offering a stable, persistent, accessible and extendable digital language infrastructure.

Comment:

- "Språkteknologin har betydelse för att ta fram digital infrastruktur för hela det humanvetenskapliga (och till viss del också det socialvetenskapliga) forskningsområdet. Språkteknologin kan bidra med metoder och verktyg för att samla in, strukturera, märka upp, lagra, hantera och tillgängliggöra stora digitala text- och taldata-baser med betydelse för många discipliner som språkvetenskap, litteraturvetenskap, filosofi, filologi m.m. Språkteknologin kan dessutom bidra med kunskaper om hur man hittar och söker i dessa. CLARINs vision är att språkteknologin ska få en sådan nyckelroll för den humanvetenskapliga forskningen inom EU. Det skulle förändra synen på språkteknologi som ett udda och marginellt område till ett angeläget område med konsekvenser för den humanvetenskapliga forskningen framåtskridande. Det här är något som innebär stora möjligheter också för nordisk språkteknologi." -- Rickard Domeij*

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
NEALT start-up	50 kEUR	NMR for funding	association
BLARK Report	10-25 kEUR per language	NorDokNet, NEALT	national projects coordinated at the Nordic level

---

# LT resources

## Current situation in 2006

On the whole, there is a shortage of adequate LT resources both in terms of their quantity and quality. There are not enough speech and text corpora, especially those with proper annotation, i.e. treebanks. Programs or LT modules exist for many languages, but they are incompatible. Some necessary tools for building LT modules and parsers are not available or they have severe restrictions on their use. On the whole, the environment is far from favorable for LT research and product development.

Language resources are an essential part of the LT infrastructure, and they are necessary for building further parts of the infrastructure. Corpora and dictionaries are necessary and useful in building parsers and analyzers, and they are equally useful for statistically oriented and rule based LT methods whether they are used for academic or commercial purposes. The language resources are also needed for creating new applications and products. Furthermore, language resources are often needed for evaluating the performance and quality of applications and systems.

In most countries, there are few public funding channels suitable for building LT infrastructure and LT resources, because building LT resources are neither like machinery nor equipment, nor are they comparable to commercial product development, nor even like usual basic research. LT infrastructure is more like ongoing public service processes or road building and maintenance, so **new forms of funding are needed**.

Comments:

- Obstacles are the availability of adequate language resources and the access to existing language resources.
- The proprietary nature of many LT resources for the region's languages is a major weakness: language processing resources as well as lexica and other databases are only made available to a few persons and groups, often at very high price levels (remarkably, this also applies to resources that have been developed with public funding).
- Existing resources are not necessarily adapted to LT purposes.
- For further development, we need willingness to fund and maintain and renew already established resources.
- An infrastructure to support the distribution of the language resources will also be needed, it may be centralized or distributed, but it has to be set up. This could be a Nordic effort, or it could be done at a European level (e.g. by making special agreements with ELRA, or by joining other initiatives).
- It is also important to assist smaller language communities in building basic resources.

Currently there are ongoing efforts to create open-source runtime support for LT modules, e.g. spellers for OpenOffice. We need additional efforts to **create open source tools for building LT modules**. The LT modules built with open source tools can either be proprietary or open source.

Comments:

- Business friendly open source alternatives such as MIT or LGPL licenses should be promoted.
- We should remember that open source does not necessarily imply free of charge, it only implies access to the source code.
- When financing research, it is important to have explicit requirements on making the results and resources available.
- To be able to share information and speed up development the infrastructure development needs to be accompanied by analysis software and methods for easy access.
- The announcement of an open source project does not necessarily create a community of users to take part in the development, and national funding programmes would not be sufficient to support 'various application areas', so one or two focused projects that invites (i) public funding, (ii) private funding, and (iii) public interest (i.e. a community of 'volunteers'). An example may be something like a talking robot that any user could teach new words, or new languages.
- A coordinating function is an important prerequisite for organizing cooperation and conflicts of interest between researchers, industry, and IPR owners when making resources publicly available.
- Assessing quality and quality assurance of LT resources and products are underdeveloped disciplines.

#### **LT modules**

**Parsers, analyzers, taggers, recognizers, generators** and other LT modules exist for major Nordic languages - for some languages there are even several competing modules. Most of them are proprietary and some can be licensed either for academic use or for commercial use - but usually as binaries which cannot and may not be modified. For different applications and for research, the ability to modify and tune would often be necessary. There seem to be excessive obstacles in the further development and integration of LT modules.

Comments:

- The LT modules are often incompatible with each other using different application programming interfaces and different tags and tagging principles.
- The further development and variation of existing LT modules for research or production purposes is mostly possible only for the owner who usually has no interest to develop the product further at its own cost and initiative. Development may be possible if a customer pays the costs.
- Using LT modules in different applications might require changes or further developing, but this may result in a stalemate.

- Even if the source-code is available, the LT modules are often built on different principles, using different tools.
- Applications for a wide Nordic audience presuppose that LT modules are developed for the smaller Nordic communities (Greenlandic, Faroese, Sámi, etc.)

### LT tools

The tools include generic programs for building parsers, analyzers, taggers, recognizers, generators and other LT modules. Several tools represent substantial development efforts, sometimes up to 100 person years. Currently, many widely used LT tools are proprietary. Open source tools exist, but they represent lesser efforts (maybe 2 to 5 person years per tool). Even if they are less complete and mature, their availability is guaranteed with no time limits and there are no restrictions on the use of LT modules created with them.

### Comments:

- There are no guarantees for the **long term availability** of proprietary tools. Even big companies may lose their interest in them while still preventing others from getting them. In the worst case, those companies may go bankrupt, and it may become extremely difficult or impossible to extend the licenses.
- SMEs do not have enough capacity to develop good LT tools or compile full dictionaries themselves even for official languages, not to mention languages for smaller communities.
- We lack learner tools and tools adapted to the requirements of the mobile handset industries.
- Proprietary solutions and tools will always exist, and innovative applications will often require that new tools and methods are developed.
- One reason why the tools are incompatible is that we disagree on what is the best solution, but the disagreement shrinks as the functionality criterion grows in importance.

### LT corpora and treebanks

Corpus resources include at least written language corpora, speech corpora, and multimedia corpora combining text and/or speech with video recording. Corpora may contain annotation to varying degrees including e.g. morphological, syntactic and pragmatic information. All Nordic corpus and treebank collections are modest in their volume. Some languages lack treebanks almost entirely.

### Comments:

- Parallel texts and corpora (raw as well as annotated) are important because they are necessary in order to further develop or evaluate monolingual and multilingual lexicons, taggers, parsers, and many other resources and tools.
- Currently one of the most significant obstacles is lack of linguistically annotated data.
- Large annotated and manually checked corpora with e.g. syntactic and semantic information are scarce or non-existent.

- Linguistic research is needed on spoken language varieties (registers, dialects, non-native) and on non-standard written varieties (computer-mediated communication, non-native, borderline literate)
- Availability of other language resources, i.e. huge amounts of speech and text, are needed.

### LT lexicons

Lexicons contain lexical information. In simpler cases they are just word lists containing entry words from some (possibly printed) dictionary and their part-of-speech and inflectional codes. Sometimes the full text of the word definitions is included. Dictionaries may be monolingual or bilingual. Publishers and compilers of dictionaries usually do not provide their dictionary material for academic purposes, because they fear that electronic copies of their dictionaries might be used for competing products or publications. On the whole, the lack of electronic dictionaries with sufficiently free terms for modification is severe.

### Comments:

- To the extent that there are proprietary lexicon resources, it should be considered if, and how (and to what extent) such resources can be made publicly available.
- SMEs do not have the capacity to develop tools or dictionaries on their own even for official languages, not to mention languages for smaller communities.
- Dictionaries for LT research and LT module development must often be created from scratch (and they remain less comprehensive). Current methods in LT can make the collection of dictionary content easier, but still, the duplication is a waste of effort.
- Lexicon development should be done with speech technology in mind, i.e. lexicons should include phonetic information, such as a phonetic transcriptions and stress.

## Vision for 2016

In 2016, a common understanding has been reached about the **domain of LT infrastructure** vs. applications and products, and an understanding of the **roles of the public and commercial sectors** has been established. The public sector has found ways to allocate the necessary and sufficient funds to develop the resources of the LT infrastructure. **A relevant infrastructure has been developed** for both text and speech to cover all languages and dialects in the region, and the data has been properly annotated at all levels. Building on the open-source lexicons and open-source tools, **the next step would naturally be to harmonize these resources to really benefit from one another.**

## Recommendations

The most obvious and substantial investment would be to create an appropriate infrastructure which has the sufficient LT resources for relevant languages of the area in such a manner that they can be used freely both for research, training and for

creating commercial products. The function of assessing quality and setting up quality standards should be part of the coordination and reviewing work by NEALT.

Based on the assessment of the situation in a **Basic Language Resource Kit (BLARK) report for the Nordic languages** the most urgent gaps in availability of corpora should be filled in using national funding with **cooperation on the Nordic level for exchanging best practices**, whereas **gaps in tools and methods could be filled in using funding on a Nordic level**. In addition, one should consider **opening up language resources on all levels (lexicons, grammars, written language corpora and speech corpora, etc.)** which have been created through public funding.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Basic Language Resource Kit	5-10 MEUR per language	Universities, research institutes, industry, NEALT	National projects coordinated at the Nordic level, exchange of researchers

These investments can be further subdivided into the areas related to LT modules, LT tools, LT corpora and LT lexicons.

#### LT modules

Both commercially and academically created LT modules need compatibility and capabilities for reusing other modules and resources. Language-independent tools can be used for creating both kinds of modules, and common API interfaces make it possible to utilize module combinations in order to facilitate interoperable and multilingual products and systems.

- Distributed openly available modules and APIs
- Interoperability of language modules and tools

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Openly available modules with common APIs	2-5 MEUR	open source community, universities, public and private institutions, NEALT	Nordic LT network

#### LT tools

Freely usable language-independent state of the art tools are needed so that investments in LT modules are not lost in the long term. Interoperable components and multilingual products and systems can be achieved through such tools. E.g. finite-state technology provides very efficient and modular implementations for a number of tasks.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Openly available LT	2-5 MEUR	Open source community, universities, public and private	Nordic LT network

tool

institutions, NEALT

### LT corpora

Speech and text corpora and their combinations are necessary starting points for many types of LT modules and applications. The required quantities have grown in magnitude. Different levels of annotation are necessary for various methods and research topics. The availability of corpus material is often too restricted excluding all commercial use and, at the same time, any development of LT modules. Changing the copyright **legislation** would make the collecting, annotating and sharing of corpora for research purposes more fruitful, see [LT Legislation](#).

**Model contracts** for collections of **copyright-protected corpora** should be created for all countries, and these model contracts should guarantee the necessary ways to use the materials including:

- sufficient rights for the end users to create LT modules and other results (which do not infringe on the copyright of the works),
- permission to create LT modules both for academic and for commercial purposes,
- ability to deposit the compiled corpus with one (or a restricted number of) computing centre(s) protecting the corpora from unauthorized access, and
- permission to use the corpora according to an agreement granted by the compiling party.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Model contracts	50 kEUR	Research organizations, lawyers, NEALT	Networking across countries
Corpus collection, written text	10-15 MEUR pr language	Universities, NEALT	Networking across countries
Corpus collection, spoken data	10-20 MEUR pr language	Universities, NEALT	Networking across countries

### LT lexicons

Dictionaries which have been developed with public funding ought to be published as open source material so that they can be used for creating LT modules such as parsers and analyzers. Lexemes including the part of speech and inflectional codes as well as other mark-up should be moved to the open source domain so that anybody can alter and make use of them for research or commercial purposes. More specifically, lists of headwords annotated with part of speech and inflectional class should be made available under very free conditions permitting their use in both academic and commercial contexts. The full text of dictionaries published as books may be reserved for academic use, but there must not be limitations on further use of methods, rules or programs which have been developed using such material, provided that they do not contain parts infringing on the copyright of the original work.

Key Area	Magnitude of funding	Parties involved	Mode of cooperation
----------	----------------------	------------------	---------------------

	needed		
Lexicon development	10 MEUR per language	Universities, NEALT	Networking across countries

---

# LT Research and Development

## Current situation in 2006

Multilingualism and the interplay between academic and research parties makes the reuse and interoperability more difficult and demanding than what is customary in other environments. Obviously several aspects have to be taken care of:

- Awareness of existing standards, recommendations and standardization efforts should be promoted.
- Documentation of the resources and the annotation and coding used in them is vital.
- Standardization of resources and APIs, as well as tools for interchange and conversion of data from one format to another should be readily available.
- Knowledge and information for integrating LT with other technologies and design disciplines should be easily accessible.
- Lack of low cost language resources for most small languages is a major obstacle for both research and development.
- Lack of cooperation between different research groups is a weakness in the region (both nationally and regionally).

We need **stimulating LT research** for various application areas. National funding programs should provide the basis, and a Nordic/Baltic framework program for networking could provide the necessary regional infrastructure and communication.

Comments:

- Preference should be given to research funding that integrates all research groups in a given area for a given country, or the Nordic area rather than supporting a centralized funding approach.
- Sufficient funding for both long term (university) research and support for industrial development.
- Good progress in the LT field needs support for joint projects and networks on the Nordic level.
- In addition to open source, we also need open standards and publicly available APIs.

## Vision for 2016

In 2016, basic **tools and resources are available as open source** and provide a **platform for further innovation and new products** due to a **substantial economical effort** provided from the governments in the Nordic and Baltic countries. **Availability of necessary language resources** improves the quality of LT research and application development and **LT research and applications can develop freely** in several directions in a stimulating research and business environment. **Mono- and multilingual LT modules with uniform APIs for a wide array of languages are**

**smooth and easy to integrate** into software products and services. LT modules will be **integrated in multimedia systems** (e.g. aligned with video systems for video retrieval) and the **use quality of LT systems is high**, so that the citizens of the region are able to access software-mediated services in their mother tongue. **Permanent LT research and development forums** have been set up in the bigger Nordic countries in support of Nordic and Baltic languages with lesser volume in economic as well as human terms. For public funding of research and development projects, it is required that the projects either make the publicly funded efforts openly available or contribute resources to some ongoing open source software project.

## Recommendations

The academic funding institutions ought to adopt recommendations or rules concerning linguistic resources which will be (or have been) developed using public funding. It ought to be a normal requirement that the researchers make the linguistic resources (e.g. tools and annotated corpora) available for the rest of the research community with as free conditions or licenses as possible. There ought to be a common goal in all Nordic countries to collect, produce and make available linguistic resources using terms which allow both academic use and the use of the resources for creating language technological products, even commercial ones, provided that the resources are used within the limits of copyright laws. In addition we may need to open up language resources on all levels (lexicons, grammars, written language corpora and speech corpora, etc.) which have been created through public funding. Common interfaces and tools should be created in cooperation between both commercial and academic parties.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Recommendations for research result materials	50 kEUR	funding organizations, universities, NEALT	working groups
Joint effort for standardization	15 MEUR	universities and industry	Academia/industry collaboration
Basic technology research	15 MEUR	Universities	Joint programme, Researcher exchange, workshop, division of research tasks
R&D Funding	50-80 MEUR	Universities, Research institutes, industry	Nordic projects

The R&D funding can be further specified into various fields of services and applications for the society:

- (statistical) machine translation and automatic methods for multilingual information processing
- information retrieval

- public information tools adapted to the mobile life of users
- cross-language information retrieval (CLIR) tools, focused CLIR tools for recent immigrants
- bioinformatics
- speech technology in multimodal applications
- language learning

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
<i>Several</i>	5-10 MEUR per area	public bodies, research partners, industry	projects

---

# LT Training and Education

## Current situation in 2006

**All parties**, i.e. the researchers, teachers, students as well as developers of applications and products in commercial companies, **need to be aware of the basic possibilities of LT** and where to find resources, partners and other information. The information on contacts and references must be available with up to date facts and pointers. The Nordic area, especially the Nordic-Baltic area, is not so small that all parties would know each other in advance.

Comments:

- In several Nordic countries, formal language knowledge in schools has been a low priority over several decades, which may hamper LT development and market uptake in the long run due to lack of basic formal linguistics skills.
- Potential users in all sectors and walks of life must be convinced that LT is something they need. Only powerful demand from the public will make politicians prioritize the area in question.
- It is necessary to raise public awareness about the importance of LT in our daily lives, and to get commercial companies interested in LT research and development.
- We need commercial and industrial recognition of the advantages of LT and a broad involvement of these parties through all phases of development.
- Development and deployment of LT modules presupposes a technical staff with a high level of competency in computational linguistics.
- Documentation of language resources is a prerequisite, if they are to be open source. If the user does not understand the categories used, he/she will fail in the use of the data and in their further development.

Each Nordic (and Baltic) country is a rather small unit for creating curricula for Master's level and PhD level teaching in language and speech technology. Some have more established Bologna system Bachelor's and Master's level studies available, but perhaps equally many cannot offer such education in their own country. The first level PhD courses offered by the Swedish GSLT have actually been courses which could be part of a Master's program in LT, and they have been used by students from countries where LT is not offered at the MA/MSc level. By **adjusting the university teaching** to the needs, we may achieve better quality and wider availability of teaching and supervision on all special areas through cooperation at master's level teaching (perhaps as a Nordic/Baltic masters program beginning through cooperation between neighboring universities) and in a Nordic/Baltic PhD teaching network (NGSLT).

Comments:

- For fruitful cooperation involving all the Nordic languages, it is necessary to create some minimal common ground by funding exchange of education.

- One should reach people already working in the industry that will integrate LT modules, and universities must create programmes for lifelong learning in LT.
- There is a need for cooperation in master's level teaching - both cooperation between universities and countries, and also cooperation between different fields such as linguistics, computer science, statistics, etc.
- We should include the BA-level as well and try to develop common teaching material, compendia and curricula using the idea of a common core with local variations.

Two kinds of problems can be identified:

1. not enough students receive the training needed for development of the LT field and
2. unnecessarily much effort is needed for creating materials and delivering similar courses at different sites.

## Vision for 2016

In 2016, skilled IT staff has a **high level of LT competency** for careful tuning of the modules to the application context. There is **focus on language awareness and multilingual awareness in primary and secondary schools**, as well as better school training in analytical and formal aspects of native and foreign languages - as a prerequisite for a strong LT competency in the upcoming generation of application builders.

## Recommendations

As a part of the Nordic Language Technology Research Program 2000-2004, a [LT documentation centre](#) was established in each of the five Nordic countries. Some continuation for them is needed, either in conjunction with some world-wide effort such as the [LT world](#) or as a Nordic or Nordic-Baltic effort. In contrast to the previous effort, only a single implementation for collecting, storing and disseminating the data, would be preferable, possibly based on Wiki techniques. This would let the national units concentrate on keeping the info up to date and maintaining its accuracy. It would be quite natural to apply the best methods of LT to make this kind of information easier to access and use. Such a site might also be a showroom of the infrastructure, applications and products.

More cooperation is needed in academic training among the universities in the Nordic/Baltic region. A sufficient number of highly skilled PhDs and Masters ought to be trained to master the best skills and all countries and language groups should be participating, including the minorities and small communities:

- Coordinated PhD education: NGSLT
- Master's level education: Distance education, exchange programs for teachers and students, common curriculum, programming skills with LT competency
- A set of introductory distant learning courses on LT directed to commercial developers and decision makers in all Nordic and Baltic countries.

- Language awareness and formal language knowledge in schools: development and empirical studies in a cross-institutional framework
- Strengthen and modernize formal mother tongue training at all levels in education: national and Nordic support at the attitude level
- Popularization: Professional PR assignment, 'sell' the idea of diversity to a much wider audience

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Nordic LT documentation	1 MEUR	NMR, NEALT	network of LT documentation centres
Journal start-up	50 kEUR	NEALT, Nordisk Publiceringsnämnd	scientific electronic journal
Coordinated PhD education	1 MEUR	Nordic/Baltic universities	NGSLT
Master's level education	2 MEUR	Nordic/Baltic universities	Distance education, exchange programs for teachers and students, common curriculum
Distant learning courses for commercial developers	50 kEUR	Nordic/Baltic universities	Production of the material
Popularization	1 MEUR	R&D, Government, Industry, Secondary Education	Professional PR assignment

---

# LT Legislation

## Current situation in 2006

The copyright and other IPR legislation has been an obstacle for collecting research materials and sharing them for academic purposes. Schemes and model contracts exist for collecting text and speech corpora, but they are laborious to use and often limit the use of the materials. Some recent changes in copyright legislation have made it even more difficult to collect and digitize material (by forgetting research and development uses).

Patenting of computer programs and algorithms has become harmful for LT. Early publishing of research results and applying open source policies will help in part but do not fully solve the problem. Lots of careful study and new research is needed because some patents protect the most obvious ways to solve common problems. It is beyond the financial resources of researchers and the small and medium-sized enterprises to resolve software patent conflicts even if the patent is obviously invalid.

Comments:

- Current copyright law and IPRs are an obstacle to the creation of quality resources.
- LT modules require complicated and costly licensing.
- The tools for creating LT modules are difficult and costly to acquire.
- Many development efforts are in stand still, as others will not or cannot develop proprietary resources or products owned by a competitor.

## Vision for 2016

In 2016, there is legislation and an infrastructure where **text and speech corpora can be freely collected, annotated and used for the purposes of research and development**. The arrangements make it possible for any published source to be stored and processed for the purpose of creating research results and LT products without compromising the copyright of the source. In addition, patenting obvious ways of solving problems with programs is no longer possible, and such patents have been declared invalid.

## Recommendations

The survival of cultures and languages with a relatively small number of speakers depends on the ability to use the language in daily life. This depends more and more on the availability of LT. The development of LT tools depends on the availability of language resources such as corpora. The **copyright legislation should enable collecting, annotating and sharing of resources for research purposes**. Currently certain privileges are granted to a few national libraries to archive electronic copies of

books, journals etc. and similar privileges are needed for developing LT resources. E.g. the Finnish library for the blind has a privilege to make electronic copies of copyrighted materials for the purposes of that library. In a similar vein, it is recommended that the legislation be changed so that the collection of text and speech corpora for the purposes of research and production of LT tools is possible. The use of such corpus collections would be deemed to conform to the principles of copyright when no longer passages are republished. Changing the copyright **legislation** would make collecting corpora more productive by guaranteeing that corpora and annotated material are available for research and development purposes. Availability can be achieved either by allowing centres (such as national language banks) share materials with each other or by allowing individual researchers share them.

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
Preparation of changes in the legislation	10 kEUR	Relevant Ministries, Universities, NEALT	working groups

---

# LT Business Aspects

## Current situation in 2006

There are quite a number of small (and medium-sized) commercial enterprises in the Nordic and the Baltic area. Many of them have an academic or research origin. Few of them are capable of major investments in LT tools or resources.

The **roles of the public and the commercial sectors need clarification** and their cooperation and interplay should be strengthened. The public sector needs to know its responsibility and provide adequate funding and continuity. The commercial sector is essentially needed for creating some of the products and applications. The commercial entrepreneurs use the infrastructure for building products. The infrastructure and the applications and services must meet each other **in a well understood way** and there must not be significant gaps between the two. The following might be a guideline for this partition:

- **Long and medium term research** of LT is and will be funded by various public sources and part of it will contribute to the building of the infrastructure. The research feeds the industry with new methods and ideas for new applications.
- Short term **applied research** and **product development** is funded by the commercial side with possible partial support from the public industrial funding agencies.
- The development of the **LT infrastructure** ought to be coordinated and mostly funded by the public sector on open source principles with shared efforts from the commercial side. Collecting corpora for languages with so few speakers as the Nordic languages have is clearly a public matter for the local governments and the Nordic Council of Ministers. The initial investments in open source software tools of the infrastructure are a matter of public funding, but the later investments will be shared with the commercial players.
- Publicly funded resources are **freely available on equal terms for everybody**.
- The opportunity to be able to make money on LT IPRs must be protected to attract people and money to this field.

Comments:

- It is also important to increase cooperation between universities and research institutes on the one hand and private companies on the other.
- Few LT endeavors and LT entrepreneurial businesses have found the means to grow and prosper.
- Currently the market for LT is small. We need to develop viable business models.

- If LT is to be a viable option for attracting talent and funds, the business potential will need to be developed and represent an interesting enough prospect.

## Vision for 2016

In 2016, the availability of compatible LT modules and interfaces give the software industry and the service providers in the Nordic/Baltic region a competitive edge in the global market place, by **facilitating the process of tailoring products and services to language-specific requirements** in new international markets. The **Nordic language councils** continue their long and successful cooperation and have extended this to **cooperation with LT companies**. Applications develop freely in a business-friendly environment, but **applications to the benefit of people with special needs, e.g. the elderly and impaired may develop in a non-competitive environment with public support**.

The principles of open source are widely understood and various parties are aware of the practices. Commercial enterprises have adopted viable business strategies for living side by side with and benefiting from the **open source efforts**, which **are seen as an important part of the third sector** in the language communities of the Nordic/Baltic region. We have **viable business models for sustaining the LT business** despite small market sizes and the limited availability of common resources. The joint efforts in the Nordic countries have resulted in **healthy industries that can support applications** in all Nordic languages with a command of spontaneous spoken interaction.

## Recommendations

The licensing conditions of LT resources must allow and encourage both their commercial and academic use. Medium term applied research projects together with industrial partners should continue. Funding should be provided for creating and purchasing LT applications and services for the public sector. This funding is intended to stimulate the LT service and application market by allowing for competition (and possible cooperation) among commercial players while aiming for real and useful public service. Such services could include more ambitious goals using LT-enhanced applications.

- Web services: tool sharing, hosted products
- LT module distribution

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation
LT module uptake	5 MEUR	industry, universities and language councils	Action plan managed at Nordic level
Web services	5 MEUR	industry and universities	Academia/industry collaboration

---

# Initial Action plan

The aim of the report was to identify key areas, magnitude of funding, parties involved and modes of cooperation. However, we are still left with questions regarding further specification of the plans as well as priorities and time-frames within the 10-year period. Some answers have been sketched for the organization of the work, but more detail is needed as well as some further consideration of the division of national and Nordic funding. To implement the goals and to further specify the areas and their time-frames in the 10-year plan, we suggest the following steps in allocating resources:

1. Establishing NEALT and its working groups
2. Commissioning BLARK reports for the Nordic languages
3. Nordic funding for cooperation on LT training and education
4. National funding of medium-term applied research projects involving university and industrial partners

When the BLARK reports have been delivered, resources coordinated by NEALT should be allocated for

1. Nordic funding of LT tools according to the recommendations of the BLARK reports
2. Nordic and national funding of corpora, treebanks and lexicons based on the BLARK report recommendations

---

# Acknowledgements

We are grateful to the following persons for contributing time and comments to this Expert Panel Report. The original ideas and contributions of the persons below can be found in [an initial vision for LT in 2016](#) and its [prerequisites](#) as well as [current obstacles for LT development](#) and [general trends influencing LT](#) development and its applications. In the questionnaire we also asked for their [recommendations](#). The synthesis of these opinions is that of the editors of the report.

<i>Name</i>	<i>Affiliation</i>
<i>Knut Aasrud</i>	Microsoft Norway a.s.
<i>Lars Ahrenberg</i>	Linköping University
<i>Eckhard Bick</i>	University of Southern Denmark
<i>Lars Borin</i>	Dept. of Swedish Language and Språkbanken, Göteborg University
<i>Bernt A. Bremdal</i>	CognIT a.s, Norway
<i>Rolf Carlson</i>	KTH, Royal Technical University, Stockholm
<i>Rickard Domeij</i>	Svenska språknämnden
<i>Tron Espeli</i>	Research Council of Norway (Innovation Division)
<i>Björn Gambäck</i>	SICS, Swedish Institute of Computer Science, Stockholm
<i>Arnor Gudmundsson</i>	Ministry of Education, Science and Culture, Norway
<i>Henrik Holmboe</i>	Aarhus School of Business
<i>Timo Honkela</i>	Helsinki University of Technology
<i>Jan Hoel</i>	The Norwegian language council
<i>Janne Bondi Johannessen</i>	University of Oslo
<i>Jussi Karlgren</i>	SICS, Stockholm
<i>Kimmo Koskenniemi</i>	University of Helsinki
<i>Mikko Kurimo</i>	Helsinki University of Technology, Finland
<i>Per Langgård</i>	Oqaasileriffik, Greenland
<i>Krister Lindén</i>	University of Helsinki
<i>Bente Maegaard</i>	University of Copenhagen
<i>Sjur Nørstebø Moshagen</i>	Sámi Diggi
<i>Joakim Nivre</i>	Växjö University and Uppsala University
<i>Torbjørn Nordgård</i>	NTNU Trondheim, Norway
<i>Mikael Reuter</i>	Forskningscentralen för de inhemska språken, Finland
<i>Eiríkur Rögnvaldsson</i>	University of Iceland
<i>Koenraad de Smedt</i>	University of Bergen
<i>Torbjørn Svendsen</i>	NTNU Trondheim, Norway
<i>Trond Trosterud</i>	University of Tromsø, Norway
<i>Martti Vainio</i>	Department of Speech Sciences, University of Helsinki

We are also grateful to the Nordic Council of Ministers for sponsoring the Department of Linguistics at the University of Helsinki when working on the Report.

---

## References

Nordisk Sprogteknologisk Forskningsprogram 2000-2004. Epilog. Editor: Henrik Holmboe. Copenhagen.

Nordisk Sprogteknologi 2001/2002/2003/2004/2005. Editor: Henrik Holmboe. Copenhagen.

Språk i Norden 2006. Språkenemdene i Norden. Oslo.

Eckhard Bick. [LT-tools such as parsers and corpora for 8 languages](#) (Research tools). [http://beta.visl.sdu.dk]

Alea M. Fairchild and Bruno de Vuyst. 2004. [Hot Spot Implosion: The Decline and Fall of Flanders Language Valley](#). [http://portal.acm.org/citation.cfm?id=962756.963192]

[Survey of the State of the Art in Human Language Technology](#). Eds. Ron Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Batista Varile, Annie Zaenen, Antonio Zampolli, Victor Zue. Cambridge University Press and Giardini 1997. [http://www.dfki.de/~hansu/HLT-Survey.pdf]

Steven Krauwer. 2003. [The Basic Language Resource Kit \(BLARK\) as the First Milestone for the Language Resources Roadmap](#). [http://www.elsnet.org/dox/krauwer-specom2003.pdf]

Joakim Nivre and Koenraad de Smedt and Martin Volk. 2005. [Treebanking in Northern Europe: A White Paper](#). [http://ling.uib.no/desmedt/papers/whitepaper-yearbook2004.pdf]

Hans Uszkoreit. [Language Technology. A First Overview](#). Accessed 2006. [http://www.dfki.de/~hansu/LT.pdf]

## Nordic Organizations

[NordForsk](#) - Institution for Nordic cooperation within research and research training. [http://www.nordforsk.org/index.cfm]

[NorDocNet](#) - Nordic network of documentation centers for language technology. [http://www.nordoknet.org/]

[NGSLT](#) - Nordic Graduate School of Language Technology. [http://ngslt.org/]

## Language Policy Documents

### Denmark

[Sprog- og Taleteknologi](http://www.vtu.dk/cgi-bin/theme-list.cgi?theme_id=9835). Ministeriet for Videnskab Teknologi og Utveckling, Danmark. [http://www.vtu.dk/cgi-bin/theme-list.cgi?theme\_id=9835]

[Strategisk satsning på dansk sprogteknologi. 2005](http://forsk.dk/pls/portal/docs/PAGE/FORSKNINGSSTYRELSEN/FORSKNINGSS TYRELSEN_FORSIDE/DET_FRIE_FORSKNINGSRAAD/FORSKNINGSRAADE T_KULTUR_KOMMUNIKATION/FKK_PUBLIKATIONER/STRATEGISK%20S ATSNING.PDF). Forskningsrådet for Kultur og Kommunikation, Danmark. [http://forsk.dk/pls/portal/docs/PAGE/FORSKNINGSSTYRELSEN/FORSKNINGSS TYRELSEN\_FORSIDE/DET\_FRIE\_FORSKNINGSRAAD/FORSKNINGSRAADE T\_KULTUR\_KOMMUNIKATION/FKK\_PUBLIKATIONER/STRATEGISK%20S ATSNING.PDF]

## **Finland**

Kieliteknologia Suomessa (*Language Technology in Finland*). Ed. Manne Miettinen, Report No. R02/98, CSC.

[Kieliteknologian koulutuksen laajentaminen](http://www.ling.helsinki.fi/users/koskenni/kieliteknologia/opm-raportti.html) (*Extending Language Technology Education*), Report No 23:1999. Ministry of Education. [http://www.ling.helsinki.fi/users/koskenni/kieliteknologia/opm-raportti.html]

[Puheentutkimuksen resurssit Suomessa](http://www.csc.fi/raportit/puhe/), (*Speech Research Resources in Finland*). Eds. Manne Miettinen and Juhani Toivanen. 2001. CSC. [http://www.csc.fi/raportit/puhe/]

## **Norway**

[Planer og utredninger](http://www.sprakradet.no/templates/Page.aspx?id=684). Språkrådet, Norge. [http://www.sprakradet.no/templates/Page.aspx?id=684]

[Norsk språkbank](http://www.sprakrad.no/templates/Page.aspx?id=685). Språkrådet, Norge. [http://www.sprakrad.no/templates/Page.aspx?id=685]

[Artiklar og utgreiningar](http://www.sprakradet.no/templates/Page.aspx?id=3166). Språkrådet, Norge. [http://www.sprakradet.no/templates/Page.aspx?id=3166]

## **Sweden**

[Språkpolitiska dokument](http://sprakteknologi.se/dokument). Språkteknologi.se. [http://sprakteknologi.se/dokument]

# APPENDIXES

## Invited Experts for the Expert Panel Report

Country		
Person	Agency	E-mail
<b>Denmark</b>		
Grete Kladakis	Danish Agency for Science, Technology and Innovation	<a href="mailto:gk@forsk.dk">gk@forsk.dk</a>
Sidse Ægidius	Ministry of Science, Technology and Innovation, Dep. International ICT policy	<a href="mailto:sae@vtu.dk">sae@vtu.dk</a>
Jørn Lund	Det Danske Sprog- og Litteraturselskab	<a href="mailto:jl@dsl.dk">jl@dsl.dk</a>
Bente Maegaard	Center for Sprogteknologi	<a href="mailto:bente@cst.dk">bente@cst.dk</a>
Børge Lindberg	Aalborg Universitet, Taleteknologi	<a href="mailto:lindberg@cpk.auc.dk">lindberg@cpk.auc.dk</a>
Henrik Holmboe	Aarhus School of Business	<a href="mailto:hh@asb.dk">hh@asb.dk</a>
Eckhard Bick	Aarhus University	<a href="mailto:lineb@hum.au.dk">lineb@hum.au.dk</a>
Sabine Kirchmeier Hansen	Copenhagen Business School	<a href="mailto:ska@id.cbs.dk">ska@id.cbs.dk</a>
Daniel Hardt	Copenhagen Business School	<a href="mailto:dh@id.cbs.dk">dh@id.cbs.dk</a>
Frans Gregersen	Københavns Universitet	<a href="mailto:fg@hum.ku.dk">fg@hum.ku.dk</a>
Per Langgaard	Oqaasileriffik - Grønlands sprogsekretariat	<a href="mailto:pela@gh.gl">pela@gh.gl</a>
Hulda Zober Holm	Nordic Council of Ministers	<a href="mailto:hzh@norden.org">hzh@norden.org</a>
<b>Finland</b>		
Marja Granlund	Finansministeriet, Avdelningen för utvecklande av förvaltningen	<a href="mailto:marja.granlund@vm.fi">marja.granlund@vm.fi</a>
Kristiina Pietikäinen	Ministry of Transport and Communications	<a href="mailto:kristiina.pietikainen@mintc.fi">kristiina.pietikainen@mintc.fi</a>
Anita Lehikoinen	Ministry of Education	<a href="mailto:anita.lehikoinen@minedu.fi">anita.lehikoinen@minedu.fi</a>
Mikael Reuter	Kotimaisten kielten tutkimuskeskus	<a href="mailto:mikael.reuter@kotus.fi">mikael.reuter@kotus.fi</a>
Gyrid Högman	Ålands lyceum	<a href="mailto:gyrid.hogman@lyceum.aland.fi">gyrid.hogman@lyceum.aland.fi</a>

Matti Sihto	TEKES - Finnish Funding Agency for Technology and Innovation	<a href="mailto:matti.sihto@tekes.fi">matti.sihto@tekes.fi</a>
Arto Mustajoki	Academy of Finland	<a href="mailto:anneli.pauli@aka.fi">anneli.pauli@aka.fi</a>
Kimmo Koskenniemi	University of Helsinki	<a href="mailto:kimmo.koskenniemi@helsinki.fi">kimmo.koskenniemi@helsinki.fi</a>
Lauri Carlson	University of Helsinki, Kouvola	<a href="mailto:lauri.carlson@helsinki.fi">lauri.carlson@helsinki.fi</a>
Martti Vainio	University of Helsinki	<a href="mailto:martti.vainio@helsinki.fi">martti.vainio@helsinki.fi</a>
Helena Ahonen-Myka	University of Helsinki	<a href="mailto:helena.ahonen-myka@cs.helsinki.fi">helena.ahonen-myka@cs.helsinki.fi</a>
Timo Honkela	Technical University of Helsinki	<a href="mailto:timo.honkela@hut.fi">timo.honkela@hut.fi</a>
Mikko Kurimo	Technical University of Helsinki	<a href="mailto:mikko.kurimo@helsinki.fi">mikko.kurimo@helsinki.fi</a>
Tero Ojanperä	Nokia	<a href="mailto:tero.ojanpera@nokia.com">tero.ojanpera@nokia.com</a>
<b>Iceland</b>		
Gudbjörg Sigurdardottir	Prime Minister's Office, Department of Information society	<a href="mailto:gudbjorg.sigurdardottir@for.stjr.is">gudbjorg.sigurdardottir@for.stjr.is</a>
Eiríkur Rögnvaldsson	Universty of Iceland	<a href="mailto:eirikur@hi.is">eirikur@hi.is</a>
<b>Norway</b>		
Torbjörg Breivik	Språkrådet	<a href="mailto:torbjorg.breivik@sprakradet.no">torbjorg.breivik@sprakradet.no</a>
Sylfest Lomheim	Språkrådet	<a href="mailto:lomheim@sprakradet.no">lomheim@sprakradet.no</a>
Bernt Erik Heid	The Research Council of Norway	<a href="mailto:beh@forskningsradet.no">beh@forskningsradet.no</a>
Tron Espeli	The Research Council of Norway	<a href="mailto:te@forskningsradet.no">te@forskningsradet.no</a>
Eivind Lorentzen	Ministry of Trade and Industry	<a href="mailto:eivind.lorentzen@nhd.dep.no">eivind.lorentzen@nhd.dep.no</a>
Fred-Arne Ødegaard	Fornyings- og administrasjonsministeriet	<a href="mailto:Fred-Arne.Odegaard@fad.dep.no">Fred-Arne.Odegaard@fad.dep.no</a>
Espen Dennis Kristoffersen	Fornyings- og administrasjonsministeriet	<a href="mailto:espen.dennis.kristoffersen@mod.dep.no">espen.dennis.kristoffersen@mod.dep.no</a>
Risten Aleksandersen	Sámediggi	<a href="mailto:risten.aleksandersen@samediggi.no">risten.aleksandersen@samediggi.no</a>
Torbjørn Nordgård	Norwegian University of Science and Technology	<a href="mailto:torbjorn@hf.ntnu.no">torbjorn@hf.ntnu.no</a>
Torbjørn Svendsen	Norwegian University of Science and Technology	<a href="mailto:torbjorn@iet.ntnu.no">torbjorn@iet.ntnu.no</a>
Lars Hellan Norwegian	University of Science and Technology	<a href="mailto:lars.hellan@hf.ntnu.no">lars.hellan@hf.ntnu.no</a>
Jon Atle Gulla	Norwegian University of	<a href="mailto:jon.atle.gulla@idi.ntnu.no">jon.atle.gulla@idi.ntnu.no</a>

	Science and Technology	
Tor Andre Myrvoll	Norwegian University of Science and Technology	<a href="mailto:myrvoll@iet.ntnu.no">myrvoll@iet.ntnu.no</a>
Helge Dyvik	University of Bergen	<a href="mailto:helge.dyvik@lil.uib.no">helge.dyvik@lil.uib.no</a>
Koenraad de Smedt	University of Bergen	<a href="mailto:desmedt@uib.no">desmedt@uib.no</a>
Britt Helle Aarskog	University of Bergen	<a href="mailto:brit@ifi.uib.no">brit@ifi.uib.no</a>
Gisle Andersen	University of Bergen	<a href="mailto:gisle.andersen@aksis.uib.no">gisle.andersen@aksis.uib.no</a>
Janne Bondi Johannessen	University of Oslo	<a href="mailto:jannebj@hedda.uio.no">jannebj@hedda.uio.no</a>
Jan Tore Lønning	University of Oslo	<a href="mailto:jtl@ifi.uio.no">jtl@ifi.uio.no</a>
Stephan Oepen	University of Oslo, Stanford	<a href="mailto:oe@csl.stanford.edu">oe@csl.stanford.edu</a>
Trond Trosterud	University of Tromsø	<a href="mailto:Trond.Trosterud@hum.uit.no">Trond.Trosterud@hum.uit.no</a>
Bernt Bremdal	CognIT	<a href="mailto:bernt.bremdal@cognit.no">bernt.bremdal@cognit.no</a>
Bente Moxness	LingIT	<a href="mailto:bente@lingit.no">bente@lingit.no</a>
Knut Morten Aasrud	Microsoft Norge	<a href="mailto:knutaa@microsoft.com">knutaa@microsoft.com</a>
Bjørn Seljebotn	Nynodata	<a href="mailto:bjorn@nynodata.no">bjorn@nynodata.no</a>
Knut Kvale	Telenor Taleteknologi	<a href="mailto:knut.kvale@telenor.com">knut.kvale@telenor.com</a>
<b>Sweden</b>		
Staffan Jonson	Näringsdepartementet, Enheten för IT, forskning och utveckling	<a href="mailto:staffan.jonson@industry.ministry.se">staffan.jonson@industry.ministry.se</a>
Rickard Domeij	Språknämnden	<a href="mailto:Rickard.Domeij@spraknamnden.se">Rickard.Domeij@spraknamnden.se</a>
Ola Karlsson	Språknämnden	<a href="mailto:Ola.Karlsson@spraknamnden.se">Ola.Karlsson@spraknamnden.se</a>
Lars Borin	Göteborg University	<a href="mailto:lars.borin@svenska.gu.se">lars.borin@svenska.gu.se</a>
Robin Cooper	Göteborg University	<a href="mailto:cooper@gsht.hum.gu.se">cooper@gsht.hum.gu.se</a>
Kirsti Hansen	Göteborg University	<a href="mailto:kirsti.hansen@svenska.gu.se">kirsti.hansen@svenska.gu.se</a>
Rolf Carlson	KTH Talteknologi	<a href="mailto:rolf@speech.kth.se">rolf@speech.kth.se</a>
Lars Ahrenberg	Linköping University	<a href="mailto:lah@ida.liu.se">lah@ida.liu.se</a>
Björn Gambäck	SICS - Swedish Institute of Computer Science	<a href="mailto:gamback@sics.se">gamback@sics.se</a>
Jussi Karlgren	SICS - Swedish Institute of Computer Science	<a href="mailto:jussi.karlgren@sics.se">jussi.karlgren@sics.se</a>
Martin Volk	University of Stockholm	<a href="mailto:volk@ling.su.se">volk@ling.su.se</a>
Anna Sågvall Hein	University of Uppsala	<a href="mailto:anna@lingfil.uu.se">anna@lingfil.uu.se</a>
Joakim Nivre	University of Växjö	<a href="mailto:joakim.nivre@lingfil.uu.se">joakim.nivre@lingfil.uu.se</a>
Veikko Hara	TeliaSonera	<a href="mailto:veikko.hara@teliasonera.com">veikko.hara@teliasonera.com</a>

## Danish LT projects

*(Note. This is not necessarily an exhaustive list of projects, but it is the best we could do in the time available and initial feed-back confirms that it gives a fair view of the activities.)*

Denmark	2003	2004	2005	Project
Rigsarkivet	2000			Udvikling af emnebaserede søgemuligheder til Statens Arkivers samlinger
Handelshøjskolen i København	4500			Center for Computational Modelling of Language (CMOL)
Københavns Universitet	300			Tillægsbevilling til: Den medieafhængige tekst og den elektroniske boghandel
Syddansk Universitet	1700			Global kommunikation i danske virksomheder
Københavns Universitet	400			IDANNA - IDentifikation og ANonymisering af NAVne
Handelshøjskolen i København		750		Language technology derived from spoken language resources
Københavns Universitet		2500		Oversættelse fra leksem- til tekstniveau. Innovation via synergi mellem sprogteknologi og komparativ forskning inden for vesteuropæiske sprog
Københavns Universitet		3000		Dansk leksikalsk-semantic ordnet (DanNet)
Roskilde Universitetscenter		420		CONTROL: CONstraint based Tools for ROBust Language processing
Københavns Universitet			2718	Center for Computational Cognitive Modeling
Københavns Universitet			740	Vidensbaseret leksikalsk disambiguering
Sum kDKK	8900	6670	3458	Total <b>19.0</b> MDKK
Sum kEUR	1194	895	464	Total <b>2.6</b> MEUR

## Finnish LT projects

(Note. This is not necessarily an exhaustive list of projects, but it is the best we could do in the time available and initial feed-back confirms that it gives a fair view of the activities.)

<b>TEKES</b>	2003	2004	2005	Project
Helsinki University	250	290	660	Finnish Semantic Web Ontologies
Helsinki Institute of Information Technology		400	268	Search-INA-Box
University of Tampere	700	540	440	New Methods in Speech Technology
VTT		120	121	Rich semantic media for personal and professional users
Helsinki Institute of Information Technology		145		Intelligent Web Services
Helsinki University	320	300	320	Mobile and Multi-Lingual Maintenance Man
<b>Finnish Academy</b>				
Helsinki University of Technology	143	176		SA-PUHE - Integrated resources for speech technology and spoken language research
Helsinki University of Technology	33	33	33	Quality of speech in hands free communication
Helsinki University of Technology		40	40	Multidisciplinary studies of the production and perception of speech - Development of new linear predictive methods for parametric modelling of speech
Helsinki University	45	45	45	Ihmisen kuulojärjestelmän kognitiiviset prosessit: puheen tuottaminen ja havaitseminen
University of Turku	40	40	40	Akustinen ja perkeptuaalinen tutkimus äänenlaadun merkityksestä puheen tunneilmaisussa
University of Turku		62	62	Puheen tuoton ja havaitsemisen monitieteinen tutkimushanke - oppimisen ja aivojen plastisiteetin vaikutus puheen havaitsemiseen ja tuottoon
Helsinki University		35	35	Puheen prosessoinnin neuraalinen mallinnus
University of Tampere		81	163	Monikielinen dokumenttien haku ja hallinta sekä tehtäväkeskeinen tiedonkulku
Helsinki University of Technology	35	35		Tiläänen tuotto ja sen havaitseminen
Helsinki University of Technology	30	30	30	Coding and Modeling of Phonemes in Speech

University of Tampere	40	40	NLP-based information retrieval systems for the biological literature
Sum kEUR	1596	2412	2297
			Total <b>6.3</b> MEUR

---

# Icelandic LT projects

*(Provided by Eiríkur Rögnvaldsson)*

	2003	2004	2005	Project
University of Iceland	14800			Isolated word recognition system
University of Iceland			17100	Text-to-speech system
Institute of Lexicography	5100	1700		Full-form morphological database
Institute of Lexicography	3600	1200		Grammatical tagger
Institute of Lexicography		3700	5600	Grammatically tagged corpus
Sum kISK	23500	6600	22700	Total <b>52.8</b> MISK
Sum kEUR	354	117	252	Total <b>0.7</b> MEUR

# Norwegian LT projects

*(Note. This is not necessarily an exhaustive list of projects, but it is the best we could do in the time available and initial feed-back confirms that it gives a fair view of the activities.)*

<b>KUNSTI</b>	2003	2004	2005	Project
University of Bergen		774	1061	TREPIL: trebankpilotprosjekt
University of Tromsø		1222	1531	Disambiguering av morfologisk tagga samisk tekst
Norwegian School of Economics and Business Administration		865	789	KB-N: Kunnskapsbank for norsk økonomisk-administrativt domene
Norwegian University of Science and Technology, NTNU	1985	2612	2041	FONEMA - Metodeutvikling for naturtro norsk talesyntese
University of Bergen	624	1430	936	BREDT - Behandling av referensielle enheter i diskursteori
University of Oslo	4868	5869	6035	LOGON - Leksikon, Ordsemantikk, Grammatikk og Oversettelse til norsk
Norwegian University of Science and Technology, NTNU	5468	5837	6107	Brukergrensesnitt med naturlig tale - BRAGE
<b>Övriga</b>				
Norwegian University of Science and Technology, NTNU		416	574	Brukerorientert elektronisk pasientjournal
University of Oslo	70	1480	1524	SPRIK - Språk i kontrast
University of Oslo	3000	3000	3000	Parallellkorpusprosjektet ved HIT-senteret
University of Bergen	1364	160	160	Fra parallellkorpus til ordnett
University of Tromsø	647	160	160	Prosjekt for utvikling av samisk språkteknologi
<b>IKT-2010</b>				
Norwegian University of Science and Technology, NTNU	1179	2213	2363	VOCALS - konvergensen mellom kommunikasjonssystemer, avansert dialogkontroll og språkteknologi
Sum kNOK	19205	26038	26281	Total <b>71.5</b> MNOK
Sum kEUR	2465	3342	3373	Total <b>9.2</b> MEUR

## Swedish LT projects

(Note. This is not necessarily an exhaustive list of projects, but it is the best we could do in the time available and initial feed-back confirms that it gives a fair view of the activities.)

<b>Vetenskapsrådet</b>	2003	2004	2005	Project
Lund University			860	Svensk dialektsyntax (SveDiaSyn)
Lund University			500	Direkt Profil: Ett program för utvecklingsgångar och utvecklingsstadier i skriften inlärafranska
Lund University			650	Grammatik, Prosodi, Diskurs och Hjärnan. ERP-studier i språkbearbetning
Linköping University		539		Lingvistisk mikro- och makroanalys av en översättningskorpus
Göteborg University	670	696		Korpusbaserad Talspråksbeskrivning
Royal Institute of Technology	500	500		Fel och missförstånd i människa-maskindialogsystem
Umeå University	880	910		Gräns och gruppering - Strukturering av talet i olika kommunikativa situationer
Royal Institute of Technology	680	680	680	Svensk informationssökning med språkteknologi och matrisberäkningar
Royal Institute of Technology	1100	1100	1100	Språkliga datorstöd och andraspråksinläring
Växjö Universitet	608	608	608	Stokastiska dependensgrammatiker för grammatisk analys av naturliga språk
<b>KK-stiftelsen</b>				
Karolinska Institutet	385	385		REFTERM - Referensterminologi för vård, forskning och uppföljning - ITHS 2
<b>VINNOVA/IKT-användning</b>				
Skövde University		1079	1025	EKLär - Effektiv kunskapshantering och lärande i kunskapsintensiva verksamheter
Linköping University	698	1 410	1 629	SWEBPROD - Semantisk webb för produkter
Swedish Institute of Computer Science	572	884	971	FetchProt - Hämtning av information ur texter om proteiner
Linköping University	232	692	792	Från metadata till uppmärkning av komplexa document - Ett ramverk för semantisk documentproduktion
Uppsala University		1000	1000	Märkning av utbildningsinnehåll
Royal Institute of Technology	770	770		CrossCheck - Svensk grammatikkontroll för andraspråksskribenter
Linköping University	1200	1200		KOMA - Korpusbaserad

				maskinöversättning
Lund University	700	700		The role of function words in spontaneous speech processing
Royal Institute of Technology			50	Seminar: Research challenges in speech technology
Swedish Institute of Computer Science	1064	830		DUMAS - Dynamic universal mobility for adaptive speech interfaces
Lund University	1330	1140	740	Intelligenta komponenter i ett distribuerat digitalt bibliotek
Linköping University	900	900	450	SwebButler - Semantic webb services based on butler agents
Göteborg University	300	300		Swedish language technology documentation centre
Linköping University	330	330		Generiska resurser för språkteknologi
Linköping University	700	700		Multimodal interaktion för publika informationstjänster
Chalmers University of Technology	1200	1200		Interactive language technology
Uppsala University	225			En svensk systranmodul
Summa kSEK	15044	18553	11055	Total <b>44.6</b> MSEK
Summa kEUR	1615	1992	1187	Total <b>4.8</b> MEUR

---

# Weaknesses in or obstacles for LT development

*(Note. Some initially identified weaknesses were circulated among leading experts on LT in the Nordic countries and their comments on these can be found below.)*

Presently there are LT modules for most of the languages widely used in the Nordic/Baltic area. However,

- The LT modules are often **incompatible** with each other, built on different principles, using different tools.
- The tools for creating such LT modules are **difficult and costly to acquire** and there is no long term guarantee for the availability of the tools.
- **No common runtime code or application interface** for the Nordic/Baltic and the major world languages exist. For modules built with some proprietary tools, the runtime requires complicated and costly licensing.

The **further development and variation** of existing LT modules for research and production purposes is mostly possible only for the owner. Proprietary LT modules can be licensed for research and development purposes, but not improved or altered by the researchers or others.

**SMEs do not have the capacity to develop tools or dictionaries on their own** even for official languages, not to mention minority languages. Many efforts are in stand still, as others will not or cannot develop proprietary resources or products owned by a competitor.

## **Are there other significant obstacles you know should be removed to realize the vision? Are any of the above of lesser importance?**

*(Quotes in order of submission:)*

I wholeheartedly agree with the above. LT modules with clear interfaces are urgently needed. Moreover we need large annotated and manually checked corpora with syntactic and semantic information.

-- *Martin Volk*

Development and deployment of LT modules in different contexts presupposes a technical staff with a high level of competency in computer linguistics, a solid schooling in the LT modules capabilities and limitations, as well as profound knowledge of the application needs. Applications for a wide and inclusive Nordic audience presuppose that new LT modules are developed for the lesser languages

(Greenlandic, Faroese, Sámi, etc.)

-- *Koenraad de Smedt*

LT endeavours and LT entrepreneurial businesses have not found the means to grow and prosper. A solid business potential is currently not visible, outside certain areas where the public invest money to seed development and create tools to remedy problems. If LT is to be a viable option for attracting talent and funds, the business potential will need to be developed and represent an interesting enough prospect.

-- *Knut Aasrud*

Proprietary solutions and tools will always exist, and innovative applications will often require that new tools and methods are developed. Again, the most significant obstacle is lack of linguistic data for these languages, not tools and standardized APIs.

-- *Torbjørn Nordgård*

1. So called 'lesser used languages' e.g. minority languages in the Nordic countries do not have sufficient LT resources, not even in terms of data.
2. Copyright law and IPRs (or perhaps rather the actions of copyright holders) is an obstacle to the creation of quality resources.

-- *Lars Ahrenberg*

One problem is that some of the smaller language communities in the area still do not have all basic LT modules and resources. It is just as expensive to build these modules and resources for the small language communities as for the larger ones, and enough national funding for such development may not be available. For fruitful cooperation involving all the languages in question to be possible, it is necessary to create some minimal common ground, and that means that the smaller language communities need some external support in the beginning. This support can be in the form of direct funding from Nordic funds or programs, but it can also involve exchange of research and knowledge.

-- *Eiríkur Rögnvaldsson*

For further development, we need willingness to fund and maintain and renew already established resources.

-- *Henrik Holmboe*

I believe that we do not always know the existence of all language resources and tools, because there is no incentive to make such information available, and there is no common format (metadata) for it.

-- *Bente Maegaard*

Lack of low cost language resources for most small languages is a major obstacle for both research and development.

-- *Torbjørn Svendsen*

Speech tools. Learner tools. Tools adapted to requirements of the mobile handset industries (desktop interaction will continue to grow but at a lesser rate than other interaction modes!)

-- *Jussi Karlgren*

In several Nordic countries, formal language knowledge in schools has been a low priority over several decades. This can potentially affect the recruiting base for LT-related education and research in the adult Nordic community.

-- *Eckhard Bick*

What you say about tools is good. One reason why the tools are incompatible is that we disagree on what is the best solution. The disagreement shrinks as the functionality criterion grows in importance, though. The accessibility of linguistic resources is a further obstacle.

-- *Trond Trosterud*

It is necessary to convince politicians that LT is vital for the viability, and even survival, of 'smaller' languages, even more so today than only a few decades ago. In this context it is also of paramount importance that politicians with budgetary power are made to realize that coordinated and publicly financed efforts to accumulate large language resource banks are vital nationally, and that development and use of standardized and interoperable technical methods is a prerequisite in a Nordic context.

-- *Jan Hoel*

Upphovsrättsliga frågor och licensavtal är ett stort problem, särskilt vad gäller publicerat material i elektronisk form. Det finns ingen samlad information om vilka resurser som finns och hur de är tillgängliga. Mycket är heller inte anpassat för språkteknologiska ändamål. Det saknas bra metoder för att bedöma och kvalitetssäkra språkteknologiska resurser och produkter, särskilt ur språkligt perspektiv.

-- *Rickard Domeij*

Both 'the availability of adequate language resources' and 'the access to existing language resources' could also be listed as obstacles.

-- *Tron Espeli*

Lack of industrial recognition of Nordic language capacities does not provide the necessary focus for R&D, product development and standardization, which seems to be driven from abroad, primarily the US.

-- *Bernt A. Bremdal*

Small market for LT and a need to develop viable business models.

-- *Arnor Gudmundsson*

There are two kinds of incompatibility, one that has more to do with software and one that has more to do with (conceptualizations of) knowledge of language and linguistic interaction. Both must be addressed.

-- *Lars Borin*

The two main obstacle to progress in LT in the Nordic region have always been two:

1. the proprietary nature of the LT resources for the region's languages: language processing resources as well as lexica and other databases are only made available to a few persons and groups, and at what's often ridiculously high price levels (most amazingly, this also applies to resources that have been developed with public funding), and

2. the lack of cooperation between different research groups in the region (both nationally and regionally).

-- Björn Gambäck

---

## Opportunities or Threats for current LT

*(Note. Some initially identified opportunities or threats were circulated among leading experts on LT in the Nordic countries and their comments on these can be found below.)*

Current technology drivers that push the development:

- **Open Source** tools for developing and maintaining language resources such as corpora, lexicons, rule sets or language models
- **Network Technology** for virtual networks and web applications
- **Content Management** using meta data and the semantic web

Customer needs that pull language technology:

- **Tacit Knowledge Growth** creating a need to share and maintain increasing amounts of corporate knowledge
- **Continued Globalization** requiring multilingual solutions
- **Process Speed-Up** demanding automated production procedures and a shifting of manual production efforts to quality control
- **Social Needs** for empowerment and cooperation, i.e. individually having full control of key resources and procedures in the work environment while cooperating in a team

## What other technology trends or customer needs do you think will affect Language Technology?

*(Quotes in order of submission:)*

Technology trends:

- Statistical Machine Learning techniques for efficient and adaptive knowledge acquisition and representation.
- Multimodal Systems that integrate language technology with pattern recognition and machine vision.

-- Timo Honkela

Other important trends:

- multimedia systems (= the combination of audio and video data with written texts)

- increasing demand to digitize archived information (making historic documents from libraries and archives available online). We will soon have a situation of a document being either electronically available or being ignored.

-- *Martin Volk*

The strongest pull factors are:

1. information access through search and
2. pervasive online services.

-- *Koenraad de Smedt*

I believe, frankly, that Open Source **tools** in this area is not important. What is important is available linguistic resources (lexicons, corpora (written and spoken) and perhaps grammar fragments) which make it possible for companies to develop solutions for the Nordic countries.

-- *Torbjørn Nordgård*

Embedded software is a continuing technology trend, which enables new interface solutions where LT can play a role.

-- *Lars Ahrenberg*

T.ex. behovet av bättre språkkontrollverktyg för tidningarna är en aktuell utmaning.

-- *Mikael Reuter*

The majority of information from local national governments will be accessible via the web; reduced availability of human assistance. Easy access to information will require LT, both due to information overload and to reduce/eliminate the effects of the digital divide.

-- *Torbjørn Svendsen*

Needs of the public system (as argued above, with responsibilities and the will to provide information to all), to level the playing field between accomplished writers and readers on the one hand and those with less experience on the other: schoolchildren, educationally less advantaged groups, recent immigrants etc. This especially concerns the educational, the administrative, and the legal system - but media houses will also have vested interests in making their information publicly available.

-- *Jussi Karlgren*

Customer needs: Machine translation between Nordic languages and English

-- *Eckhard Bick*

Other technology trends:

- Hardware development, notably memory size and processor speed, makes more advanced language technology solutions possible, in more contexts.
- Increased use of electronic rather than paper archives will force the development of LT tools for marking, storing and retrieving data.

- A growing European integration requires multilingual solutions.

-- *Trond Trosterud*

Trends:

- Automated public information services, monolingual or multilingual, for fully functional or disabled persons Speech synthesis or recognition embedded in future ubiquitous computer tools or practical gadgets

Customer needs:

- The need for voice operated (input and output) computer tools and gadgets, both amongst fully functional and disabled persons

-- *Jan Hoel*

Mobila system och allestädes närvarande datorer (ubiquitous computing) är en trend som säkert håller i sig och där språkteknologi har en viktig roll att spela. Det finns många behov hos användare som inte är uppfyllda och där bättre användaranpassning krävs av språkteknologiska produkter, t.ex. för funktionshindrade, olika användarkompetenser, olika verksamhetsområden som t.ex. undervisning.

-- *Rickard Domeij*

'Technology trends':

- Trend towards multi-modal and more flexible interfaces to information systems, will increase the demand for LT, in particular Speech Technology.

'Customer needs':

- Increased usage of and dependence on the world wide web as a general source of information.

-- *Tron Espeli*

Multi-media technology will be a driver for both content management and language technology

-- *Bernt A. Bremdal*

Management of large-scale digital speech, text, image and video corpora: While the available resources continue to grow almost exponentially, the challenge of language technology is to provide new efficient multilingual language processing algorithms

-- *Mikko Kurimo*

Customer needs will pull the LT forward such as Information access for all including people with functional handicaps. The aging population will put new demands on LT services for increased life quality. Language learning will require LT support for training.

-- *Rolf Carlson*

The merging of television, telephony and other media through IP. Related to this the use of various end-devices, e.g. mobile phones that call for the use of LT e.g. in mediating text. Cultural needs to express one identity through language.

-- *Arnor Gudmundsson*

There are also citizen and societal needs (not all - or even most - human relationships should be thought of as involving a customer). Multilingual, cross-lingual, multimodal access to public services and participation in public life.

-- *Lars Borin*

What the semantic web is or will be can be debated (as well as whether it exists or ever will exist). More important is, however, the possibilities for distributed data storage (over the net), as well as the possibilities for wider and easier access (e.g., through mobile devices).

-- *Björn Gambäck*

---

## Vision

*(Note. An initial vision was circulated among leading experts on LT in the Nordic countries and their comments can be found below.)*

In 2016, after a 10-year period of focused investments in making the Nordic Countries a Leading Region in Language Technology (LT),

- **Multilingualism is seen as a strength** of the Nordic/Baltic region.
- **LT research and applications develop freely** in several directions in a stimulating research and business environment.
- **Availability of necessary language resources** improves the quality of LT research and application development.
- **Mono- and multilingual LT modules with uniform APIs for a wide array of languages are smooth and easy to integrate** into software products and services.
- Available LT modules give the software industry and the service providers in the Nordic/Baltic region a competitive edge in the global market place, by **facilitating the process of tailoring products and services to language-specific requirements** in new international markets.

**Do you have additions or modifications to the vision for strengthening the Nordic Countries as a leading region for LT in 2016?**

*(Quotes in order of submission:)*

I agree with the above. I would like to add: - LT modules will be integrated in multimedia systems (e.g. aligned with video systems for video retrieval).

-- *Martin Volk*

The claim that 'LT modules are smooth and easy to integrate' is too optimistic. Integration will always require a high level of competency and careful tuning to the application context.

-- *Koenraad de Smedt*

Nordic and Baltic languages are generally small language communities, and countries in these regions will need to be in the forefront to maintain and protect their local language existence.

-- *Knut Aasrud*

The vision is OK, but it presupposes a quite substantial economical effort from the governments in the countries in question.

-- *Torbjørn Nordgård*

I would stress the use quality of LT systems, and that citizens of the region are able to access software-mediated services in their mother tongue. I absolutely agree with statement 2.

-- *Lars Ahrenberg*

It is important that the less widely used languages in the region will be able to participate in this development. For that to be possible, they may need some external support since sufficient national funding for building the necessary resources may not be available.

-- *Eiríkur Rögnvaldsson*

Något kunde väl också sägas om det nordiska språksamarbetet och dess långa traditioner, inklusive möjligheterna till samarbete mellan språknämnder och språkteknologiföretag.

-- *Mikael Reuter*

I would like to make sure that multilingualism comprises the very small languages in our Nordic context, and also in principle all, but in practice a diligently chosen number of our immigrant languages. As for applications: I agree, that they should develop freely in ..a.. business environment, but let us add, that applications to the benefit of the elderly and impaired should be developed in a non-competitive environment with public support.

-- *Henrik Holmboe*

Public information is freely and openly available and disseminated in several languages. The public administrative bodies of the Nordic countries take their information dissemination task seriously. The language care tradition of the Nordic countries has strong support from language users. Bodies, both publicly funded, such as Kotus, Svenska akademien, Nämnden för svensk språkvård and industrially funded bodies such as TNC, with no legislative but established and accepted status on

questions about language usage, have a voice in the public arena.

-- *Jussi Karlgren*

Focus on language awareness and multilingualism awareness in primary and secondary schools, as well as better school training in analytical and formal aspects of native and foreign languages - as a prerequisite for a strong LT base in the upcoming generation of students

-- *Eckhard Bick*

I think the list identifies both our current strengths and (by listing their opposites) some of our current obstacles to being in a leading position already today. Short: Our strength is the multilingualism (including typologically diverse languages), combined with a long tradition of linguistic research). Our weaknesses are the relatively small market sizes, the limited availability of common resources.

-- *Trond Trosterud*

'Multilingualism' as the designation is used here goes beyond 'Nordic languages' and the wider designation 'languages of the Nordic countries'. It is difficult to see how this widening of scope could possibly be to the benefit of the LT situation for the Nordic languages and other traditional languages of the Nordic countries.

-- *Jan Hoel*

All basic tools and resources available as open source will provide a platform for further innovation and new products.

-- *Sjur Nørstebø Moshagen*

Språkteknologi bidrar till ökad demokratisk delaktighet i samhället bl.a. genom att göra information tillgänglig via t.ex. 24-timmarsmyndigheten. Språkteknologi är till för alla, oavsett språk, kön, klass, etnisk tillhörighet, kognitiv och fysisk funktionsduglighet, språklig och teknisk kompetens, verksamhetsområde m.m. Språkteknologi bidrar till att stärka de nordiska/baltiska språken och den nordiska språkgemenskapen i en mångspråkig värld. Till 'multilingualism' bör inte bara de nordiska/baltiska huvudspråken räknas, utan också minoritetsspråk, teckenspråk och invandrarspråk. Det är viktigt att tänka på att utveckla resurser också för kommersiellt mindre gångbara språk som t.ex. samiska.

-- *Rickard Domeij*

Other factors may be decisive for software industry and service providers to become global leaders - this part of the vision should therefore focus on the actual contribution to this aim from the LT domain.

-- *Tron Espeli*

Permanent LT research - and LT development forums have been set up in the bigger Nordic countries in support of Nordic and Baltic languages with lesser volume in economic as well as human terms. Comment: It should be an attractive alternative to mother tongue research for Danish, Finnish, Norwegian, Swedish scholars to work with Faroese, Greenlandic, Samic, (Icelandic?) where money is scarce but humans scarcer or with Baltic, Romani, etc. where humans are scarce but money scarcer.

-- *Per Langgård*

After 10 years, a relevant infrastructure should have been developed for both text and speech to cover all languages in the region and dialects. The data should have been properly annotated at all levels. After 10 years, the joint efforts in the Nordic countries should have resulted in healthy industries that can support applications in all Nordic languages. This requires for example a good knowledge of spontaneous spoken interaction.

-- Rolf Carlson

The available LT modules shouldn't just have uniform APIs but be open source. The market in the Nordic countries is too small to support attempts to make business out of the resources per se (companies that aim to make a profit in the LT field would be better advised to make it from selling consultancy services and support to their products - and give the products themselves away for free to encourage usage).

-- Björn Gambäck

---

## Prerequisites for implementing the Vision

*(Note. Some prerequisites for the initial vision were circulated among leading experts on LT in the Nordic countries and their comments on these can be found below.)*

**Opening up lexicon resources** which have been created through public funding. Lexemes including the part of speech and inflectional codes as well as other mark-up should be moved to the open source domain so that anybody can alter and make use of them for research or commercial purposes.

Cooperation in **creating open source tools for building further LT modules** which, in turn, can be either proprietary or open source. Cooperation in creating open-source runtime support for the LT modules built with those tools.

**Stimulating LT research** for various application areas. National funding programs would provide the basis, and a Nordic/Baltic framework program for networking would provide the necessary regional infrastructure and communication.

**Adjusting the university teaching** to the needs. Better quality and wider availability of teaching and supervision on all special areas through cooperation at master's level teaching (perhaps as a Nordic/Baltic masters program) and in a Nordic/Baltic PhD teaching network (NGSLT).

**Do you find that there are additional prerequisites for progress in LT? Are any of the above irrelevant?**

*(Quotes in order of submission:)*

I agree and would like to add opening up language resources on all levels (lexicons, grammars, written language corpora and speech corpora, etc.) which have been created through public funding. ... I like the Nordic cooperation in PhD education. NGSST is very good. But I am skeptical about coordinated Masters programs. I assume such coordination will happen rather on the local level (= neighboring universities).

-- *Martin Volk*

Education should not stop with a masters or PhD degree, but one needs to reach people already working in the industries that will integrate LT modules. Universities must create programmes for lifelong learning in HLT.

-- *Koenraad de Smedt*

Lexicon resources are important, but parallel texts and corpora (raw as well as annotated) are even more important because they are necessary in order to develop further monolingual and multilingual lexicons, taggers, parsers, and many other resources and tools.

-- *Janne Bondi Johannessen*

Alternatively, or in addition, the prospect of being able to financially benefit from language technology should not be jeopardized by opensourcing too much IP. The opportunity to be able to make money on LT IPR must be protected to attract people and money to this field.

-- *Knut Aasrud*

Regarding 'lexicon resources' they should be made available with no requirement for sharing additions, i.e. MIT license or Extended GPL.

-- *Torbjørn Nordgård*

Open source is a good idea, but the announcement of an open source project does not necessarily create a community of users to take part in the development. National funding programmes would not be sufficient to support 'various application areas'. I believe a better idea is a few (one or two) focused projects that invites (i) public funding, (ii) private funding, and in the best of all worlds (iii) public interest (i.e. a community of 'volunteers') say, something like a talking robot that any user could teach new words, or new languages. I would also add integrating LT with other technologies and design disciplines.

-- *Lars Ahrenberg*

Opening up (and creating) resources should include more than lexicon resources, notably corpora, possibly also grammar resources.

-- *Joakim Nivre*

I think all of this is very important. Especially, I want to emphasize the need for cooperation in master's level teaching - both cooperation between universities and countries, and also cooperation between different fields such as linguistics, computer science, statistics, etc. It is also important to assist smaller language communities in building basic resources, as pointed out above. Furthermore, it is necessary to raise public awareness about the importance of LT in our daily lives in the future, and to get commercial companies interested in LT research and development. It is also

important to increase cooperation between universities and research institutes on one hand and private companies on the other.

-- *Eiríkur Rögnvaldsson*

I think we should include the BA-level as well; - and try to develop common teaching material, compendia and curricula using the idea of a common core with local variations.

-- *Henrik Holmboe*

If we want to change the status of Language Resources to Open Source, there is absolutely no need to limit ourselves to lexical Language Resources. Grammars, Parsers, Named Entity Recognizers, etc. are no different. So, I disagree that tools (made under the same conditions, i.e. public funding) could be left proprietary. However, the current policy of the Danish Ministries urges universities to invoice everything! We may of course recommend they behave differently. We should also remember that Open Source does not necessarily imply FREE, it only implies access to the source code. Promotion of standards would be beneficial, but is not a necessity. Documentation of Language Resources is a prerequisite if they are to be Open Source. If the user does not understand the categories used, he/she will fail in the use of the data and in their further development. An infrastructure to support the distribution of the Language Resources and tools will also be needed, it may be centralised or distributed, but it has to be set up. This could be a Nordic effort, or it could be done at a European level (e.g. by making special agreements with ELRA, or by joining other initiatives)

-- *Bente Maegaard*

- Availability of other language resources, i.e. huge amounts speech and text.
- Sufficient funding for both long term (university) research and support for industrial development.

-- *Torbjørn Svendsen*

I'm all for open source but open standards and open APIs are more important. Bring in the industrial players.

-- *Jussi Karlgren*

Preference should be given to research funding that integrates all research groups in a given area for a given country, or the Nordic area as such, rather than supporting a centralized (e.g. capital university based) funding approach.

-- *Eckhard Bick*

Potential users in all sectors and walks of life must be convinced that LT is something they need. Only powerful demand from the public will make politicians prioritize the area in question.

-- *Jan Hoel*

Standardization of resources and APIs, as well as tools for interchange and conversion of data from one format to another. Building on open-source lexicons and open-source tools, the next step would naturally be to harmonize these resources to really benefit

from the available resources.

-- *Sjur Nørstebø Moshagen*

En samordnande funktion är en viktig förutsättning för att organisera ett samarbete, och inte minst för att överbrygga intressekonflikter och problem mellan forskare, industri och rättighetsinnehavare i tillgängliggörandet av resurser. Vid finansiering av forskning måste det finnas tydliga krav på tillgängliggörande av resultat och resurser.

-- *Rickard Domeij*

To the extent that there are lexicon resources established by private funds it should be considered if, and how (and to what extent) these resources could be made publicly available.

-- *Tron Espeli*

Commercial and industrial recognition of the advantages and broad involvement of these parties through all phases.

-- *Bernt A. Bremdal*

The 'go-west-maxim' that prevails at the political level and even in many academic institutions must be addressed. As long as factual policy as well as public opinion in reality sees uniformity in English as a necessity while validating diversity as a beautiful but slightly anachronistic dream in academic ivory towers LT as a roadpaver for multilingualism will never obtain the support needed.

-- *Per Langgård*

The development of lexicons should be done with speech technology in mind. That is, lexicons should include phonetic information, such as a phonetic transcriptions and stress.

-- *Martti Vainio*

A good progress in the LT field needs support for joint projects and networks on the Nordic level. To be able to share information and speed up development the infrastructure development needs to be accompanied by analysis software and methods for easy access. This is a research topic by itself.

-- *Rolf Carlson*

- Opening up all kinds of linguistic resources (not only lexicons): corpora, grammars, speech databases, lexicons, etc.
- Linguistic research on spoken language varieties (registers, dialects, non-native) and on non-standard written varieties (CMC, non-native, borderline literate)

-- *Lars Borin*

Obviously, the first two points apply to all types of language processing resources - and are very important. Whether national funding is important is a question which depends on which roads the EU research funding takes.

-- *Björn Gambäck*

# Key areas with magnitudes of investment

In order to implement the vision, some key areas with magnitudes of investments and modes of cooperation were identified. These were proposed by the leading experts on LT in the Nordic countries. Below is a list of their suggestions arranged into thematic tables.

## LT policy

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
----------	-----------------------------	------------------	---------------------	--------------

## LT resources

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
Basic Language Resource Kits	>10 MEUR per language	Universities, copyright owners	infrastructure, standards, API:s	Lars Ahrenberg
BLARKs for the Nordic Languages	~5 MEUR per language	Universities, research institutes, LT companies w. coordination at Nordic level	National projects, coordinated at the Nordic level	Joakim Nivre
Basic LT resources and tools for individual languages	10 MEUR	Governments, national and Nordic funding bodies	Funding programs, exchange of researchers	Eiríkur Rögnvaldsson
Language resources	> 40 MEUR	Universities, research institutes, industry	Harmonization of data, contents, formats and availability	Torbjørn Svendsen
Linguistic resources	10 MEUR	open source, gov, univ, companies	Develop and standardise linguistic and multilingual resources	Sjur Nørstebø Moshagen
Opening and developing the language resources for international and	> 10M	leading universities	projects	Mikko Kurimo

interdisciplinary use				
Basic linguistic resources	100M			<i>Lars Borin</i>
LT resource development		Key actors in national projects.	Action plan at Nordic level, complemented by projects to facilitate a common approach	<i>Tron Espeli</i>

## LT modules

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
Open source resource development	2 MEUR	University coordinator with partners	Institutions	<i>Timo Honkela</i>
Distributed open source and APIs				<i>Lars Borin</i>
Interoperability of language modules/tools				<i>Arnor Gudmundsson</i>

## LT tools

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
Open source LT tool		one or two institutions per country	Nordic LT network	<i>Martin Volk</i>
Coordinated development of suitable open-source LT tools	1 MEUR	universities, enterprises	on a Nordic basis	<i>Jan Hoel</i>
Tools, applications	2-5 MEUR	open source community, univ, public and private institutions	Develop open source, basic tools and applications	<i>Sjur Nørstebø Moshagen</i>
Cooperation in creating LT tools				<i>Arnor Gudmundsson</i>
Create cross-linguistic platforms in all public funded resources		Individuals, individuals' affiliations, funders	networking	<i>Per Langgård</i>

## LT corpora

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
Multilingual annotated language corpora	1 MEUR	one or two institutions per Nordic language	coordinated annotation project	<i>Martin Volk</i>
Corpus collection, written text	~15 MEUR pr language	Universities	Networking across countries	<i>Torbjørn Nordgård</i>
Corpus collection, spoken data	~20 MEUR pr language	Universities	Networking across countries	<i>Torbjørn Nordgård</i>
National corpora for the Nordic languages	~10 MEUR per language	Universities, research institutes, LT companies	National projects, coordinated at the Nordic level	<i>Joakim Nivre</i>
Collection of linguistic material, speech and text, for a public 'language bank'	1 MEUR	public authorities, universities	nationally	<i>Jan Hoel</i>
Marking and other preparation of the above material	1 MEUR	universities	nationally	<i>Jan Hoel</i>

## LT lexicons

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
Lexicon development	~10 MEUR per language	Universities	Networking across countries	<i>Torbjørn Nordgård</i>

## LT research and development

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
Academia/industry collaboration	15 MEUR	Universities, industry	Joint effort for standardization	<i>Torbjørn Svendsen</i>
Basic technology research	15 MEUR	Universities	Joint programme, Researcher exchange, workshop, division of research tasks	<i>Torbjørn Svendsen</i>
R&D Funding	50-80 MEUR	Universities, Research institutes,	Nordic projects	<i>Bernt A. Bremdal</i>

		industry		Joint Nordic Research projects, funded and organised on Nordic level	<i>Tron Espeli</i>
LT research					
Statistical machine translation	1.5 MEUR	Helsinki University of Technology as coordinator with partners	Projects		<i>Timo Honkela</i>
Multimodal systems	1 MEUR	Helsinki University of Technology as a consortium member	Projects		<i>Timo Honkela</i>
Language Learning	10 MEUR	Universities, industry, volunteers	open source		<i>Lars Ahrenberg</i>
Public information tools adapted to the mobile life of users	2-5 MEUR	consortium of research partners, telecom service providers, handset makers, public bodies	IST-IP-like project		<i>Jussi Karlgren</i>
CLIR tools, focused CLIR tools for recent immigrants	-'	public bodies, research partners, media houses	-'		<i>Jussi Karlgren</i>
Biotek informatics	-'	academic partners, research intensive corps	-'		<i>Jussi Karlgren</i>
Machine translation	> 5 MEUR	Universities & business community	shared research		<i>Eckhard Bick</i>
Research on automatic methods for multilingual information processing	> 20M	leading universities	projects		<i>Mikko Kurimo</i>
Multimodal applications					<i>Lars Borin</i>
Multilinguality					<i>Bente Maegaard</i>
Developing applications					<i>Arnor Gudmundsson</i>

## LT training and education

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
Coordinated PhD education			NGSLT	<i>Martin Volk</i>
Master's level education	1 MEUR	Nordic/Baltic universities	Distance education, exchange programs for teachers and students, common curriculum	<i>Eiríkur Rögnvaldsson</i>
Language awareness and formal language knowledge in schools	> 1 MEUR	universities, schools	development and empirical studies in a cross-institutional framework	<i>Eckhard Bick</i>
Popularization	1 MEUR	R&D, Government, Industry	Professional PR assignment	<i>Bernt A. Bremdal</i>
'Sell' the idea of diversity to a much wider audience		All parties one could think of	Nordens sprogråd for instance	<i>Per Langgård</i>
Strengthen and modernize formal mother tongue training at all levels in education		Ministers of education, L1 teachers, applied linguists	national + Nordic support at the attitude level	<i>Per Langgård</i>

## LT legal aspects

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
Corpus material for research purposes		governments	committee	<i>Kimmo Koskeniemi</i>

## LT business aspects

Key Area	Magnitude of funding needed	Parties involved	Mode of cooperation	Suggested by
Web services	> 3 MEUR	universities, business community	tool sharing, hosted products	<i>Eckhard Bick</i>
LT module distribution		industrial and academic players	Action plan managed at Nordic	<i>Tron Espeli</i>

## **Additional comments**

I don't find myself in a position to say anything about the magnitude of funding needed, but I firmly believe that for a small language community like Iceland, continuing Nordic cooperation within Language Technology is vital. The Nordic Language Technology Programme 2000-2004 was very important for us. However, we would have benefitted more from the programme if it had started a couple of years later. The reason is that its start coincided with the start of the national Icelandic LT Program, which literally marked the beginning of Icelandic LT. Thus, we were not prepared to participate as much in the Nordic Programme as we would be now, but some kind of a continuation of that program would be very beneficial for us.

-- *Eiríkur Rögnvaldsson*