EUR MAP
LANGUAGE TECHNOLOGIES

Benchmarking HLT progress in Europe

in Europe

The EUROMAP Study
Andrew Joscelyne and Rose Lockwood

Copenhagen 2003

Table of Contents

Information Society Technologies

Preface

The report concludes that a visible presence for European HLT activities should be established

Human Language Technologies (HLT) enable humans to communicate with computers and to use computers in a more natural way and in their own language, i.e. to participate in the information society in a totally natural way. HLT is particularly important for Europe as no other advanced economic area enjoys a similar cultural and linguistic diversity. The need and ability to use multiple languages in every-day life is an increasingly familiar aspect of business, leisure, government and civil society in the EU and the Candidate Countries. Actually, being able to do business in several languages has become a com-mercial necessity.

The EUROMAP Language Technologies project has investigated the state-of-the-art of HLT research and take-up in Europe, as well as the background for the present situation in each country. Building on data collections of research centres, suppliers, national research policies, and on market analyses, the European countries have been compared in a bench-mark analysis. This analysis shows e.g. that the sig-nificant and steady investment made by the authori-ties in Germany, UK and the Netherlands has paid off - these countries are the European 'Leaders' in HLT. The situation in other countries is described as well and suggestions made for the future development.

The report concludes that a **visible presence** for European HLT activities should be established, and that it should have a strong relationship to the European Research Area. The goal should be to have a set of robust, stable, multilingual HLT modules, capable of being embedded into emerging IST appli-cation environments. A **Language Technology Agency** should be established to supervise and monitor the transition from national HLT efforts to a truly European technology level of language parity. **Infrastructural funds** for the provision of language resources and basic language technology modules for all languages should be made available and should be monitored by the Language Technology Agency.

It has been extremely interesting to work with these matters and to see in which way different European countries have tackled the challenge of language in the information society, and what the consequences are. We do hope that the data and the analysis provided, as well as the recommendations, will be taken up by policy makers at the national and European level.

Finally, while acknowledging the support of the European Commission, we should stress that this report is the result of the EUROMAP project, and opinions herein do not necessarily reflect the opin-ion of the European Commission.

Bente Maegaard
Co-ordinator for EUROMAP
Language Technologies
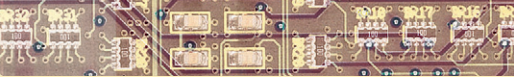
Introduction

What is HLT?

***The combination of speech and NLP provides
powerful technology for interaction***

The term Human Language Technologies (HLT) covers the group of software components, tools, techniques and applications that process natural, human language. HLT comprises two broad areas: speech processing and natural language processing (or NLP). Speech technology replicates the human ability to hear and utter spoken language, for example in speech recognition applications. NLP technology models the human capacity to comprehend and process the content of human language - i.e. to understand and transform written text. Automatic translation for example is a common application of NLP. The combination of speech and NLP provides powerful technology for improving the interaction between humans and machines, and between humans *using* machines.

In the last ten years, HLT has grown from a highly specialised, theoretical research topic to a core technology of the Information Society. By contemporary standards - in an environment where cycles of technology innovation have been reduced to months rather than years - the advance of HLT may seem slow, if not glacial. This impression is misleading. Basic research in the field has been underway for 50 years or more. After decades of research, the conditions for exploiting HLT began to emerge in the 1990s, and advances in the use of HLT have grown steadily ever since.

HLT thrives in the conditions that support the information revolution - high levels of relatively affordable computing capacity, and virtually universal connectivity. For decades, the complexity and computational demands of HLT were a barrier to development, and this barrier in turn limited the scope of research. The happy convergence of information society technologies has both provided the infrastructure that can support HLT, and driven the need for exactly the kinds of products and services that HLT can support.

Why is HLT important for Europe?

***HLT plays a unique role in the EU, because of the
cultural and linguistic diversity***

HLT plays a unique role in the European Union, due to the unusual cultural conditions that pertain to Europe, both socially and economically. There is no other advanced economic area that enjoys the cultural and linguistic diversity of Europe. The 11 official languages of the EU will grow to more than 20 as the next round of Candidate Countries join the Union. There are dozens of additional languages in common use in the Union, including regional languages (such as Catalan and Basque in Spain), non-official national languages (such as Welsh in the UK), and immigrant languages (such as Urdu in the United Kingdom, Maghrebi Arabic in France, and Turkish in Germany).

The ability - and willingness - to use multiple languages in everyday life is an increasingly familiar aspect of business, leisure, government and civil society in the EU. This reflects the aspirations of European citizens to integrate, alongside their deeply held respect for locale. Europeans have become increasingly aware that active support for linguistic diversity protects the rights of all citizens to maintain their own languages - not to the exclusion of others, but as part of the common cultural assets of the Union.

The information revolution, therefore, brings particular challenges to the EU. In an increasingly dense information environment - for citizens and consumers, governments and businesses - language transparency becomes vital. If all citizens across the expanded Union are to participate fully in the information society, the products and services of that society must be available in all their languages. If Europe is to operate successfully as a single market, and if the goals of the eEurope vision are to be achieved, those products and services must be delivered cross-lingually, making it as easy to move across languages as it is across borders.

From another point of view, the language challenge in Europe will almost certainly prove to be an advantage for the EU. As the EUROMAP Study confirms, many of the products and services of the information society will be built on core HLT components. The importance of HLT goes well beyond the obvious, and penetrates into the deepest layers of the Internet and the web, where the ability to process the components of language - coding knowledge and intelligence into the information infrastructure - will be the basis for next-generation technology.

The EU has already established its credentials as the most advanced research location in the HLT field. The very difficulty of developing HLT for many languages gives European researchers and technology developers a natural advantage in one of the most crucial technologies for the next generation of information and communication technology. As a consequence, commitment to the future of HLT in Europe is perhaps most important for the contribution it will make to the strength of the European ICT sector. A study by Booz-Allen & Hamilton, *The Competitiveness of Europe's ICT Markets: The Crisis amid the Growth* (presented at the Ministerial Conference in March, 2000) documents the challenges to Europe's global competitive position from weakness in several key segments of ICT, including software. A recent study by The Conference Board, *Productivity, ICT and Service Industries: Europe and the United States*, assesses the impact of ICT on productivity. Europe's productivity gap in ICT-using industries, especially services, is notable. Thus the EU faces competitive challenges in ICT from both the supply and demand perspectives.

HLT is a 'small' technology in pure market terms, but its potential impact - on accessibility, innovation, and integration - is significant, and its crucial role in unlocking the potential for eEurope is unchallenged. The EUROMAP report outlines the current state of language technology in Europe, including its role in new paradigms for next-generation ICT. It assesses the progress, and the barriers, for HLT communities in each Member State, and recommends actions to

secure the future of this important technology in Europe.

Background to the EUROMAP Study

The study offers suggestions for the next generation IST research agenda

EUROMAP Language Technologies is a European Commission supported initiative dedicated to promoting greater awareness and faster take-up of HLT within Europe. Since 1996 the project has served as a central resource and marketing-support unit providing information to all communities involved in the language technology field, from researchers and developers to suppliers and users. EUROMAP supports national HLT communities-of-interest through National Focal Points that provide direct services to their constituencies. Through pan-EU initiatives - such as the www.hltcentral.org web site and the LangTech Conferences - EUROMAP knits together the various stakeholders in the European HLT community.

The EUROMAP Study brings together the experience gained in more than five years serving Europe's HLT community. It draws on the resources and knowledge of many experts and practitioners, from every Member State and from a number of New Accession Countries. The EUROMAP network has documented the state of the HLT research community, and tracked the steadily growing number of new companies operating in the HLT field. Through seminars and fieldwork, the network has documented the evolving market for HLT technologies. Through consultation with leading HLT visionaries and co-ordination with the HLT research network ELSNET, EUROMAP has developed a wide view of the many opportunities and challenges in the field.

This study follows on from a report published in 1998 that provided the first pan-EU perspective on this emerging technology, at the beginning of the Fifth Framework Research and Technology Development Programme. EUROMAP has continued to track the

progress of HLT, and has developed a prototype benchmarking method that measures progress in the field.

The results of the current study point toward the future of HLT in Europe, and identify policies and practices that have yielded successful outcomes. The study offers suggestions for leveraging the success of previous research investments into the next-generation IST research agenda for Europe.

State-of-the-Art: the technologies

Some 300 European companies have been identified

The EUROMAP project has identified some 300 European companies offering HLT-based products and services. Most of these companies are based in current Member States of the EU, but there is also a small but growing base of companies in Candidate Countries in Eastern Europe. Many of these companies offer combinations of different HLT features and functions, ranging from basic components to advanced solutions based on speech and text processing.

Components and resources

All HLT relies on core language processing components that digitally model or replicate the way humans process language. These components can be based on linguistic rules (such as grammar), on statistical analysis (e.g. to measure the probability that a text or an utterance has a particular meaning), or on a mix of the two. In addition, all HLT techniques need a source of linguistic data as a reference, such as a lexicon (a dictionary coded with grammatical information), or a 'corpus' that provides a large database of the raw material of language, either text or speech. The existence and availability of these basic components provide the baseline for development of HLT. EUROMAP has identified around 120 European companies offering core HLT components and language resources in roughly 25 languages.

Knowledge processing

HLT components can be embedded in a wide range of what are generally called knowledge applications - i.e. products and services that process information using some level of linguistic intelligence. Search engines use HLT components to improve the matching of search terms, e.g. by retrieving different morphological forms of a word, or even synonyms. More advanced applications, such as knowledge mining, can use complex combinations of HLT tools to find, analyse, and create reports on the content of text or document repositories. Increasingly, large companies are developing taxonomies (i.e. structured trees of linguistic concepts) to organise and manage their content assets. EUROMAP has identified around 120 European companies offering HLT-based knowledge processing products and services, in some 25 languages.

Interface and Interaction

Interface and interaction technologies are often speech-based. The most familiar uses of speech technology are telephone-based speech recognition systems that eliminate the need for a keypad, commonly used in call centres and telephone transaction systems. Speech recognition systems are also used in dictation systems that bypass the keyboard. On the other hand, speech synthesis (Text-To-Speech) systems are increasingly used for applications such as 'listening to email', after having served for a long time as reading support tools for the visually impaired. More advanced applications include voice authentication, where a person's identity is verified from a voiceprint. Speech systems have gone beyond traditional platforms, and are now embedded in common consumer items, and in telematics systems in cars. EUROMAP has identified around 130 European companies offering HLT-based Interface and Interaction products and services, in some 25 languages.

Cross-linguality

Automatic translation (machine translation, MT) was the earliest NLP application, and remains one of the

most technically challenging. Nevertheless a large number of products have been developed for many different language pairs, and free 'gisting' translation is widely available on the web. Aside from MT, cross-lingual applications can overlap with both knowledge and interface applications. A cross-lingual search engine can translate a term in order to search repositories in different languages, retrieve the 'foreign' language text, and provide a rough translation, or even a summary, in the language of the original search term. Several prototype systems exist that provide cross-language speech applications, such as telephone reservation systems that allow people speaking different languages to communicate. EURO-MAP has identified around 60 European companies providing cross-lingual products and services in 25 languages.

The Multilingual Semantic Web

The next-generation Internet will embed core linguistic data at the heart of the web. The Semantic Web initiative aims to capture and encode the semantics (i.e. the meaning) of all types of digital content, and use that embedded knowledge to enable more predictable levels of interaction between different systems and services. Agent technology armed with semantic knowledge about a user will interact with virtually any electronic system that shares its semantic knowledge of the world. This knowledge will be captured in ontologies - structured sets of concepts with agreed relationships that represent real-world knowledge. European HLT should be capable of sustaining its position as thought-leader in the development of the Multilingual Semantic Web, assuring that all European language communities participate in the development of semantic resources, and that the services that use them are expressed in all the languages of Europe.

Visionary technologies

HLT will be a key embedded technology as next-generation ICT products and services emerge from the lab. Visionary work on information processing is focused on ambient intelligence for ubiquitous computing, where knowledge is embedded in devices throughout the environment, responding to human activities in natural modes of interaction. Research on 'e-sense' will model the way all the human senses are processed in the experience of communication. Thus the boundary between 'knowledge' processing and 'interfaces' will blur, and machines will cease to dominate the modality of electronic communication. Machines will interact with humans in a more human way, and humans will interact with humans using machines that are more transparent. New ICT paradigms will process information about human experience, through all the human senses, in the most natural coding and representational system, i.e. language, creating what has been called the 'perceptually aware cross-lingual human interface'.

Europe's HLT companies: a showcase of competencies

The following showcase offers a very small selection of the 300 or so commercial enterprises that offer language technology products across Europe. They have been chosen to represent the mix of products and capabilities that characterises the very hybrid world of HLT. Application areas covered by these companies include translation automation, speech recognition and text-to-speech, text mining for information, pronunciation learning, taxonomy management, basic linguistic tools and components, dialog management, assistive systems for the disabled and more. And most EU countries are represented.

Many of these companies began as spin-offs from university research centres in the last five years, while others have spun off from large IT companies, usually offering a broader range of integration expertise than purely niche language products. This whole process of new business creation exemplifies the recent maturing of the technology, and the growing role that technology transfer processes and support

play in bringing language technology to the market. The evidence suggests that there are many more suppliers of speech technology solutions in Europe than specifically text-based technology suppliers, and that the speech supplier community is on the whole better funded.

Although this showcase focuses on stand-alone technology companies, it should be remembered that a number of large IT players in Europe (e.g. Philips, IBM, Bosch, Ericsson and various incumbent telecommunications carriers) have also developed ancillary speech & language technologies as part of their larger technology development programmes, and in some cases have branded and marketed them. Here again, the recent tendency has been either to spin off the product unit into a separate company, while maintaining an in-house HLT R&D facility, or simply to sell off the language technology arm and refocus on core business lines.

Today there are no across-the-board language and speech technology players in the European market-place offering the full range of basic technologies (as the Belgian company Lernout & Hauspie tried unsuccessfully to achieve in the 1990s). There are, on the other hand, a number of early signs of convergence between speech and text technologies – for example the number of text-to-speech applications that now require more advanced integration of sentence-level semantics to determine appropriate intonational contours – that may prompt more technology partnerships and other forms of innovative business collaboration between Europe's language and speech technology suppliers.

Innovators in HLT-based Knowledge Processing

Ankiro (Denmark) – User-centric dialogue and knowledge robots

Drawing on a strong R&D base in computational linguistics and computer science, Ankiro has developed a range of 'robots' to simplify the human tasks of searching and communicating with knowledge bases over the web. Founded in 1999, this Copenhagen-based company has built up a leading position in Denmark in the field. Its Dialog Robot Technology is built around a broad set of component language tools, ranging from dictionaries and ontologies, to spell checkers and parsers, that work to optimise the user interface and the search and dialog experience. The company's robots are then used for guidance, FAQ support and CRM services, while the search technology is adapted to the user's interface with a variety of networked knowledge applications. www.ankiro.com

Language & Computing (Belgium) - Semantics for medical knowledge

Language and Computing NV (L&C) is one of the very few companies in the world to use advanced semantics-based language technologies in a commercial context. L&C was formally incorporated in the spring of 1998 to convert academic success into commercial solutions for the delivery of healthcare information. Its key products are LinKBase, the world's largest medical ontology, and TeSSI Indexing and Search components that can automate the process of semantically indexing and retrieving information from documents. L&C is currently adapting these technologies to function in the mobile world of portable devices.
www.landcglobal.com/index.php

Xtramind Technologies (Germany) – Intelligent enterprise information processing

Xtramind uses advanced language and machine learning technologies to develop intelligent solutions for enterprise communications management. The company's major products include an automatic e-mail reader that can learn to route content to the appropriate location based on content profiles, a competitive business intelligence tool that collects and transmits critical knowledge, and a suite of software components to process and disclose linguistic content. Xtramind was formed as a spin-off from the DFKI (German Research Center for Artificial Intelligence) in 2000 and maintains strong ties with Germany's flagship R&D base.
www.xtramind.com

Knowledge Concepts (Netherlands) - Boosting cross-lingual access to corporate content

Knowledge Concepts BV develops technologies that enhance existing document management solutions and search engines to access relevant information both inside and outside the organisation, whatever the language. Founded in 1998 to meet a perceived business need rather than exploit an existing technology, the company works in partnership with major content management platform suppliers to deliver knowledge management solutions for the English, Dutch, German, French, Italian, Spanish, Swedish, Russian, Polish and Arabic languages. They embed semantic networks that enable concepts to be translated instantaneously across languages during searches, making for a transparent flow of information for users. In 1999, Knowledge Concepts was shortlisted for the European IST prize.
www.knowledge-concepts.com

Wordmap (UK) – Enterprise Taxonomy Management Systems

Founded in 1998, Wordmap has become one of the world's leading developers of taxonomy management systems, which enable large organisations to structure their content semantically and thereby improve access and searching. The company uses advanced techniques in linguistics, indexing and lexicography to compile controlled terminology lists that underlie the centralised taxonomies tailored to its clients needs. One key client is the automotive giant DaimlerChrysler who are building a complete new web-based content architecture, using Wordmap's taxonomy tools to identify and manage valuable content in a rigorous, uniform way. Wordmap has also developed special tools to handle metadata and taxonomies in various branches of government for the UK e-government initiative.
www.wordmap.com

Innovators - Interface and Interaction

Telisma (France) – Speech recognition for telecommunications voice services

Telisma is a world-class French speech-technology supplier that develops software applications for the demanding telecommunications marketplace. Building on the track record of France Télécom's pioneering R&D centre, the company develops speech recognition software for mass market applications by focusing on the human factors that drive user satisfaction. Telisma has sales offices in Spain, Italy and Germany and provides speech recognition solutions to telecommunications companies through a web of technology partnerships. By paying close attention to costing and maintenance issues in deploying speech communications, the company acts to make the technology more accessible to a broader range of enterprises and end users. In 2002, Telisma was listed as one of Europe´s 100 most successful start-ups.
www.telisma.com
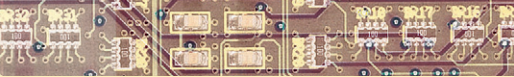
Auralog (France) Speech technology for the language learning industry

Acknowledged as one of the world's most innovative technology suppliers to the language learning industry, Auralog integrates speech technology into software solutions to coach pronunciation and dialog skills in foreign language learners. The company was founded in the late 1980s, and marketed the very first language-learning application (Talk to Me) based on speech recognition in 1991. By pioneering an IT solution in a potentially large market, Auralog has managed to maintain its leading position through its product quality, careful marketing and multilingual outlook.¨
www.auralog.com

Acrolinx (Germany) – Advanced content management tools

Acrolinx GmbH is an example of a spin-off company that has successfully transferred technology from Germany's prestigious DFKI Language Technology Group to the marketplace. Based in Berlin, the company's core competency lies in providing tools and services for handling unstructured text content for knowledge management tasks. It provides term extraction and engineering, and especially language checking components for the multilingual

documentation process, with special attention to workflow design and the deployment of controlled language editors.
www.acrolinx.de

Loquendo (Italy): A global speech technology powerhouse

A world-class speech technology company, Loquendo is a Telecom Italia Group company headquartered in Turin that has successfully transferred decades of R&D experience into best-of-breed speech recognition and synthesis systems. It provides a full range of voice-driven technologies in nearly ten languages, as well as a development platform and integration environment for enterprise solutions. Loquendo markets its products in Europe, Asia, Latin America and the USA, and claims to process over a million calls a day. The company is particularly proud of its range of both male and female voices for its text-to-speech products.
www.loquendo.com

Rhetorical Systems (UK) - High quality speech synthesis in multiple languages

Rhetorical Systems develops very high quality text-to-speech solutions in a variety of languages and voices for a growing range of online applications. The company was founded in 2000 by language and speech engineers from the University of Edinburgh, has a US subsidiary, and regularly announces new customers. Rhetorical's core product – rVoice – offers a broad range of synthesised voices, with 20 different regional accents and speaking styles in English, German and Greek. This choice enables customers to tailor their voice services to customer preferences and even to 'brand' their voice identity. Application domains for Rhetorical technology include online flight and weather information, driving instructions and phone-access email.
www.rhetoricalsystems.com

Innovators in Cross-Lingual Applications

ESTeam (Sweden/Greece) - Resource-driven translation automation

ESTeam AB is one of the small number of dedicated translation automation companies in Europe. Founded in 1995, the Swedish company has a business site in Gothenburg and development site in Greece. In 2002, ESTeam launched ESTeam Translator, a comprehensive, Unicode-compliant client-server translation environment that deploys both Translation Memory and Machine Translation to exploit any available data at various levels of processing (paragraph, sentence and sub-sentence) for the language pairs in question. The system's Machine Translation module, which combines linguistic rules with data-driven methods, is available in all combinations of over 12 European languages.
www.esteam.gr

Sail Labs Technology (Austria) - An advanced language understanding agenda

Sail Labs has emerged as one of the world's leading players in the field of natural language understanding of both text and speech and one of the largest dedicated language technology companies in Europe. Founded in the mid 1990s as part of a strategy to create commercially-oriented R&D centres on a global scale, the company has since focused its efforts on developing a language technology infrastructure that underpins a broad range of cross- and multi-lingual systems and applications. Major products include the Conversational System, which enables speech dialogue, and the Media Mining System, an innovative indexing and retrieval system for the broadcast media industry, which won the 'Merkur' award from Austrian enterprise organisations.
www.sail-technology.com

Synthema (Italy) – Tools for multilingual knowledge management

Founded in 1994 by scientists from Italy's IBM Research Centre, Synthema leverages its natural language engines and components over a number of information management applications, including

computer-assisted translation, text and knowledge mining, knowledge-based decision aid, document analysis and voice interfaces. These tools in turn enable services and products for competitive intelligence and customer care in several languages. Strongly research-driven, Synthema participates in a number of national projects ranging from subtitling systems to legal linguistic resources, as well as in European projects.

www.synthema.it

Systran (France) – Industrial-strength machine translation

The world's venerable dedicated machine translation supplier, Systran, has continually managed to reinvent itself to stay in the forefront of commercial translation technology. It offers one of the widest ranges of language pairs available and has an unrivalled track record in supplying translation solutions to government, defence and high technology organisations the world over. It adapted swiftly to the Internet revolution and today offers different products, from gisting to post-edited text, for different end users. Systran translates millions of Web pages every day into 35 language pairs.

www.systransoft.com

Aixplain (Germany) – Cross-lingual solutions for speech and text

The German supplier AIXPLAIN AG was founded by leading researchers at the Department of Computer Science of the Aachen University of Technology. It applies advanced stochastic methods to the three fields of translation, language-driven knowledge management tools and speech technology. Mixing these three basic technology resources, the company can provide various enterprise solutions. Its speech translation system PLAINbabel currently translates spoken utterances between English and German, while its interactive text translation system covers all EU languages.

www.aixplain.com

Grounded in the basics: suppliers of Componentware and Resources

Neurosoft (Greece) – Greek language components for text mining

Neurosoft S.A. was founded in 1994 to design, develop, customise and maintain integrated software systems. In the last few years it has expanded its expertise towards language technology, by developing the lexical and corpus tools to act as a platform for basic text application components. It has now productised these into proofing tools such as spell checkers and hyphenators for various word processing and document management packages. It has also leveraged its language resource base to design its flagship Envisioner system knowledge mining software, which offers a comprehensive Greek language solution for text mining applications.

www.neurosoft.gr

Polderland Language & Speech Technology (Netherlands) - Core components for multiple languages

Founded in 1993 as a Nijmegen University spin-off, Polderland has positioned itself as the leading core language-technology player on the Dutch and Flemish language market. It has developed a set of basic linguistic technologies that help establish the core components that underpin such word-and-rule applications as spell checkers, grammar checkers, dictionaries and thesauri, as well as text-to-speech readers and language identifiers. Its products are found in some of the most popular word processing software solutions, and the company is extending its language coverage from Dutch, Flemish, Frisian, Afrikaans and Papiamentu (Dutch Caribbean) to several dozen global and local languages.

www.polderland.nl

Connexor (Finland) – Embedded multilingual language analysers

Connexor develops and market embeddable language analysers that add linguistic value to OEM information technology products. Founded in 1997, the company first worked on end-user products for
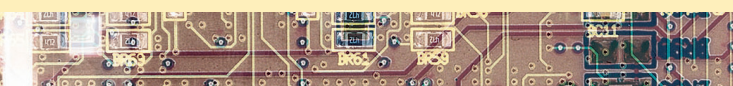
terminology and proofing. It later shifted to component production with its Connexor Machinese product line, which offers fast, light-weight phrase tagger, syntax, semantics and metadata modules that can be deployed in various processing tasks, from text indexing to machine translation. Connexor's clients range from small software houses to very large high technology corporations, and its products operate on operating systems and platforms ranging from PDAs to mainframes across English, French, German, Spanish, Italian, Dutch, Swedish and Finnish.

www.connexor.com

Daedalus (Spain) - Document processing tools for Castilian

Founded as a university spin-off in 1998, Daedalus provides language-processing components for Castilian language software, together with proofing tools for document management. Originally developed within the framework of government technology funding, these tools can be integrated into web-based services to improve the quality of searches and document filtering. The company's flagship product is STILUS Modular, which acts as a general text corrector for spelling, grammar, punctuation, concordance etc. It includes an exploration robot, an indexing unit and an information collection unit, and can be used for a wide range of online activities, from auditing document or file quality to filtering content.

www.daedalus.es

The HLT market

Crossing the chasm with HLT

The transfer of HLT to market has now been through one full 'crossing the chasm' cycle. Use of basic tools and technologies (such as spell-checkers or simple speech-recognition systems) has moved from innovators, to early adopters, and into the mainstream (early majority).

In the first-generation language technology market, *innovators* either had a uniquely compelling requirement (e.g. the use of machine translation by the European Commission and the US Defense Department), or experimented with component technology in innovative ways (e.g. Reuters' early use of NLP for information retrieval for news resources). *Early adopters* exploited the increasing maturity of some HLT products for very specialised purposes, e.g. the use of MT for technical publishing by Xerox and Caterpillar, and the introduction of speech recognition in medical transcription systems. HLT has now entered the mainstream, in *early majority* embedded capabilities. Telephone-based speech recognition is now widely used; most major search engines have embedded HLT components; and millions of web pages are translated automatically every day using MT.

This progress is reflected in market spending for HLT. Datamonitor puts the worldwide speech technology market for 2003 at just short of €1 Billion. IDC estimates the current NLP market at around €400 Million. By 2005, the combined speech/NLP market is forecast to exceed €2 Billion. While these are respectable market opportunities in themselves, they do not reflect the multiplier effect of embedded HLT. The value added to products and services employing HLT creates markets worth many times the value of the core technology itself.

The next market stage for HLT - exploiting language knowledge in more complex and advanced applications - will initiate a new cycle of development. In second-generation HLT, innovators will experiment with new combinations of components and tools, while the mainstream market will wait for proven embedded solutions. It is unlikely that second-generation HLT will produce many standalone 'pure HLT' products; instead, language technology will be incorporated into other applications, creating innovative features or superior performance to provide differentiation for mainstream products and services. This likely shape of the future HLT market should direct future language technology research, which
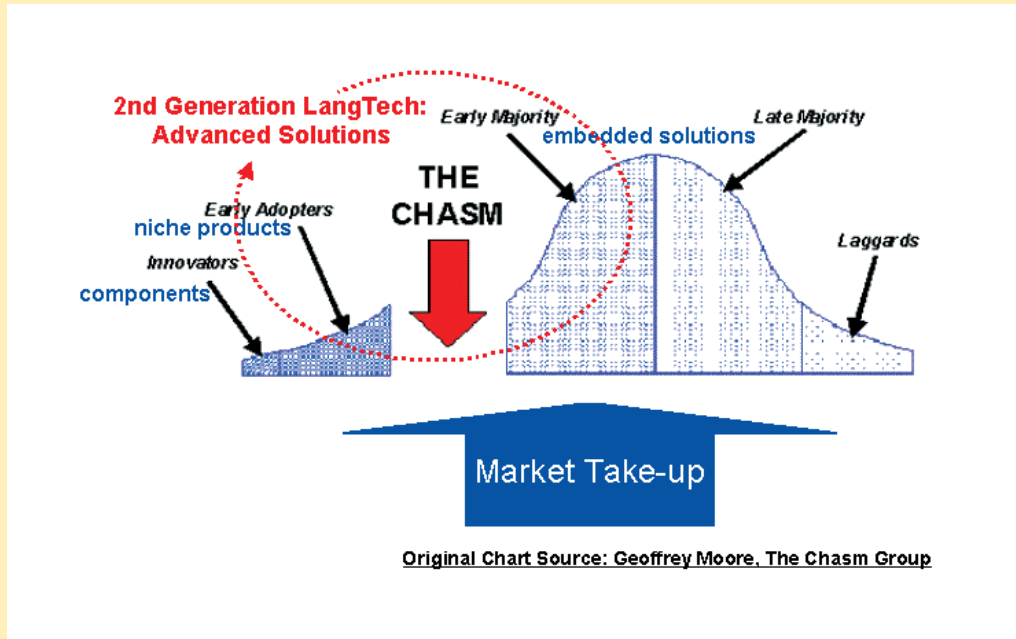
Figure 1: Crossing the Chasm with HLT

European Projects: meeting the needs of future markets

The framework research programmes funded by the European Commission have been well crafted to help the HLT community meet the needs of the market in the HLT field. With this support, and the initiatives of national governments, European HLT research is already actively addressing the future needs of the market for language-enhanced products and services. The following is a showcase of HLT projects that are being carried out (or in certain cases have been completed) under the IST section of the Fifth Framework Programme. They illustrate the breadth of research topics, the range of R&D organisations committed to HLT research in Europe, and the variety of languages involved. They also demonstrate the role that commercial companies have played in the application development and demonstrator phases of certain projects.

will need to be carried out in the context of advanced research in companion, or hosting, technologies.

Case in point: HLT in the Travel & Tourism sector
Projects addressing the needs of the travel & tourism sector illustrate how HLT-enabled IT platforms might work in future advanced electronic services in the EU. Integrated projects with initiatives operating at European, national and local level can have a major impact on future commercial developments in the field, by exploring problems and solutions in life-size applications at relatively low risk to the community.

CATCH 2004
CATCH-2004 aims to provide citizens throughout the European Union with multilingual access to a wide range of transactional and information services and systems offered by providers from both the public and the private sector. Access will be provided using a unified architecture across a range of devices including kiosks, telephones and Wireless Application Protocol (WAP) enabled appliances. Multilingual speech recognition and synthesis together with voice browsing methods based on the WAP architecture form the key features of a system that will interact with users through speech, text and graphics depending on the characteri-

stics of the access device. Information systems and service providers in Athens and Helsinki will test and validate CATCH-2004 components in specially selected application scenarios.

http://www.catch2004.org/

Participants

IBM France (F)	Organising Committee for the Olympic Games Athens 2004 (GR)
Nokia Corporation (FIN)	
Elisa Communications (FIN)	National Technical University of Athens (GR)
Stadt Köln (D)	
Gerhard-Mercator-Universität-Gesamthochschule Duisburg (D)	Hellenic Telecommunications Organisation (GR)

M-PIRO

The M-PIRO project is developing personalisation solutions based on HLT. In order to provide information which is tailored to user interests or preferences, it needs to be stored in a different, non-textual fashion, and turned into a spoken or written message in the language of the user when accessed. This generation process can take context and personal preferences of the user into account, so that it looks as if the object has been stored all along as a Personalised Information Object. The project is developing the HLT aspects of Personalised Information Objects concentrating on multilingual information delivery in virtual and other museum settings.

http://www.ltg.ed.ac.uk/mpiro/

Participants

University of Edinburgh (UK)	National Centre for Scientific Research "Demokritos" (GR)
Istituto Trentino di Cultura (IT)	National and Kapodistrian University of Athens (GR)
System Simulation Ltd (UK)	Foundation of the Hellenic World (GR)

MIETTA

By combining advanced language and information processing technologies, MIETTA provides a flexible, cross-lingual system that allows users to search the Web for tourist information in their own language, and simultaneously supports the information provider with an integrated solution to generate and offer this information in different languages.

http://www.dfki.de/lt/mietta/

Participants

University of Edinburgh (UK)	Advance Bank (DE)
Lloyds-TSB Group (UK)	Deutsche Bahn Reise & Touristik (DE)
British Midland Airways (UK)	Loquendo (IT)
Periphonics Voice Processing Systems (UK)	Comune di Roma (IT)
	SARITEL (IT)

NESPOLE

The NESPOLE project is developing a solution for multilingual and multi-modal negotiation in e-commerce and e-service by providing a robust, flexible, scalable and portable speech-to-speech translation system. It will deliver a first system geared towards tourism that will be used to develop another with a larger coverage of the domain and richer interaction modalities.

http://nespole.itc.it/

Participants

Istituto Trentino di Cultura (IT)	Université Joseph Fourier (FR)
Universität Karlsruhe (DE)	AETHRA S.r.l. (IT)
Carnegie Mellon University (US)	The Trentino Tourist Board (IT)

SPOTLIGHT

The project aims are to research innovative methods to extend the spoken natural language and speech recognition capabilities of these European services and bring their successes under public 'spotlight'. The results of the research will push the technology boundaries of telephone-based spoken natural language user interface capabilities within Europe beyond today's limits of information provision - where customers can find out details of a bank account or travel timetable - to new mass market eCommerce channel interfaces using speech recognition.

http://spotlight.ccir.ed.ac.uk/

Centre for Communication
Interface Research (UK)
bmi British Midland (UK)
Deutsche Bahn (DE)
Comune di Roma (IT)

Lloyds TSB (UK)
Nortel Networks (UK)
Loquendo (IT)
Lastminute.com (UK)

ARISE

The ARISE-project (Automatic Railway Information Systems for Europe) started in October 1996 and lasted 24 months. The aim of the project was to improve speech recognition, understanding and user-oriented dialogue strategies. The partners of ARISE had to develop prototypes in four languages (English, French, Italian and Dutch). the Italian system (INFORMA) and the Dutch system (VIOS) already work since 1998.

http://www.compuleer.nl/arise.htm

Participants

CSELT (IT)
Ferrovie dello Stato S.p.A. (IT)
Institut de Recherche en
Informatique de Toulouse (FR)
Katholieke Universiteit
Nijmegen (NL)
KPN Research (NL)
Rheinisch Westfälische
Technische Hochschule (DE)

LTV - La Tipografica Varese (IT)
Nederlandse Spoorwegen (NL)
LIMSI (FR)
Philips (DE)
Saritel (IT)
SNCF (FR)
Vecsys (FR)
Openbaar Vervoer
Reisinformatie (NL)

Pushing the envelope: European research for advanced HLT applications

European HLT research is already actively addressing the future needs of the market for language-enhanced products and services. The following is a selection of HLT projects that are being carried out (or in certain cases have been completed) under the IST section of the Fifth Framework Programme. They illustrate the breadth of research topics, the range of R&D organisations committed to HLT research in Europe, and the variety of languages involved. They also demonstrate the role that commercial companies have played in the application development and demonstrator phases of certain projects.

NAMIC: News Agencies Multilingual Information Classification

NAMIC is using intelligent information extraction technologies to develop a system for multilingual news customisation and dissemination services based on XML and news industry standards. The project's components include a user profile manager, automatic (multilingual) authoring tools, a cross-linguistic linker and a hypernews Engine, which are being combined into an integrated NAMIC prototype. The project is carrying out user trials at media sites in the UK (*The Financial Times*, for international financial news in English), Italy (ANSA a news agency working in Italian), and Spain (EFE, a news agency working in Spanish).

http://namic.itaca.it

Participants

ITACA (IT)
University of Sheffield (UK)
University of Roma Tor
Vergata (IT)
Universitat Politècnica de
Catalunya (ES)

Vrije Universiteit Brussel - VUB (BE)
Agenzia ANSA (IT)
Agencia EFE (ES)
Financial Times (UK)

D'HOMME: Dialogues for the Home Machine Environment

D'HOMME has developed methods and architectures for building and configuring speech interfaces to networks of small programmable devices in the domestic machine environment. The project addresses questions related to the nature of dialogues between humans and domestic devices, the processing architectures & representations for dialogues, and, the impact of reconfigurable device networks and language processing components. The project is building baseline demonstrators in English, Spanish & Swedish and exploring and evaluating reconfigurability methods for plug-and-play device networks.

http://www.ling.gu.se/projekt/dhomme

Participants

SRI International (UK)
Universidad de Sevilla (E)
Telia Research (S)

Göteborg University (S)
University of Edinburgh (UK)
NetDecisions (UK)

CORETEX: Improving Core Speech Recognition Technology

CORETEX is dedicated to improving core speech recognition technologies in order to lower the cost and effort of porting speech applications to new languages and environments. The project is developing generic, application-neutral speech recognition technologies that work well for a wide range of tasks. Key project foci include techniques for producing enriched symbolic speech transcription for higher level (symbolic) processing, methods for improving language models and for extending vocabulary size by automatic pronunciation generation, and developing an evaluation framework to assess improvements. The focus languages are Italian and German. The project has also set up a user group of representatives of European companies in different fields - information technology, telecommunication, broadcasting, and multimedia archives, to help identify the most pressing research issues and problems of industrial application domains.

http://coretex.itc.it

Participants

Rheinisch-Westfaelische Technische Hochschule Aachen - RWTH (DE)	University of Cambridge(UK)
	Istituto Trentino di Cultura - IRST (IT)
	Centre National de la Recherche Scientifique - CNRS (FR)

DEEP THOUGHT: Hybrid and Shallow Methods for Knowledge Retrieval

Modelling the meaning of texts to ensure better knowledge retrieval is a major challenge for natural language processing. DEEP THOUGHT is using a novel combination of deep and shallow methods that will extend mainstream information over multiple language retrieval operations, using high precision concept indexing and relation detection. The project will apply statistical methods for tokenisation, IR indexing and search, POS-tagging and chunk parsing; word nets and domain ontologies for conceptual indexing and detection of ambiguity and polysemy; weighted finite-state transduction technology for named entity and simple relation detection; and shallow parsing for the detection of complex and covert relations. By automating many of these tasks, the project hopes to boost performance in knowledge management tasks spanning the information society spectrum.

http://www.eurice.de/deepthought/index.htm

Participants

Saarland University (DE)	Norges teknisk-naturvitenskapelige universitet (NO)
The University of Sussex (UK)	CELI (IT)
The University of Cambridge (UK)	XtraMind (DE)

E-MATTER: E-Mail Access through the Telephone Using Speech Technology Resources

E-MATTER has developed a multilingual spoken language dialogue system providing access to e-mail through the telephone network. The system is able to operate in all official languages in Spain —Spanish, Catalan, Galician and Basque — together with English and French. The system can automatically identify an email's language, correct any misspelling errors and read the message via a text-to-speech converter in the appropriate language. The highly configurable system also allows users to connect to a Windows-based interface to configure such profile parameters as native language and e-mail filters. A prototype has been successfully integrated into Telefónica's I+D voice portal, and by Terra in Catalonia.

http://www.ub.es/gilcub/e-matter/index.html

Participants

Telefónica Investigación y Desarrollo, Unipersonal (ES)	Universitat de Barcelona (ES)
	Terra Networks (ES)
Universidad Politécnica de Madrid (ES)	

FASiL: Flexible and Adaptive Spoken Language and Multi-Modal Interfaces

One of the most generously funded projects under the Fifth Framework Programme, FASiL aims to pilot a full multi-modal voice portal application that is 3G mobile-network ready, along with tools for rapid development of new applications. It will develop a robust and scalable Virtual Personal Assistant to manage e-mail, calendar and agenda through an intelligent, friendly adaptive multi-modal interaction in at least three European languages - English,

Portuguese and Swedish. The project will call upon state of the art speech and language technologies and includes two charities as partners to ensure that the results will meet the needs of the highest assistive technology standards.
http://www.fasil.co.uk

Participants

Portugal Telecom Inovação (P)	Royal National Institute of the Blind (UK)
SpeechWorks UK (UK)	
University of Sheffield (UK)	Royal National Institute for Deaf People (UK)
Media Lab Europe(IE)	
Vox Generation (UK)	Cap Gemini Ernst & Young (SE)

i-EYE: Interacting with Eyes: Gaze Assisted Access to Information in Multiple Languages

i-Eye is designing an innovative interface that utilises gaze tracking, together with speech input, for systems that respond and react to user eye movements. This is intended to pave the way for new applications of responsive interaction technology. It will evaluate eye tracking in two applications. iDict is a translation tool that tracks and is triggered by cues in users' gaze patterns, offering language aid in areas where they are having particular difficulties. ITutor is a multimedia application to support hand-free maintenance activities. The application will communicate either automatically in response to the user's gaze and/or via a speech interface and will also make use of wearable computing technology. The first prototype supports English to Finnish, German and Italian.
http://www.cs.uta.fi/research/hci/ieye/

Participants

University of Tampere (FI)	GIUNTI Interactive Labs (IT)
Connexor (FI)	University of Nottingham (UK)
Sensomotoric Instruments (DE)	

LC-STAR: Lexica and Corpora for Speech-to-Speech Translation Technologies

The advent of operational speech-to-speech translation will be particularly demanding on high quality, relevant language resources. LC-STAR is creating the lexicons and corpora needed for the basic system components - flexible vocabulary speech recognition, high quality text-to-speech synthesis and speech centred translation. These components will be inte-

grated into speech driven interfaces embedded in mobile devices and network servers. The lexicons for twelve languages (Catalan, Finnish, German, Greek, Hebrew, Italian, Mandarin Chinese, Russian, Spanish, Standard Arabic, Turkish and US-English) will include phonetic, prosodic and morpho-syntactic content, and the corpora will cover bilingually aligned text in the tourism domain for three prototype languages - Catalan, Spanish and US-English. The tourist scenario will also be used to demonstrate the speech-to-speech prototype in these languages. LC-STAR will make the project data available to research institutes and companies worldwide for further exploitation in research and commercial applications.
http://www.lc-star.com/

Participants

Siemens (DE)	Rheinisch-Westfälische Technische
IBM Deutschland (DE)	Hochschule (DE)
Universitat Politècnica	Natural Speech Communication (IL)
de Catalunya (ES)	Nokia Corporation (FI)

M4: MultiModal Meeting Manager

Imagine generating the minutes of a multi-party meeting automatically. M4 is an ambitious project dedicated to developing a demonstration system to enable structuring, browsing and querying of an archive of a meeting in a room equipped with multimodal sensors. This includes the creation of a 'smart' meeting room, and its associated multi-modal meetings database, the recognition of speech, actions and emotions from multiple audio/video streams, and the subsequent management of all the data and knowledge produced, including retrieval, summarisation and access. The technology to be developed crosses several disciplines and will be implemented in a demonstrator.
http://www.dcs.shef.ac.uk/spandh/projects/m4/

Participants

University of Sheffield (UK)	TNO (NL)
EPFL (CH)	University of Geneva (CH)
Technische Universitaet	University of Twente (NL)
Munchen (DE)	Brno University of Technology (CZ)
IDIAP (CH)	

MKBEEM: Multilingual Knowledge Based European Electronic Marketplace

MKBEEM adds comprehensive multilinguality support to robust e–commerce platforms, including multilingual content generation and maintenance, automated translation and interpretation and enhancing the natural interactivity and usability of the service with unconstrained language input. The project is validating its system on Tourism, Mail order, and Business-to-Business portals. The aim is to facilitate the flow of information independently of the language of the user, the service, or the content provider. Ontologies will be used for classifying and indexing catalogues, for filtering user's queries, for facilitating multilingual man-machine dialogues, and for inferring information that is relevant to the user's request. The system supports Finnish, English and French as well as Spanish and Swedish, and has been trialled on active websites.

http://mkbeem.elibel.tm.fr

Participants

France Télécom - CNET (FR)	Universidad Politécnica de Madrid (ES)
Sema Group (ES)	Technical Research Centre of
University of Montpellier (FR)	Finland -VTT (FI)
Société FIDAL (FR)	Ellos Postimyynti (FI)
Tradezone International (UK)	Société Nationale des Chemins
National Technical University	de Fer Français - SNCF (FR)
of Athens (GR)	

MUMIS: Multi-Media Indexing and Searching Environment

MUMIS aims to produce a prototype for accessing dynamic multimedia information (sound, image and text) via any channel. The technologies under development will automatically create indexes in multimedia content, using formal representations of contents from free text, noisy spoken accounts and image understanding. Information from these sources will be combined into a multi-layered data structure, based on an ontology for the soccer match domain. Using a search engine, users will be able to search for specific sets of events and retrieve the corresponding multi-media fragments. The project involves developing new uses for existing automatic speech recognition and information extraction to create for-

mal annotations for each data source and language. A novel type of merging tool will be developed to maximise the coherence of the event descriptions, and a user interface for users to interact with the domain knowledge will be built as a web showcase.

http://parlevink.cs.utwente.nl/projects/mumis/index.html

Participants

University of Twente, Centre	The University of Sheffield -
for Telematics and	USFD (UK)
Information Technology (NL)	Max Plank Institute (NL)
Stichting Katholieke	Deutsches Forschungszentrum für
Universiteit Nijmegen -	Künstliche Intelligenz - DFKI (DE)
KUN (NL)	ESTeam AB (SE)

SALT: Standards-based Access service to multilingual NLP-Lexicon and human-oriented Terminology resources

Locale customisation is increasingly a technology-intensive and collaborative activity. SALT has developed a range of XML-based formats and tools for modelling, representing, and exchanging terminological data so that humans (terminology managers, translators, technical writers, localisers) and technology systems (machine translation and translation memory) can (re)use and share this data more cost-effectively and efficiently. The two key international formats produced are OLIF (Open Lexicon Interchange Format), which focuses on the interchange of data among lexbase resources from various MT systems and MARTIF (MAchine-Readable Terminology Interchange Format, ISO 12200), which facilitates the interchange of termbase resources with conceptual data models ranging from simple to sophisticated. The project also emphasised the derivation, integration, and interfacing of ontologies and data structures in translation and localisation environments.

http://www.loria.fr/projets/SALT/

Participants

Institut für Übersetzer und	SEEITM (UK)
Dolmetscherausbildung,	nstitut National de Recherche en
University of Vienna (AT)	Informatique et Automatique -
Institut für Informations-	INRIA-LORIA (FR)
management, University	Universität des Saarlandes (DE)

of Cologne (DE)
Accademia Europea
di Bolzano per la ricerca
applicata ed il perfeziona-
mento professionale (IT)
University of Surrey -

Brigham Young University
Translation Research Group (US)
Kent State University Institute for
Applied Linguistics (US)

MUCH.MORE: Multilingual Concept Hierarchies for Medical Information Organisation and Retrieval

MUCH.MORE has developed a framework for integrating existing and emerging technologies in order to boost the efficiency of automating cross-lingual information organisation and access for the medical domain, a critical requirement in the information society. The main focus has been on combining statistical, knowledge-based and heterogeneous methods and resources. The approach relies on existing rich concept hierarchies in the medical domain (International Classification of Diseases (ICD), Medical Subject Headings (MESH) and the Unified Medical Language System (UMLS)), and on classified document collections. The project's languages are German and English, with some work on Chinese.
http://muchmore.dfki.de/

Participants

Deutsches Forschungs-
zentrum für Künstliche
Intelligenz - DFKI (DE)
Carnegie Mellon
University - CMU LTI (USA)
Eurospider Information
Technology - ETI (CH)

Xerox Research Centre Europe -
XRCE (FR)
Stanford University - CSLI (USA)
ZInfo Universitätsklinikum (DE)

VICO: Virtual Intelligent Co-Driver

In-vehicle interactive communications are a major social, psychological, technological and economic challenge for the research and development community. The VICO project is developing a virtual intelligent co-driver interface that enables conversational interaction between the human driver and digital devices and services in the car. The project is developing and testing natural-sounding access to services such as interactive hotel reservation, customised sightseeing tours, on-the-fly route planning, and an electronic car manual. Technologies will include robust speech recognition for adverse envir-

onments, a natural language understanding component, a safe-to-use vocal interface, adaptive dialogue management strategies and multilingual conversational information and communication services within Europe.

Participants

Robert Bosch (DE)
DaimlerChrysler (DE)
Phonetic Topographics (BE)

Istituto Trentino Di Cultura -
IRST (IT)
University Of Southern Denmark (DK)

TRUST

This project is developing a multilingual, semantic and cognitive search engine for text retrieval using semantic technologies. Four interactive semantic multilingual search engine prototypes are under development in French, Italian, Polish and Portuguese. They are aimed at the general consumer market within the cost range of 30 to 90 Euro and will feature queries in natural language, cross-lingual capabilities and human like interpretation, the ability to retrieve relevant data without overload or underload in the four languages and the ability answer refined questions in delivering semi or fully automatic corpus synthesis. The test domains for these search engines will be in the environment and historical demography.
http://www.trustsemantics.com

Participants

Expert System
Solutions Srl (IT)
Priberam Informatica
Lda (PT)

Synapse Développement (FR)
Convis GmbH (DE)
TiP sp. zo.o (PL)

TT2: TransType2 - Computer-Assisted Translation

TT2 is developing a robust Computer-Assisted Translation (CAT) system to help meet the growing demand for high-quality, high-throughput translation services. The consortium is embedding a data-driven machine translation engine into an interactive translation environment, combining human quality control with translation automation productivity gains. TT2 can support both text and speech input, and will cover six translation pairs from French, Spanish and German to English and back. The proto-

type will be evaluated by two professional translation agencies to ensure that the system meets the need of the professional translation environment. http://tt2.sema.es

Participants

SEMA Group (ES)	RALI - University of Montreal (CA)
RWTH Aachen (DE)	Celer Soluciones (ES)
Instituto Tecnológico	Société Gamma (CA)
de Informática (ES)	Xerox Research Centre Europe (FR)

SIRIDIUS: Specification, Interaction and Reconfiguration In Dialogue Understanding Systems

SIRIDIUS is improving understanding of what is needed for reusable, robust and user-friendly spoken dialogue systems by building two demonstrators - an automated telephone operator in Spanish, and an integrated toolset for building dialogue systems. The aim is to handle unpredictable speech in noisy environment, develop generic strategies for dialogue management that can be applied to a wide range of dialogues including command and negotiative dialogues, and provide architectures which allow appropriate sharing of information between modules, in particular enabling dialogue systems to be sensitive to how users stress individual words. http://www.ling.gu.se/projekt/siridus/

Participants

Goeteborg University (SE)	Telefónica Investigación y
University of Saarland (DE)	Desarrollo (ES)
Universidad de Sevilla - USE (ES)	SRI International (UK)

The European HLT research base

The language technology community in Europe has managed to remain competitive against strong HLT research initiatives in the US and Japan, as well as the growing levels of R&D in other parts of Asia (especially translation technology in China, Korea, and India). Indeed, HLT is one of the few areas of software research where European research is clearly world class.

Importance of EU support

Language technology research has been supported for many years within EC Framework Research Programmes, and the timing and structure of that support has been well suited to the needs of the HLT domain. Up to the mid-1990s the research programmes had a technology-push focus that was very effective for HLT, as market conditions were not yet favourable. Funding for language engineering in FP4, and the HLT action in FP5, has been more market-focused, and has tracked the evolution of market opportunities for language technology very closely. EU funding has been particularly important for the HLT domain, and has been largely responsible for the creation of a coherent research community in Europe. Industry-sponsored research in HLT has been weak, though stronger in speech than in NLP. Moreover, national-level public support for HLT research has been highly variable, and with a few notable exceptions, somewhat inconsistent. EU funding for machine translation research produced a network of NLP researchers across the EC, and spawned a variety of research efforts in different languages, as well as an established academic base for MT experts. In addition, the tendency to fund a larger number of smaller projects (compared to the practice in the US and Japan) has had the effect of broadening the technical base across the Union; at the same time the structure of FP projects, requiring cross-border collaboration, has created a genuinely pan-European research base.

EU funding has, in addition, had a significant impact on technology transfer in the HLT field, though not always a direct one. The number of suppliers active in the HLT market has expanded exponentially in the last decade, from fewer than 30 companies in 1993, to 10 times that many in 2003. Almost all these European suppliers have some roots in EU-funded programmes, either through technology inherited (often through several generations) directly from projects, or through the technical capacities of experts who have been involved in projects.

HLT Research Showcase

There are some 300 R&D labs carrying out research in various aspects of language and speech technology in Europe. These range from dedicated language technology research centres to small departments or groups in university computational linguistics departments. Every country in the Union has active R&D groups in the language and speech technology field, and the total active headcount of researchers working in the various disciplines of language and speech technology totals more than 3,500. Industrial research laboratories, funded within the private sector, represent some 10-15 % of this research base in Europe.

The following is a very small sample of Europe's language technology research centres. It is intended to show the variety of organisations and activities in this sector, and is in no way intended to reflect a value judgement on the relative performance of different R&D centres within each country represented.

ÖFAI - Austrian Research Institute for Artificial Intelligence - Natural Language Processing Group (Vienna, Austria)
Natural Language Processing has been a major research area at the Austrian Research Institute for Artificial Intelligence (ÖFAI) since its inception in 1984. The ten-strong group's focus is on constructing linguistic resources, processing speech and text algorithms, and developing application prototypes (such as natural language interfaces, advisory systems and concept-to-speech systems). It is a member of the EU's Network of Excellence (ELSNET) and has participated in a number of EC and national projects.
www.ai.univie.ac.at/oefai/nlu/

Centre for Computational Linguistics, Katholieke Universiteit (Leuven, Belgium)
The Centre for Computational Linguistics at the Louvain Catholic University was founded in September 1991 to promote basic research in formal and computational linguistics, and the application of this research in natural language processing. The Centre builds on the expertise acquired by the Leuven Department of Linguistics during the 1980s within the framework of various NLP projects, and today focuses on machine translation, computational syntax and semantics, corpus linguistics, automatic transcription, NLP tools and resources especially for Flemish and computer-aided language learning & course development.
www.ccl.kuleuven.ac.be/

CST - Center for Sprogteknologi (Copenhagen, Denmark)
The Danish Centre for Language Technology is a government research institute with 20 employees whose mission is to carry out and promote strategic research and commercial development in the areas of language technology and computational linguistics in Denmark, especially for the Danish language. Focus is on machine translation, lexicography, semantic web, ontologies, HLT tools, Danish and a number of foreign languages. The Centre has participated on several dozen HLT projects that range from EC support projects to Nordic country language technology projects. It has been particularly associated with machine translation through its development of the PaTrans system.
www.cst.dk

Speech-based and Pervasive Interaction Group, Tampere University (Tampere, Finland)
The Speech-based and Pervasive Interaction Group forms part of the Tampere Unit for Computer Human Interaction and Department of Computer and Information Sciences. Its multilingual research agenda covers speech and audio-based applications and mobile and ubiquitous systems, with a strong focus on the ergonomy of computer-human interaction as well as on user interface models, techniques and architectures. It is closely involved in a national university-industry User-Oriented Information Technology programme alongside numerous research groups.
http://www.cs.uta.fi/research/hci/spi/

VTT Information Technology –
Language Engineering Group (Espoo, Finland)
VTT Information Technology is one of the six research institutes at the 60-year old VTT Technical Research Centre of Finland, an independent organisation. The eBusiness group carries out research in language technology for developing and transferring new language processing techniques for industrial uses. Recent focus has been on developing authoring and translation systems, ontology management and other knowledge-related applications for web-based business applications. The group has a strong multilingual focus, and has participated in numerous EC projects, most recently MKBEEM.
http://www.vtt.fi/tte/language/

LIMSI - Spoken Language Processing Group
(Paris, France)
Part of the Human-Machine Communication Department, the LIMSI Spoken Language Processing Group is a world-class laboratory of some 25 people with a historical role in the development of speech technologies in France and Europe, and with strong ties to research centres in the USA. The group's three key research directions are acoustic and lexical modelling, linguistic modelling for dynamic contexts, and recognition and dialog systems, including a focus on audio indexing and searching. Languages covered include English, French, German, Mandarin, Portuguese, Spanish, and Arabic. The Group participates in many EC and industry-driven projects, and has a number of demonstrators available for indexing and large vocabulary speech recognition.
http://www.limsi.fr/Recherche/TLP/PageTLP.html

XRCE - Xerox Research Centre Europe
(Grenoble, France)
One of a number of Xerox Research facilities, this XRC based in Grenoble, France is dedicated to being a Centre of Excellence for pure and applied research into multilingual document technologies. The centre's research competencies are organised around long-term activities for contextual computing, with a special focus on finite state technology, machine learning, robust parsing, semantics and document content models. It also runs a Programme of Advanced Technology Development, which engineers market-ready knowledge services, linguistic tools and document management solutions for technology transfer.
www.xrce.xerox.com/

Daimler Chrysler Research and Technology Speech Understanding Group (Ulm, Germany)
Daimler Chrysler's Speech Understanding Group, located at the car-maker's R&D centre in Ulm, has long been active in research in speech recognition, speech synthesis and language analysis. As an industrial laboratory, the Group's aim is to conduct solutions-driven R&D into vehicle interfaces. One example was the fundamental research that led to the Linguatronic system, first integrated into a Daimler vehicle in 1996 via Daimler's TEMIC business unit. The Speech Group has some 30 researchers who work closely with academic researchers in fields such as sub symbolic information processing in adaptive sensorimotor systems.

DFKI - Language Technology Lab
(Saarbrücken & Kaiserslautern, Germany)
Germany's prestigious DFKI (AI Research) includes a well-resourced Language Technology Lab based in Saarbrücken, one of Europe's major facilities for exploring advanced language technology. It conducts research across the complete range of fundamental and advanced language tools, and works closely with the Departments of Computational Linguistics and Computer Science at the University of Saarland, and has been involved in many national and international projects which cover languages such as English, French, Japanese, Chinese, Italian, Dutch, Spanish languages. The current focus is on ontology engineering, enriching deep processing with statistical methods, and exploring new ways of authoring and visualising documents.
www.dfki.de/lt

ILSP - Institute for Language and Speech
Processing (Athens, Greece)
The Institute for Language and Speech Processing
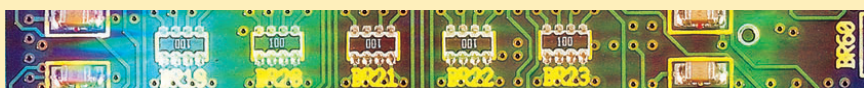
was founded in Athens with the aim of becoming a Greek centre of excellence in language technology. With its 90-strong staff, it has departments of e-lexicography, language technology applications, educational technology, speech technology, machine translation and has developed a critical mass of expertise in transferring research results into language technology products, ranging over speech modules, language learning programs, and proofing and translation tools. ILSP has also participated in numerous EC and national projects in these domains ever since the Eurotra project in 1985.

www.ilsp.gr

NCLT - The National Centre for Language Technology, Dublin City University (Dublin, Ireland)
Ireland's leading dedicated facility in the field, the National Centre for Language Technology has been exploring a broad range of language technology areas since the early 1990s. With a staff of around 15, research foci include speech recognition and production, translation, human-computer interfaces, information retrieval and extraction from the worldwide web, the teaching and learning of languages using computers and software localisation and globalisation. The centre has also forged strong links with industry.

www.computing.dcu.ie/research/nclt/

ITC-irst - Istituto Trentino di Cultura, Cognitive Communication Technologies and Interactive Sensory Systems Divisions (Trentino, Italy)
The Centre for Scientific and Technological Research (ITC-irst) at Trentino has since 1988 focused on various aspects of natural language processing, and the development of intelligent interfaces and content processing systems through two divisions - Communication and Cognitive Technologies division and Interactive and Sensory Systems. The Centre has a staff of over 50 multidisciplinary language technology-related researchers, and is currently working on intelligent interfaces, automated reasoning, computational humour. It is active in technology transfer and has worked extensively on international projects.

http://www.itc.it/

ILC-CNR: Institute for Computational Linguistics (Pisa, Italy)
The Italian Computational Linguistics Institute (ILC-CNR) is one of Europe's historic language technology R&D facilities. It was created in 1978 to develop an agenda for interdisciplinary research issues in the field of language automation in general. The institute has played a lead role not only in Italian language technology development, but also in stimulating cross-border HLT research in Europe. Today it forms a Centre of Excellence in Italy and internationally, with a special focus on developing international standards and evaluation methods for language resources and technologies.

www.ilc.it

OTS - Foundation for Language Technology, Utrecht institute of Linguistics (Utrecht, Netherlands)
The Stichting Taaltechnologie (Foundation for Language Technology or STT) was established by Utrecht University in 1986. It acts as an intermediary between the funding body and project teams at the Utrecht Institute of Linguistics OTS. Drawing on Institute and other Dutch R&D centre resources, the STT has had a long tradition of collaborative research on large-scale EC natural language processing projects. The Institute of Linguistics itself is currently undertaking a wide-ranging Language in Use programme focusing on the interface between language and cognition, and covering such fields as language acquisition, grammatical modelling and prosody.
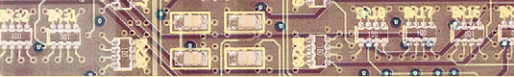
www-sk.let.uu.nl/stt/

INESC-ID: L^{2}F Spoken Language Systems Lab (Lisbon, Portugal)
The Spoken Language Systems Lab (L^{2}F - Laboratório de sistemas de Língua Falada) was created in January 2001, as part of Lisbon-based Instituto de Engenharia de Sistemas e Computadores (Institute for Systems and Computer Engineering). The aim was to harness the skills of various national research groups to promote computational processing of spoken language for European Portuguese. L^{2}F is actively involved in all aspects of speech recognition, synthesis

and coding, with a special emphasis on developing and validating spoken linguistic resources. The Lab's mission is to promote spoken language systems for Portuguese, at a national and international level, and it has participated in numerous EC and national projects covering speech technology in telecommunications and assistive contexts.

www.inesc.pt

TALP - Te Llenguatge i la Parla, Universitat Politèc-nica de Catalunya (Barcelona, Spain)

The TALP is a research centre comprising two R&D groups dedicated to Natural Language Processing and to Speech, and forms part of the Universitat Politècnica de Catalunya. It was created in the mid-1980s and today has over 40 researchers split equally between the two main R&D domains. One of TALP's key features is the fact that it provides a platform for conducting R&D into application domains where speech and text are combined in new ways. The NLP Group focuses on generating multilingual lexical resources, interface design and core language components in Spanish, Catalan and English. The Speech Group works on all aspects of speech technologies for both research and industrial development projects.

www.talp.upc.es

KTH - Centre for Speech Technology (Stockholm, Sweden)

Drawing on a long Swedish tradition of research into spoken language, the CTT was established within the KTH (Royal Institute of Technology) in 1996 as a sustainable platform for co-operation between Swedish companies, non-commercial organisations and academic research in the field of speech technology, with particular emphasis on robust systems for the Swedish language. Current projects include speech technology in interactive dialogue systems, language modelling for spoken languages, and the development and testing of advanced acoustic models for speaker characterisation in speech synthesis.

http://www.speech.kth.se/ctt/

BTexact Technologies – Future-oriented interfaces and knowledge solutions (UK)

BTexact Technologies is the advanced communication technologies business of British Telecom (BT). With facilities in the UK, USA and Malaysia, BTexact technologies is a world-class telecommunications engineering research unit with a long and rich heritage of communications technology at one of Europe's largest concentrations of communications technologists. The lab is future-oriented, venture-friendly, multidisciplinary and highly experimental, and explores areas in which language technology is embedded in next-generation solutions that add business value in a networked economy. BTexact's areas of current language technology interest include domain ontology management, knowledge discovery, intelligent agents, semantic web services, and intelligent customer-contact interfaces.

www.btexact.com

The Language Technology Group (Edinburgh, UK)

The LTG is a part of the Human Communication Research Centre in Edinburgh, which can draw on the skills and expertise of one of the largest communities of natural language processing specialists in Europe. LTG's large-volume text handling work is application oriented in the areas of text annotation, mark-up architectures (XML tools), information extraction (named entities) and automatic or computer-assisted generation of text (museum object descriptions). Projects are also underway in the area of semantics for restricted language processing, and in rich document structure mark-up. The Group has participated regularly in international benchmarking exercises such as the Message Understanding Competition.

http://www.ltg.ed.ac.uk

Natural Language Processing Research Group (Sheffield, UK)

The Natural Language Processing Research Group at the University of Sheffield has been in existence nearly nine years and is one of the largest and best known in the UK. The Group's major research foci are

the use of coded representations of meaning content, belief and knowledge; Machine learning techniques to derive data from sources such as the web; and providing software architectures to underpin NLP research. The Group's GATE architecture has been installed at over 400 sites world-wide. The Group has also been successful in international competitions for best computer conversationalist, best question-answerer system etc, and has regularly participated in EC and global projects in the field. http://nlp.shef.ac.uk/

HLT research in the New Accession Countries

When the 12 New Accession Countries (NACs) from the Baltic (Estonia, Latvia and Lithuania), Eastern and Central Europe (Poland, Czech Republic, Slovakia, Hungary, Slovenia, Bulgaria and Romania) and the eastern Mediterranean (Cyprus and Malta) join the European Union in the near future, the number of languages spoken and used as official tongues in Europe's multilingual mix will nearly double. This means that at least eleven languages (Cyprus being Greek speaking) will need to be brought up to digital cruising speed to ensure equality of access to the information society for all Member State citizens.

One of EUROMAP's tasks was to make a preliminary assessment of the human and institutional assets of the global language and speech technology communities in these countries and to contextualise this information with respect both to the EU experience and to local opportunities and barriers for growth in the domain. It was considered premature to attempt to measure HLT 'readiness' on a NAC national level, applying the same HLT Scorecard model that this report has used to benchmark progress in the current European Union. What follows, therefore, is a brief overview of the principle features of this language technology landscape.

Broadly speaking, there is a relatively small but flourishing academic research community in all NACs. This community has developed ongoing relations with R&D organisations in the EU for some years and has in many cases participated in EC-funded projects

dedicated to launching language and speech technology development in the region.

In addition to individual participation from laboratories and research centres in certain EC-funded projects, there have also been projects specifically dedicated to the NAC situation in the EC's INCO programmes under FP4 and 5, which carried out 13 projects in the region, including a broad-constituency TELRI project on corpus development, BALKANET- IST to extend Euro-WordNet, and the current BALRIC-LING focused more specifically on language and speech resource development in Balkan countries. The Swiss-funded DICO-EAST – SCOPES project has been dedicated to dictionary work in certain NAC languages.

Bulgaria

With a population of just over 6.5 million Bulgarian speakers and substantial communities of Turkish (770,543) and Roma (322,641) speakers, Bulgaria speaks a Southern Slavic national language, supported by a major R&D text processing centre (www.lml.bas.bg) and one or two of other computational linguistic departments. There are two English to Bulgarian machine translation systems on the Bulgarian market and a highly successful knowledge support tools supplier (www.sirma.bg/OntoText/AboutUs.html). Some 30% of the population have access to the Internet.

Czech Republic

The Czech Republic has a population of 10.3 million Czech speakers, as well as small populations of Polish, Slovak and German speakers. There are a number of research centres dedicated to computational linguistics and speech recognition and synthesis, especially those at Charles University in Prague (www.ckl.mff.cuni.cz). There are also certain web-oriented language component for translation and dictionaries. Some 12% of the population have access to the Internet.

Estonia

Some 67% of the country's population of 1.4 million speaks Estonian while half of it speaks Russian.
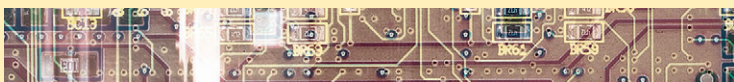
Estonia has a small research centre at Tallinn Technical University (www.phon.ioc.ee), and further groups in Tartu (www.cl.ut.ee). Two external companies supply certain basic language components, and there is also an electronic dictionary and machine translation project underway. Estonia had a nationally-funded general text language programme in the later 1990s, and currently has a text-to-speech programme (www.etf.ee). Some 10% of the population have access to the Internet.

Latvia

With a population of 2.7 million, Latvia has 1.37 million national language speakers and numerous Russian and other Slavic language speakers. It has a number of academic groups working on text and speech processing projects and one successful localisation company, Tilde (www.tilde.lv) which is making its mark as a Baltic language technology component developer. Some 10% of the population have access to the Internet.

Lithuania

Lithuania has a population of 3.6 million, 10% of them Russian speakers. Research is carried out in speech and text technology at various universities, notably Vytautas Magnus University (www.ktu.lt) and the government is currently funding a programme to promote the Lithuanian language in the information society. The Latvian company Tilde (www.tilde.lv) supplies basic language technology components. Some 11% of the population have access to the Internet.

Hungary

With a population of 10 million, and a relatively well-established ICT infrastructure, Hungary has a relatively advanced language technology research base, with corpus linguistics, text, speech and artificial intelligence groups all at work in universities (www.nytud.hu). Hungary also has a successful text component supplier, Morphologic (www.morpholoigc.hu), as well as a subsidiary of the US firms ScanSoft and Mindmaker. Some 20% of the population have access to the Internet.

Malta

The smallest of the NACs, Malta has a total English and Maltese speaking population of 384,000. The university has a dedicated language engineering group (www.mdina.cs.um.edu.mt/mike/rdg/language-engineering.html) working on Maltese and English language R&D issues. Some 6% of the population have access to the Internet.

Poland

Poland has the largest population of the NACs with 38.6 million Polish speakers, plus 10 million diasporic Poles mostly in the USA. The country also has a number of speakers of other lower-density languages. The country has half a dozen dedicated language and speech research groups (www.ippt.gov.pl/centrum-ACC/CA.html), and one company, Neurosoft (www.neuro.pl), with an R&D division. There are a number of first generation speech applications (often for assistive technology), and an emerging Polish-English machine translation system (http://poleng.amu.edu/pl/opis). Some 20% of the population have access to the Internet.

Romania

With a population of 22.1 million, Romania is the second largest of the NACs, with substantial Hungarian and German speaking populations. It has a strong tradition in linguistics research with two or three key language technology R&D centres including the Centre for Computational Linguistics of the University of Bucharest (www.racai.ro). Some 11% of the population have access to the Internet.

Slovakia

The population of 5.4 million includes a complex mix of regional languages and ethnic groups. Research into language and speech technologies is carried out in four academic centres, above all at the Slovak Academy of Science (www.sav.sk). Some 7% of the population have access to the Internet.

Slovenia

Slovenia has a population of 1.9 million, and one of the more buoyant economies in the NAC community.

It boasts a national Language Technologies Society (http://nl.ijs.si/sdjt) that networks R&D activities and provides a platform for researchers. Some 25 % of the population have access to the Internet.

Benchmarking HLT performance

This EUROMAP Study is based on a benchmarking analysis of the opportunities and achievements of the HLT research effort in Europe. The analysis compared Member States, and created indexes for two broad measures: the robustness of the opportunity to exploit HLT (the 'Opportunity Index'), and the prospects for and success of HLT research and technology transfer (the 'HLT Benchmark').

Factors measured for the Opportunity Index were based on third-party research that rates conditions such as the general environment for research innovation; supply-side factors including ease of business formation, access to key channels (as defined by the EUROMAP study) for HLT, and ability to adopt innovation; and demand-side factors including trade competitiveness, ICT infrastructure, and capacity to absorb innovation. The factors were then weighted to reflect a judgment of their relative significance as a potential success factor for HLT.

Factors measured for the HLT Benchmark were based on EUROMAP desk research and fieldwork. They included depth and breadth of HLT research (in both speech and NLP); funding commitments by both the public-sector and industry; and the breadth of language coverage in research and products (considering both the number and choice of languages processed, and coverage of low-density or minority languages). The measurement of research depth considered whether core HLT components have been fully developed, and also the extent to which more advanced applications are the subject of research or technology development projects.

The Opportunity Index was then mapped against the HLT Benchmark to create the 'HLT Scorecard' - a summary measure that captures the relationship between the two. There was a notably strong correlation between the Opportunity Index and the HLT Benchmark. In general, countries with the most favourable business environment and the most highly developed infrastructure also have the most successful HLT research efforts.

The HLT Scorecard-results

The '**Leaders**' include Germany, the Netherlands and the UK. Each of these countries has enjoyed strong national commitment to HLT research. Germany, which scored highest on the HLT Benchmark, has had consistent, long-term effective national investment in HLT from both the public and private sector ever since the SPICOS project in 1985. The Leaders are judged to be 'market ready' for advanced HLT research.
A '**Strong Potential**' group who scored near or below average on the Opportunity Index, but above average on the HLT Benchmark, includes France, Belgium and Spain. France would have clustered with Leaders on the HLT measure, but scores significantly lower on business opportunity environment measures. These countries have well-developed research communities, and a significant depth of HLT research, so they are in a strong position to exploit HLT as opportunity factors improve, e.g. as rates of Internet use rise and greater support for business creation is forthcoming.

A third '**Promising**' group includes Ireland and Denmark ranked near average on both scores, just behind Sweden which scored highest on Opportunity factors, and Finland which is above average on HLT. While all these countries stand more or less at the EU median, with comparable performances in both 'first generation' HLT R&D and transferring results to the marketplace, they need to boost both their HLT research investment and also improve their technology transfer record if they wish to aim for next generation standards.

Finally, there is a group of four countries (Greece, Italy, Portugal and Austria) that have reached the '**Structural Limits**' of their existing HLT market situation, and require a new approach to catch up with the leaders. They all scored below average on both measures, though with different profiles. Both
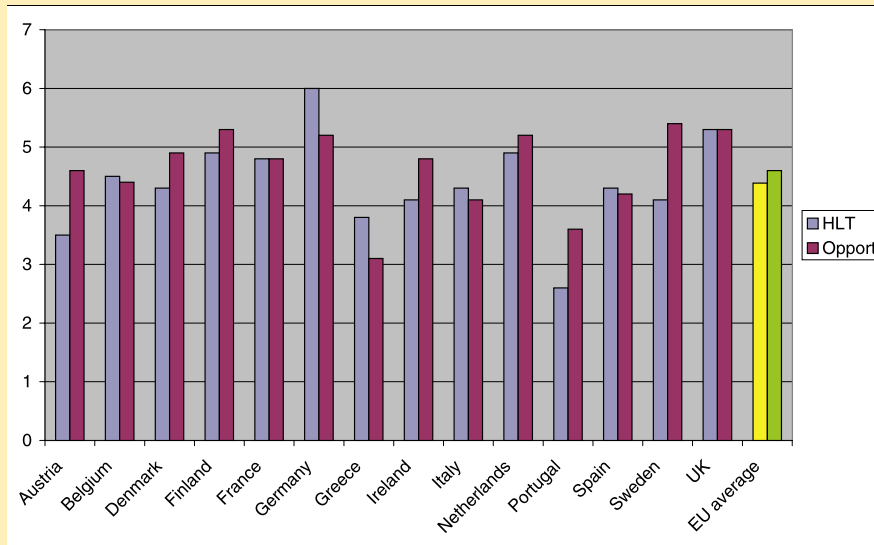
Figure 2: Comparison HLT/Opportunity

Greece and Portugal scored low on Opportunity factors, though Greece scored higher on HLT measures, due to its strong R&D base. Both these countries may need to look beyond their borders for opportunities to exploit their HLT research, and will benefit from enhanced EU collaboration. Portugal, in particular, could improve its research opportunities with more cross-border collaboration. Italy has a stronger research base than most, but with Austria is pulled down by low scores on Opportunity measures. Austria has the advantage of sharing a language with the leading HLT research country, but this very fact might also act as a disincentive when it comes to expanding its own HLT activities.
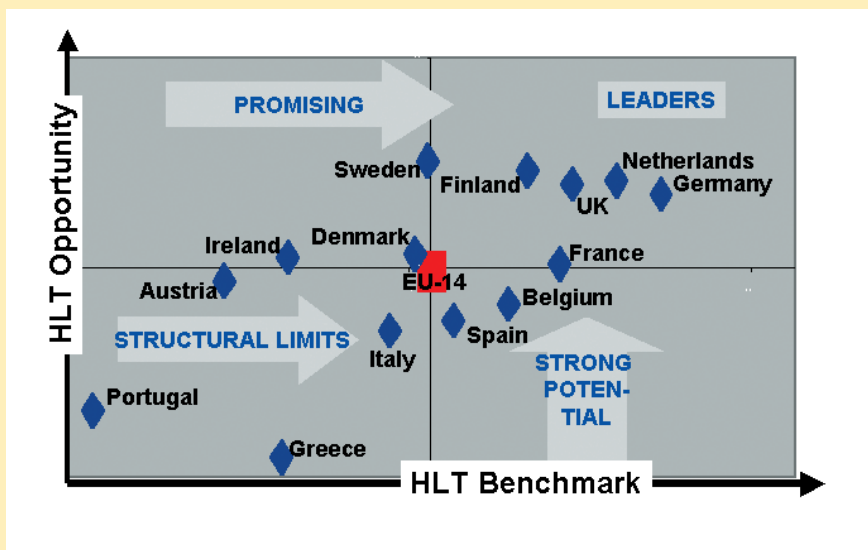
Figure 3: The HLT Scorecard

Austria

HLT Indexes: Austria

Benchmark Scale: 1 (weakest) to 7 (strongest)
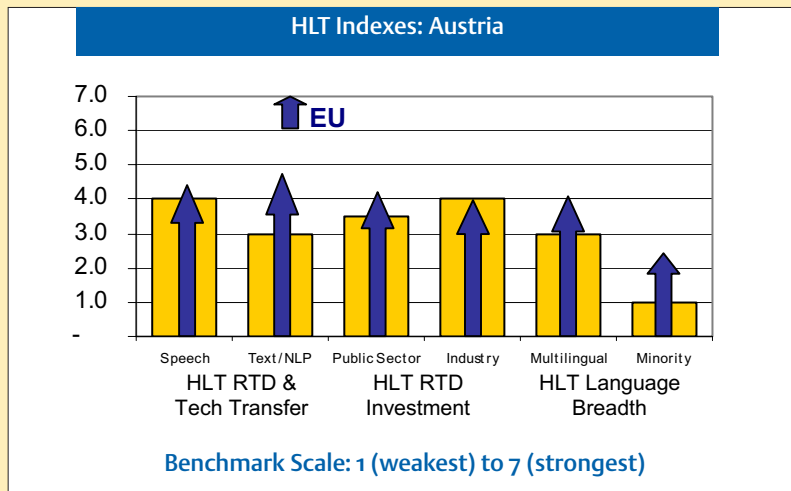
Austria has a strong tradition of RTD in the speech domain, which was supported for some years by the presence of industrial labs hosted by international companies. Both research and industrial activity have been much weaker in the text/NLP domain, and the closing of some speech research labs has placed Austria in a relatively weak position, compared to other Member States. Austrian HLT is inevitably overshadowed by the strong position of Germany, and the Austrian research community would benefit from a stronger focus on specialised applications; one industrial lab is already moving in this direction with integrated speech/NLP solutions.

In addition, Austria could exploit its geographical position and linguistic roots with a focus on Central European languages. More than 90% of Austrians are native German speakers, while nearly 6% speak regional languages (including Alemanisch, Bavarian, Hungarian, Romani, Serbian/Croatian and Slovenian). Less than 3% of Austrians are native speakers of immigrant languages. All HLT research in Austria is in the German language. More cross-border collaboration addressing less dense European languages could give more scope to the Austrian HLT research base.

HLT Benchmark: 3.6

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Austria scores below average on research and technology transfer measures, particularly in text/NLP applications, where there is little activity. There is a strong tradition of speech research; Philips ran a Speech Processing Competence Center in Austria for some years, but it is no longer active since Philips has reduced its activities in the speech research domain. Consequently, Austria scores average on the industry-investment measure.

The strong HLT performance of Germany poses a challenge for Austria, where limitations of size and resources necessarily limit the ability to compete in German-language research. (This is comparable to the position in Ireland for English-language research.) Austria has implemented progressive high-tech R&D support programmes, particularly in telecoms where its strength in speech research would be most relevant. But these have not had a significant impact on the HLT scene, and there has been no dedicated support for language technology research. Nor has Austria successfully incubated a local industrial base that could leverage specialised HLT

expertise for competitive advantage within the wider European market (as, for instance, has happened in Ireland). Scores on multi-language research, and work on minority languages, are also low for Austria.

Technology Transfer

The Austrian government's Technologie Impulse Gesellschaft (TIG) funds the A-plus-B programme to encourage academic spin-offs from Austrian academic institutions by providing professional support for scientists in the difficult process of turning a good business idea into a viable business. Start-up activity in Austria is generally not very dynamic by international standards; this is particularly true for the high-tech sector, which accounts for less than 10% of all new companies. The A-plus-B programme therefore aims to bring about a sustainable increase in the number of innovative, technology-oriented spin-offs from the academic sector. This involves not only counselling and assistance during the actual start-up phase but also establishing the idea of entrepreneurship more firmly in academic theory and practice. Close links between potential founders and their academic "home base" ensure that the new companies can exploit the know-how developed in academic institutions.

There is little evidence that the A-plus-B programme has yet made an impact in the HLT field. Austria still has a very small number of suppliers of HLT products; only four were identified by EUROMAP, one of which grew out of the research arm of L&H (in Belgium and Germany). Austria may have relied too heavily on the resources of multi-nationals, at the expense of the local scientific community.

HLT Policy

There is no dedicated HLT research programme in Austria. Funding for HLT research is available through the Technology Impulse Programme (TIG), whose K-Plus Competence Centers stimulate long-term co-operation between innovative enterprises and top-quality researchers. Pre-competitive research and development on an internationally competitive level

will be supported. K-Plus Competence Centers develop R&D competence as well as human capital in promising fields of research.

One of the key features for the establishment and operation of a Competence Center is the long-term participation of research institutions and at least five enterprises. Three centres have been developed for topics relevant to HLT, but thus far there is not a competence centre specifically dedicated to language technology, possibly because the base of companies that can channel HLT to market has been too small, or too volatile. The Telecoms Research Center (FTW) funds projects selected on grounds of their industrial relevance that will the transfer of scientific results to innovative applications of telecommunication technologies through the collaboration of researchers employed by FTW and specialists from the partner companies. This centre funds some speech research. The Knowledge Management, Software Competence centres have projects relevant to text/NLP.

The Austrian Industrial Research Promotion Fund (FFF) is Austria's most important source of finance for research and development projects carried out by industry. FFF also supports scientists working on new products together with companies. It helps companies by providing them with an objective evaluation of each project chances of success, co-operates with know-how transfer agencies and helps in the search for joint research ventures. The FFF manages the ITF (Innovation and Technology Fund) for projects involving a large element of research and development. Projects involving technology transfer and technology diffusion are mainly supported through this fund. However, ICT applications are not priority within the FFF. Moreover, the industrial base that might take advantage of this fund for HLT research is very small in Austria, and may even be shrinking.

HLT Scorecard: 4.2

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business
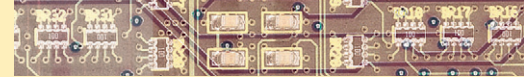
environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Austria scores above average on several significant "opportunity" factors - including competitiveness, strength of ICT infrastructure, and innovation potential. However, scores on other measures are below average. Access to channels for HLT exploitation is a particular challenge in Austria.

The combined weakness in HLT research and opportunity scores poses a significant challenge for HLT research in Austria. Remediation measures could include: a more focused HLT research programme, more specialisation for advanced applications to exploit existing core HLT components, developing niche industry positions, and expanding the research focus to new geographical/linguistic areas, such as Eastern Europe.

HLT Suppliers

EUROMAP identified the following HLT suppliers in Austria: Roodix, Sail Labs, Sonorys, and Webdynamite.

HLT Labs

Austria has around seven research labs working in the HLT field, including: Austrian AI Research Institute (ÖFAI), Vienna Telecommunications Research Centre (FTW), Institute of Translation and Interpretation (U. of Innsbruck), Department of Linguistics and Computer Linguistics Research (U. of Klagenfurt).

HLT Initiatives

No dedicated programme; some support within K-Plus Competence Centers.
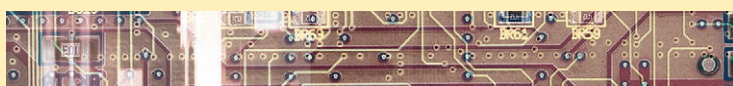
Opportunity Snapshot - Austria

Economy and Society - Austria	
△ Total Population	8,200,000
△ Languages (number of native speakers)	
German	7,500,000
Regional languages (circa 7)	475,000
Immigrant languages (circa 12)	225,000
△ % of citizens who can speak a language in addition to mother tongue	61%
△ Number of Internet users in Austria	3,700,000
△ Gross Domestic Product	
Total GDP (€ millions)	€ 188,500 M
GDP per capita	€ 23,000
RTD and Innovation - Austria	
△ Annual RTD Expenditure (€ millions)	€ 3,000 M
Total RTD as % of GDP	1.5%
Public RTD Expenditure (€ millions)	€ 1,330 M
Business RTD Expenditure (€ millions)	€ 1,721 M
△ High-Tech Patents per 1M population	
European Patent Office	9.8
US Patent Office	5.6
△ Language Technology R&D	
Number of HLT Research Centres	7
Number of Active HLT Suppliers	4
ICT Infrastructure - Austria	
△ Number per 100 inhabitants:	
PCs	27.7
Internet Users	32.9
Mobile telephone subscriptions	78.5
Telephone lines	47.4
△ Computers with an Internet connection	21.3%
△ ICT Spending	
ICT Expenditure as % of GDP	5.9%
Per capita ICT expenditure	€ 1,482

HLT Indexes: Belgium

Benchmark Scale: 1 (weakest) to 7 (strongest)
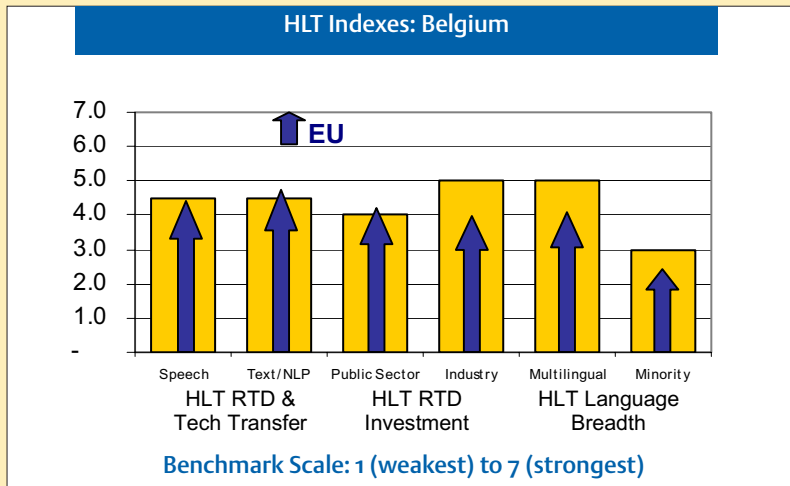
HLT has a long and established history in Belgium, where some of the earliest suppliers of language techno-logy products and services were established in the 1990s. The research base is robust, and compares well with the EU standard. Commitment to research by both the public and private sector is favourable.

Flanders benefits from close collaboration with the research community in The Netherlands, while the Observatoire du Traitement Informatique des Langues & de l'Inforoute Espace Wallonie-Bruxelles (OTIL) provides a focus for French-language HLT in Belgium.

A multi-language focus is natural to the Belgian community, which has three national languages, and is the home to citizens from all over the Union due to the administrative role of Brussels. The vast majority of Belgians have as their mother tongue either Dutch (57%) or French (37%), while only 1% has mother tongue German, the third national language. In addition, 4% have other EU languages as mother tongue, and an-other 2% speak immigrant languages. Belgium has one of the best track records in addressing cross-lan-guage issues and multi-language HLT development, in both the speech and NLP domains.

HLT Benchmark: 4.4
The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Belgium ranks above average on the HLT Benchmark, reflecting a relatively strong base of research and technology transfer. HLT research is carried out in 20 departments of eight universities in Belgium, and there are in addition five institutes or research networks that support HLT.

The HLT research base is split between Flemish and Wallonian centres. Both public and industry invest-ment in HLT research is strongest in Flanders; work on the Dutch language benefits from the close collabo-ration between Flemish researchers and pro-grammes in The Netherlands.

Belgium has a strong history of industry investment in HLT, and ranks above average on this measure, as well as on its tradition of RTD in multiple languages. Commercial investment in HLT research has been interrupted recently by the failure of the most signi-ficant national player, L&H, but the assets of this company have passed to others, and most of the technology is still available to the market, some in successor companies in Belgium.

Technology Transfer
Belgium has some of the most established compa-nies in the HLT sector, including a few that have been in operation since the early 1990s, though players in the sector continue to be small. The establishment of the Flanders Language Valley (FLV) in Ypres in 1997

was a watershed event, not just in Europe but worldwide, as the first science park and incubator established specifically for HLT. Although much of the dynamism of FLV has dissipated with the failure of L&H, who were strong backers, the FLV Fund continues to back start-up companies with a potential to exploit HLT.

EUROMAP has identified 20 companies active in the Belgian HLT market. Speech-based applications are particularly strong in Belgium, with eight companies active in this segment; another three companies are developing interface applications based on NLP, though there is as yet little interaction of speech and NLP for interface products.

Belgium is also prominent in the promotion of cross-language applications, with five companies working actively in this area. In addition, four companies produce knowledge applications based on NLP technology. None of these companies overlap with the cross-lingual suppliers, suggesting that advanced cross-lingual knowledge applications is a ripe opportunity for Belgian suppliers.

HLT Policy

The Dutch Language Union (NTU, Nederlandse Taalunie) is an intergovernmental organisation that unifies The Netherlands and Flanders in the field of language and literature and has become the pivotal institution for the promotion of HLT in the Dutch language. The NTU is responsible for implementing the language policies stipulated by the Flemish and Dutch governments and is involved in a number of HLT projects.

The NTU initiated the Platform for Dutch HLT that has established priorities for basic Dutch-language HLT components. It has set criteria for creating core components as well as a blueprint for managing, maintaining, making available and distributing the basic Dutch-language resources that can be used in education and research and for developing HLT tools and applications.

The Platform is directly supported by three Flemish

agencies, including the Science and Innovation Administration, which also manages and finances the Flemish part of the Spoken Dutch Corpus, and co-ordinated the Flemish Research Programme on Language and Speech Technology for Dutch (1993 - 1997).

The Institute for the Promotion of Innovation by Science & Technology in Flanders also supports technology R&D for Flemish companies. Specific action programmes relevant to HLT are Medialab (that funds non-technological aspects of electronic services) and the Information Technology Action Programme that promotes transfer of RTD results in the field of information technology to SMEs.

The National Fund for Scientific Research in Flanders has established research networks on computational linguistics and language technology, including the CLIF research community (Computational Linguistics in Flanders), and networks focused on advanced topics such as cognitive linguistics, contrastive linguistics and language typology.

HLT Scorecard: 4.4

The HLT Scorecard compares the HLT Benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Belgium ranks in the "strong potential" group. It lags behind others in the EU on measures such as new business formation, supply-side readiness, and innovation potential factors.

However, although it is not in the Leaders group, with an Opportunity score similar to the HLT Benchmark, Belgium is judged to be in a good position both to extend its research base in innovative ways, and to exploit the results of RTD in successful technology transfer.

HLT Suppliers

Belgium has around 20 suppliers of HLT products and services - for example: BaBel Technologies, I.R.I.S Group, Language & Computing, Language

Dynamics, Scansoft, UbiCall Communications, Xplanation.

HLT Labs
Belgium has more than 15 research labs working in the HLT field, including: Vrije U. Brussel (AI Lab), Katholieke U. Leuven (PSI, ICRI, Machine Learning Group), U. of Mons-Hainaut.

HLT Initiatives
Nederlandse Taalunie (Dutch Language Union), Dutch HLT Platform, MediaLab, CLIF, OTIL, Spoken Dutch Corpus.

Opportunity Snapshot - Belgium

Economy and Society - Belgium	
△ Total Population	10,300,000
△ Languages (number of native speakers)	
Dutch	6,000,000
French	4,000,000
Dialects and non-indigenous languages	300,000
△ % of citizens who can speak a language in addition to mother tongue	61%
△ Internet users by language	
Dutch-speaking Internet users in Belgium	2,600,000
French-speaking Internet users in Beligum	900,000
△ Total GDP (€ millions)	€ 230,000 M
GDP per capita	€ 23,000
RTD and Innovation - Belgium	
△ Annual RTD Expenditure (€ millions)	€ 4,420 M
Total RTD as % of GDP	1.8%
Public RTD Expenditure (€ millions)	€ 1,242 M
Business RTD Expenditure (€millions)	€ 3,179 M
△ High-Tech Patents per 1M population	
European Patent Office	17.6
US Patent Office	12.8
△ Language Technology R&D	
Number of HLT Research Centres	18
Number of Active HLT Suppliers	20
ICT Infrastructure - Belgium	
△ Number per 100 inhabitants:	
PCs	34.5
Internet Users	26.2
Mobile telephone subscriptions	54.9
Telephone lines	49.9
△ Computers with an Internet connection	8.6%
△ ICT Spending	
ICT Expenditure as % of GDP	5.75%
Per capita ICT expenditure	€ 1,381

Germany

HLT Indexes: Germany

Benchmark Scale: 1 (weakest) to 7 (strongest)
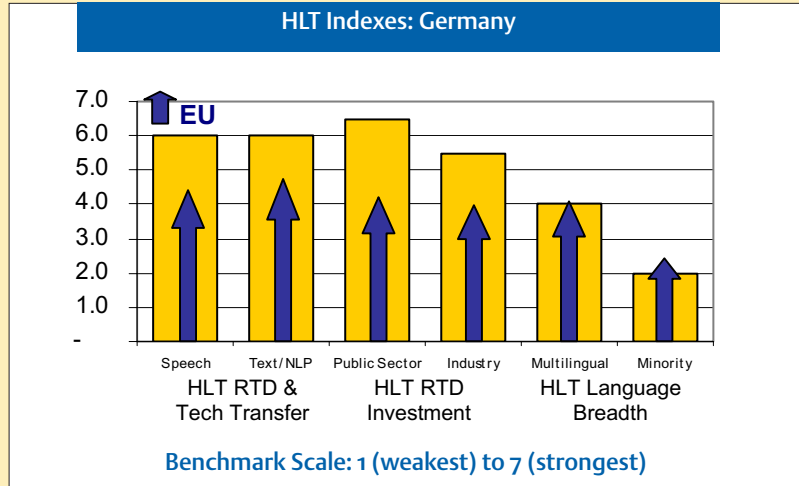
Germany has the strongest HLT R&D position in the EU, with the advantages of an exceptionally effective research environment, well developed industrial programmes, and a mature ICT infrastructure to exploit language technology products and services. Germany has a large number of academic spin-offs in the HLT field. A significant number of the 60+ commercial companies developing HLT-based products use technology developed in research programmes supported by national programmes, or in collaborative EU R&D.

Standard German is the mother tongue of 90% of the population in Germany, while three percent speak regional languages (either German variants or the languages of neighbours, such as Danish or Polish). More than six percent of the population speak "immigrant" languages, with the largest group speaking Turkish. The traditional HLT research agenda has been strongly focused on the German language, and the high quality and availability of German language technology is largely due to the extensive research programmes supported at the national level. Since the mid-1990s German research has acquired a more multilingual focus, with significant attention to high-density languages, notably English and Japanese. Development of tools for low-density languages is less advanced. Multilinguality is weakest in the speech technology domain.

HLT Benchmark: 5.1
The HLT Benchmark measures the relative maturity of language technology R&D in the EU. Germany ranks well above average in both speech and text/NLP research, and has a strong - and continuing - track record of investment from both the public and private sectors. On the other hand, Germany ranks just below average on multilingual breadth in its research base, and is below average in RTD devoted to minority, immigrant, or low-density languages.

Good industrial participation in RTD, made possible by the structure of German research programmes, has paid off in the HLT domain. The strong research record has, in many cases, led to the creation of new

products and services, and new companies exploiting language technology.

Language technology components for the German language are well developed, and Germany is among the leaders in the design of advanced applications for HLT. Innovative companies have been launched to exploit HLT in a variety of priority areas, especially for knowledge management applications. Speech interface products are relatively advanced, spurred on by the availability of suitable platforms in the automotive industry, which has been a strong supporter of the HLT domain. In addition, German companies are leaders in the integration of speech and text HLT to create next-generation interfaces for

electronic products and services. The German market continues to incubate leading developers of translation technology, which has a long tradition in Germany.

Technology Transfer

Germany has had by far the most compelling success rate in commercialisation of HLT, with more than 60 companies active in the market. The range of applications based on speech and text is reasonably well balanced, though there is a slightly larger base of companies developing speech products. The largest single application area is in speech interfaces, where 32 companies have products; this reflects the long-standing and highly effective speech R&D in Germany, particularly by significant industrial companies in, for example, the automotive sector. Germany is a world leader in informatics applications for in-car services. On the other hand, the linguistic coverage of speech products in Germany is relatively limited, and only two or three products have a broad coverage of many languages; most are targeted specifically to German. This obviously limits the market potential of such products within the EU and globally.

There are 13 German companies developing cross-language applications - more than any other Member State. Most of these are translation tools, and several products are being developed with advanced engineering approaches (such as example-based or statistical methods). Three companies are actively engaged in developing speech-based cross-language applications, reflecting the legacy of the Verbmobil initiative in the 1990s. These cross-language products cover a broad range of languages, and more than half address at least five languages. A similar number of companies (14) are producing knowledge management applications. However, only three of these products have wide linguistic coverage, namely those that are being developed by companies also involved in translation tools.

HLT Policy

Germany's success in the HLT field is built on its well established approach to research funding and exploitation: promoting a healthy balance between national and regional programmes; co-ordinating public and private-sector investment; emphasising the development of a strong pool of research scientists; and supporting international collaboration. Nowhere was this approach better exemplified than in Verbmobil, a long-term project of the Federal Ministry of Education & Research (BMBF), co-ordinated by the German Aerospace Center (DLR, which acts as a Project Management agency for BMBF) over the second half of the 1990s. The aim of Verbmobil was to give Germany a top international position in language technology by co-operation and concentration of as many specialists as possible from industry and science. The focus of Verbmobil research was the development of a mobile translation system for the translation of spontaneous speech in face-to-face situations. Germany's strong position in the HLT field is due, in no small part, to this ambitious programme.

Currently, Germany funds both generic ICT programmes which can incorporate HLT research as well as specific lang-tech initiatives. The most recent lang-tech project is COLLATE, which will develop a Competence Center for Language Technology within DFKI (the German Research Center for AI) at Saarbrücken. The aim is to shorten the path to market, focusing on information search, extraction, and summarisation, and on applications for natural interactivity in electronic services. The Competence Center comprises a virtual information center (Language Technology World), a Demonstration Center for LT systems, and an Evaluation Center for LT applications.

The IT-Research 2006 programme sets the stage for reorientation of research funding in the area of information and communication technology. The Informatiksysteme programme, supported by BMBF, aims to strengthen the scientific/technical basis of German computer science research as well as to accelerate the transfer of new technologies from the research into the economy. It will assist in making fundamental contributions to the development of

knowledge-intensive industries and services and thus contribute on a long-term basis to the creation of new High Tech jobs in Germany.

HLT Scorecard: 5.2

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Germany scores above average on most measures for HLT opportunity. While measures of "supply-side readiness" (e.g. availability of capital for new business formation) are only average, this is compensated for by business start-up programmes available for ICT innovators. Germany will continue to be a leading player in the EU HLT domain.

HLT Suppliers

Germany has more than 60 suppliers of HLT products and services - for example: Acrolinx, Aculab, Aixplain, Cortologic, Kiwilogic, Langenscheid, Mundwerk, Ontoprise, Sail Labs, Semantic Edge, Sonorys, TalkinWeb, Temis, Trados, Voice Robots, Zeres, etc.

HLT Labs

Germany has more than 80 research labs working in the HLT field. These include commercial labs such as Alcatel, IBM, Philips, Sony, Siemens, Grundig and Robert Bosch, as well as applied research institutes such as the Fraunhofer. Twenty German universities carry out HLT research in various departments.

HLT Initiatives

Various programmes funded by the German Ministry of Education & Research (BMBF), including COLLATE.

Opportunity Snapshot - Germany

Economy and Society - Germany	
△ Total Population	83,000,000
△ Languages (number of native speakers)	
German	75,300,000
Regional languages (circa 23)	2,300,000
Turkish	2,100,000
Other immigrant languages (circa 45)	3,300,000
△ % of citizens who can speak a language in addition to mother tongue	53 %
△ German-speaking Internet users in Germany	37,100,000
△ Total GDP (€ millions)	€ 1,850,000 M
GDP per capita	€ 22,250
RTD and Innovation - Germany	
△ Annual RTD Expenditure (€ millions)	€ 48,200 M
Total RTD as % of GDP	2.38%
Public RTD Expenditure (€ millions)	€ 15,200 M
Business RTD Expenditure (€ millions)	€ 33,000 M
△ High-Tech Patents per 1M population	
European Patent Office	29.3
US Patent Office	14.4
△ Language Technology R&D	
Number of HLT Research Centres	85
Number of Active HLT Suppliers	63
ICT Infrastructure - Germany	
△ Number per 100 inhabitants:	
PCs	33.6
Internet Users	31.3
Mobile telephone subscriptions	58.6
Telephone lines	60.1
△ Computers with an Internet connection	7.4 %
△ ICT Spending	
ICT Expenditure as % of GDP	5.71 %
Per capita ICT expenditure	€ 1,400

Denmark

Benchmark Scale: 1 (weakest) to 7 (strongest)
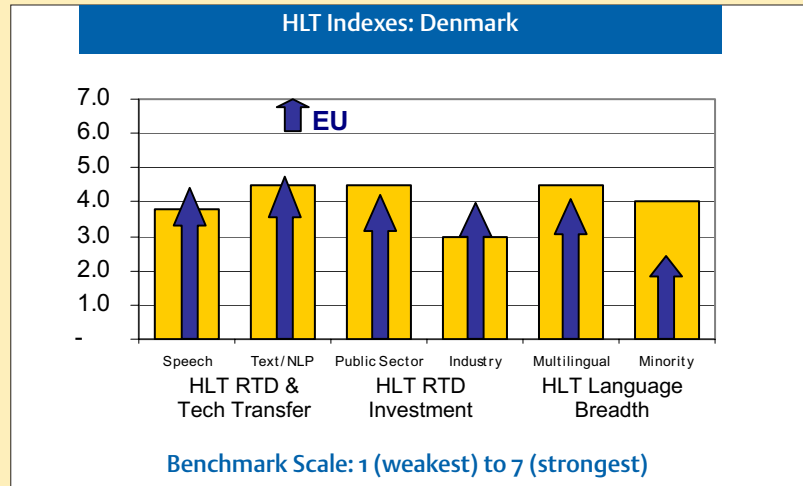
Denmark enjoys a well-trained R&D base in language and speech technology, and in addition to EU project participation, plays a leading role in regional Nordic language technology activities. The country offers a healthy business innovation environment and excellent levels of public infrastructure readiness, but due to the small national marketplace, language technology transfer has reached no more than the European norm.

As a traditionally export-focused country, Denmark has developed strong multilingual capabilities which have made it an excellent cross-border facilitator. It now faces the challenge of ensuring that its own language community can benefit from speech and language technologies appropriate to its high degree of readiness. While 95% of the population are Danish speaking, the language is highly localised, and not used widely elsewhere. A continued focus on cross-border collaboration - both within the Nordic Region, and in the wider European context - will ensure the future vitality of the Danish HLT community.

HLT Benchmark: 4.6

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Denmark scores to the EU average on measures of robustness in HLT research. It enjoys a strong tradition of text and NLP applications, and speech research is well represented. The country has five major HLT research centres, one of which acts as a national centre of excellence for language technology. Industry involvement in HLT research, however, is scarce. There is also an active professional translation community, including a dedicated machine translation system for processing patent documents from English to Danish.

As yet, however, there is a relatively limited number of products and services available on the market based on these basic tools. Denmark has about twelve HLT suppliers, though not all of them are dedicated language-technology focused companies.

Technology Transfer

Although Denmark enjoys a relatively strong innovation potential, a high degree of trade competitiveness and a well-developed digital infrastructure capable of integrating language technology applications, the transfer of first generation language technology to the marketplace is still in waiting mode.

In part this is due to the relative lack of large-scale high technology channels that can facilitate the transfer of language and speech technologies to market, and partly also to the small scale of the local

language marketplace. Denmark's IT sector, for example, is far more consultancy- than manufacturing-oriented. The fact that Danish citizens receive good foreign language training in schools could also act as a break on demand for local language technology products and services.

HLT Policy

Although Denmark has not benefited from a fully-fledged HLT programme with substantial funding, the government did support the setting up of the Center for Sprogteknologi, the country's main language technology R&D facility, which has played a major role in national and regional development in this field.

In terms of basic resource development, the Ministry of Science, Technology and Innovation funds some HLT research. It has recently funded a project to develop a lexical database for language technology. It also funded a speech technology initiative in the late 1990s that led to a successful development of a text to speech system for Danish.

In addition the Danish Research Agency funds an interdisciplinary research programme focussing on information technology. The Danish Research Council for the Humanities (SHF, housed by the Research Agency) has financially supported various HLT-oriented projects within the area of speech analysis and NLP.

Denmark also takes its role as a member of the Nordic region seriously and plays an active role in the NorDokNet (documentation centre for language technology research for Danish, Faroese and Greenlandic) and more generally in the Nordic Academy for Advanced Study (NorFA) that supports Nordic language technology research. NorFA has launched the Nordic Research Programme 2000-2004 that supports initiatives to secure the use of Nordic languages through development of the Nordic language technology environment.

HLT Scorecard: 4.6

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Denmark scores above the average on most "readiness" measures, indicating a high potential for exploitation of HLT research. However, it has so far been less successful in transferring its research results to market. It could therefore benefit from a greater effort at transfer opportunities. Given the country's high degree of Internet penetration and networked educational system, such contexts as e-government, education and training may provide new opportunities for exploiting language technology in an inherently small market.

HLT Suppliers

Denmark has around twelve suppliers of HLT products and services including: Ankiro, Nordisk Språkteknologi, Mondosoft, Prolog Development Center, Max Manus, MikroVærkstedet A/S, Center for Sprogteknologi, Navigo, KBL Sprogmagisteren, Speech-Ware, Textware, Empathy Systems.

HLT Labs

Denmark has a small number of research labs working in the HLT field, including: Center for PersonKommunikation, AUC, Natural Interactive Systems Laboratory, Center for Sprogteknologi, Institut for Datalingvistik, Copenhagen Business School, Institut for Sprog og Kommunikation SDU, Institut for Fagsprog, Kommunikation og Informationsvidenskab, SDU.

HLT Initiatives

Nordic Research Programme 2000-2004; SprogTeknologisk Ordbase - STO (Lexical Database for Language Technology).

Opportunity Snapshot - Denmark

Economy and Society - Denmark

△ Total Population	5,300,000
△ Languages (number of native speakers)	
Danish	5,000,000
Regional and immigrant	300,000
△ % of citizens who can speak a language in addition to mother tongue	85 %
△ Danish-speaking Internet users in Denmark	3,400,000
△ Total GDP (€ millions)	€ 162,000 M
GDP per capita	€ 30,000

RTD and Innovation - Denmark

△ Annual RTD Expenditure (€ millions)	€ 3,500 M
Total RTD as % of GDP	2 %
Public RTD Expenditure (€ millions)	€ 1,253 M
Business RTD Expenditure (€ millions)	€ 2,224 M
△ High-Tech Patents per 1M population	
European Patent Office	21.5
US Patent Office	17.3
△ Language Technology R&D	
Number of HLT Research Centres	6
Number of Active HLT Suppliers	12

ICT Infrastructure - Denmark

△ Number per 100 inhabitants:	
PCs	43.2
Internet Users	29.6
Mobile telephone subscriptions	61.0
Telephone lines	75.3
△ Computers with an Internet connection	14.5 %
△ ICT Spending	
ICT Expenditure as % of GDP	6.2 %
Per capita ICT expenditure	€ 2,000

Spain

HLT Indexes: Spain

Benchmark Scale: 1 (weakest) to 7 (strongest)
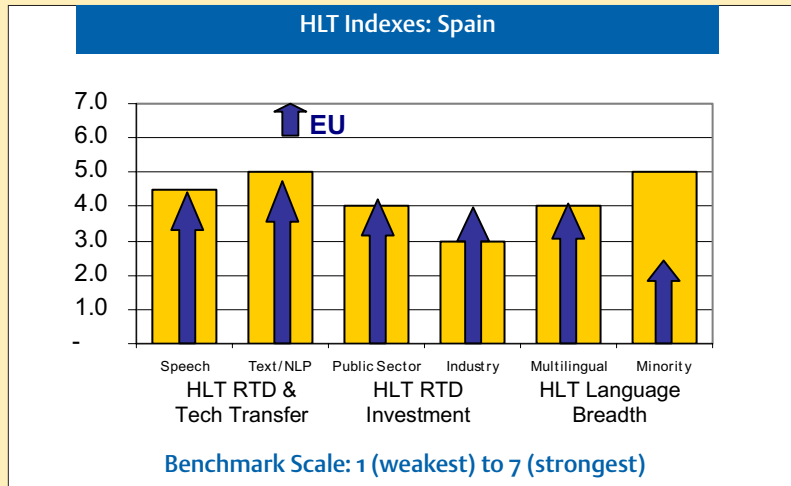
Spain has made exemplary progress in the development of HLT competence to meet the needs of its linguistically complex population. With nearly 30% of Spain's citizens native speakers of a regional language other than Castilian (i.e. Spanish, the common language nationally), the research focus is necessarily diffuse. Catalan, Basque, Galician and Gascon/Aranese are officially bilingual with Spanish in the regions where they are spoken, while both Aragonese and Asturian are protected within their Autonomous Communities (though they are not recognised as "official" languages).

Political devolution has enabled the Spanish regions to control the agenda for language technology development, with particular success in Catalan and Basque. This regionalisation, however, has the consequence of diluting the focus on Spanish; as a global language of considerable commercial significance this may represent a missed opportunity for the Spanish HLT community. The intrinsically broad linguistic focus of Spanish research provides a good model for Europe's HLT community.

HLT Benchmark: 4.1

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Spain scores well in research and technology transfer, around average for speech and above average for text/NLP applications. Several small NLP-based companies (including a market leader in translation memory), as well as translation service suppliers that develop and use machine translation (including the market leader in localisation services) have a presence in Spain.

Public-sector investment in the domain is a little below average; there is no national HLT programme, although support is quite strong in some of the regions. Investment by industry is weaker; although there is a relatively large number of smaller companies working in niche areas, HLT plays

a role in certain commercial labs working in the field.

The diffuseness of the Spanish effort in HLT is double-edged. On the one hand there is a depth of scientific experience with less dense languages that is virtually unique in Europe. Attention to developing cross-lingual products and services between regional languages is equally rare in other Member States. On the other hand, like Ireland (for English), the research community in Spain competes globally in the development of products for a world language. Spanish is well served by established first-generation tools (such as machine translation) that were developed in the US, or elsewhere in Europe, to take advantage of the market opportunity represented by a global Spanish-speaking population of four to five hundred million people.

Technology Transfer

Spain has made a good start in technology transfer in the HLT field, with around 15 suppliers identified in the market. It has a small but significant base of companies developing and selling products with cross-lingual applications, most of them focussing on Spanish, but also there is development of cross-lingual knowledge applications in Catalan, Basque and Galician, and some work on products in other languages, for example English, Portuguese, Italian, French and German. A free online MT system for Catalan is available. There are a few suppliers of speech technology, including products that process Spain's regional languages, and several that are multilingual; most of these are interface applications.

The Ministry of Science & Technology has created a technology transfer network (OTRI/OTT – Office for Transfer of Research Results).

HLT Policy

Spain is divided into 15 Autonomous Communities (ACs) and 2 Autonomous Cities (Ceuta and Melilla, in North Africa). All ACs have their own parliaments and governments. The National Government has progressively transferred political competencies to the ACs, keeping the role of establishing a basic regulatory framework that every AC must meet as a minimum. Universities and R&D are two of the competencies that have been fully transferred to the ACs, and the Science Act co-ordinates the Science Policy of national government ministries and the regions.

HLT is included in the Information Society Programme of the National R&D Plan, but not in a specific programme. The Office for Spanish in the Information Society (OESI) is part of the Instituto Cervantes (hosted by the Ministry of Foreign Affairs). OESI co-ordinates information and activities related to language technology in Spain.

ACs have their own R&D programmes and HLT is specifically included in the plans of Andalucía, Asturias, Cataluña, Islas Baleares, La Rioja and País Vasco. Some of the regional programmes have been in place for many years and have developed advanced research programmes, e.g. in Cataluña which has a Catalan Language and Language Technologies Programme. Five other ACs have ICT programmes that can accommodate HLT research. There are specific IST programmes for Basque, Catalan and Galician. Communities where these languages are spoken have implemented HLT policies to support their languages; in addition there are research centres and/or networks for Basque, Catalan and Galician HLT.

At the national level, the Centre for Industrial Technological Development (CDTI) and the Programme for Promoting Technical Research provide support for industrial R&D, and these programmes have been used by HLT researchers and developers. The Office for Science & Technology (SOST) supports Spanish participation in EU-funded programmes, which has improved in recent years.

HLT Scorecard: 4.3

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Spain scores slightly below the EU average on the overall opportunity measure. It is particularly strong in some areas, notably in potential access to channels for HLT with the large global base of Spanish speakers and the strong international position of its national telecoms supplier (with a strong presence in Latin America). On other key measures, however, Spain scores below average, e.g. the maturity of its ICT infrastructure and its innovation potential.

Overall, Spain is among the Member States with "strong potential" in HLT. Its commitment to multi-language research, especially through regional RTD programmes, is a model for the Union. As the ICT infrastructure matures - with wider access to the Internet, for example - Spain should be able to exploit its head start in inter-regional HLT capabilities. The HLT community would also benefit from a more focused approach to the development of Spanish-

language applications (perhaps at the national level), in order to compete effectively in global markets. More robust mechanisms to support the transfer and commercialisation of tools suitable for the integrated European market would benefit the HLT community as a whole in Spain.

HLT Initiatives

There are many national and regional initiatives to boost research into HLT in Spain.

Opportunity Snapshot - Spain

Economy and Society - Spain		
△	Total Population	40,000,000
△	Languages (number of native speakers)	
	Castilian Spanish (excl. native speakers of regional languages)	28,600,000
	Catalan	6,500,000
	Galician	3,000,000
	Other regional languages including Basque (circa 8)	1,600,000
	Immigrant languages (circa 8)	300,000
△	% of citizens who can speak a language in addition to Castilian and their regional language	32%
△	Internet users by language	
	Number of Internet users in Spain	9,400,000
△	Total GDP (€ millions)	€ 582,000 M
	GDP per capita	€ 14,545
RTD and Innovation - Spain		
△	Annual RTD Expenditure (€ millions)	€ 5,500 M
	Total RTD as % of GDP	0.9%
	Public RTD Expenditure (€ millions)	€ 2,600 M
	Business RTD Expenditure (€ millions)	€ 2,800 M
△	High-Tech Patents per 1M population	
	European Patent Office	2.5
	US Patent Office	1
△	Language Technology R&D	
	Number of HLT Research Centres	30
	Number of Active HLT Suppliers	15
ICT Infrastructure - Spain		
△	Number per 100 inhabitants:	
	PCs	14.3
	Internet Users	17.5
	Mobile telephone subscriptions	60.9
	Telephone lines	42.1
△	Computers with an Internet connection	7.9%
△	ICT Spending	
	ICT Expenditure as % of GDP	6.8%
	Per capita ICT expenditure	€ 973

Finland

HLT Indexes: Finland

Benchmark Scale: 1 (weakest) to 7 (strongest)
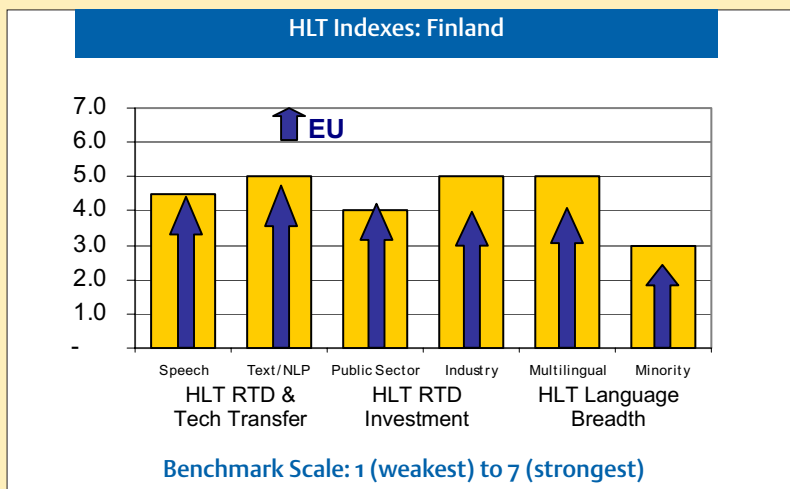
Finland has developed a strong research base in language technology, and is beginning to make an impact outside its home market. The country enjoys a healthy environment for technology development and transfer, both in its business environment and in the advanced state of its ICT markets and infrastructure. A new generation of companies is transforming the way language technology moves from Finnish labs to the market - seeking investment support, and targeting cross-border markets, as early as possible.

Finland has a particularly strong multi-language research focus, and participates in the Nordic-language collaboration initiatives. The well-known complexities of the Finnish language are credited with boosting the focus on theory in HLT research, giving Finland valuable potential for developing advanced HLT solutions in many languages. The research community benefits from programme support and investment from both the national government and the private sector.

Finland has two official languages: Finnish (spoken as a native language by 92% of the population) and Swedish (6%), and there are small communities of Saami and Roma speakers (1%). Immigrant languages play little role.

HLT Benchmark: 4.8

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Finland has developed in recent years as one of the leading HLT research communities in Europe. In spite of its small size, Finland has made significant contributions to the theory and practice of language engineering, and has innovative developers in both speech and text. In the mid-1990s Finnish companies were implementing speech technology developed elsewhere; now, with a strong research push from both the public and private sectors, Finland hosts world-class speech research efforts. With a longer track record in text/NLP development, Finland scores above average on this measure. After the recent increase in focus on speech technology, Finland now scores equal to the EU average for this measure of technology development and transfer.

Core language technology components have been developed for Finnish, Swedish, and English, as well as a number of other languages. These components provide the basis for speech interface, machine translation and knowledge management products. MT is currently limited to Finnish-English, although English-Finnish is in development. Industry investment in HLT in Finland is significant, and Finland rates above average on this measure. Government-funded support for research is now in place and

producing results, but because these programmes started much later than elsewhere in Europe, Finland scores lower on this measure.

Technology Transfer

Finland has a well developed funding structure for technology transfer assistance. Tekes (the National Technology Agency) provides funding through the commercialisation stage of product development, and also supports new companies with brokerage events and clinics, and links with the MIT Industrial Liaison Program. Sitra (The Finish National Fund for R&D) provides capital investment and pre-seed-funding, and has invested in Finnish HLT start-ups. Finnvera is a state-owned finance company that supplements commercial financing for SMEs, particularly for export and internationalisation. Finpro is a business association that supports export activities. New HLT companies also benefit from well established science parks, as well as from organisations that incubate and support licensing and commercialisation of scientific research results (e.g. Licentia and Oulutech).

Finland has been notably successful in creating small, niche companies to exploit its high-quality academic research results. The presence of Nokia (with its substantial R&D facilities that include significant speech-processing research) acts as a strong stimulant to lang-tech transfer. In addition, Finland benefits from its collaboration within the Nordic language communities in developing lang-tech components for multiple languages, and has extended this principle effectively to address major European language markets beyond the Nordic region.

Finland has niche lang-tech suppliers operating in most of the leading application areas, including early efforts to integrate speech and language in advanced products and services. However, most of the sector is still focused on component tools.

HLT Policy

Finland enjoys significant national-level support for its language-technology RTD effort. Tekes is the main financing organisation for applied and industrial R&D in Finland, providing funding and expert services for R&D projects for companies and universities. Tekes also co-ordinates and finances Finnish participation in international technology initiatives. The USIX (User-Oriented Information Technology) programme funded a number of HLT projects through 2002, and Tekes has launched a new Interactive IT programme that will continue support of research areas relevant to the HLT field. The Academy of Finland is an expert organisation in research funding and science policy and funds research projects and centres of excellence in the field.

The Ministry of Education has financed the building of the Finnish Network for Language Technology Studies (KIT Network), which links university departments specialising in language technology and related areas. The aim of the network is to increase the number of professionals and scientists working in the HLT field. The network includes 29 departments in 10 universities, teaching computational linguistics, computer science, cognitive neuroscience, information sciences, applied linguistics, various technical disciplines etc. This initiative has been very effective in awareness and community building for the HLT research base. The labs that participated in USIX R&D funding also joined the KIT network to develop university education for HLT. The result has been a ten-fold increase in the number of HLT students studying language technology either as a major or minor subject.

The Ministry of Education also funds the Research Institute for the Languages of Finland (KOTUS), and The Language Bank of Finland, which is an electronic archive of language resources. The Finnish Language Technology Documentation Centre (FiLT) operates with funding from the Nordic Language Technology Research Programme (through the Nordic Academy for Advanced Study).

HLT Scorecard: 5.2

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business

environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Finland scores above the EU average on every measure that supports healthy development of the language-technology sector, and is ranked in the "leaders" group.

Developing the pool of research scientists through the KIT network will secure the future for Finland's HLT agenda. In addition, the strong engineering and application focus of speech research will need to be supplemented with a more robust theoretical approach to the discipline in Finnish research. The challenge for Finland is to maintain the momentum of its fledging HLT industry through development and transfer of advanced tools and products, and securing its global channels via the ICT mainstream.

HLT Suppliers

Finland has around 15 suppliers of HLT products and services - for example: Connexor, Gurusoft, Kielikone, Lingsoft, Master's Innovations, Promentor Solutions, PT ControlNet, Republica, Sandstone.fi, TimeHouse.

HLT Labs

Finland has around 25 research labs working in the HLT field, including: IBM Finland, Nokia Research Centre, Helsinki University of Technology, University of Helsinki, University of Tampere, VTT (Technical Research Centre of Finland).

HLT Initiatives

KIT, FiLT, KOTUS, The Language Bank of Finland (plus Interactive IT, USIX for general funding).

Opportunity Snapshot - Finland

Economy and Society - Finland

△ Total Population	5,200,000
△ Languages (number of native speakers)	
Finnish	4,800,000
Swedish	300,000
Regional languages (circa 8)	60,000
Immigrant languages (circa 11)	40,000
△ % of citizens who speak a language in addition to mother tongue	58 %
△ Finnish-speaking Internet users	2,060,000
△ Total GDP (€ millions)	€ 121,000 M
GDP per capita	€ 23,250

RTD and Innovation – Finland

△ Annual RTD Expenditure (€ millions)	€ 4,108 M
Total RTD as % of GDP	3.1%
Public RTD Expenditure (€ millions)	€ 1,290 M
Business RTD Expenditure (€ millions)	€ 2,818 M
△ High-Tech Patents per 1M population	
European Patent Office	80.4
US Patent Office	35.9
△ Language Technology R&D	
Number of Lang-Tech Research Centres	21
Number of Active Lang-Tech Suppliers	14

ICT Infrastructure – Finland

△ Number per 100 inhabitants:	
PCs	39.6
Internet Users	38.5
Mobile telephone subscriptions	72.6
Telephone lines	54.7
△ Computers with an Internet connection	25.8%
△ ICT Expenditure as % of GDP	6.43%
Per capita ICT expenditure	€ 1,534

France

HLT Indexes: France

Benchmark Scale: 1 (weakest) to 7 (strongest)
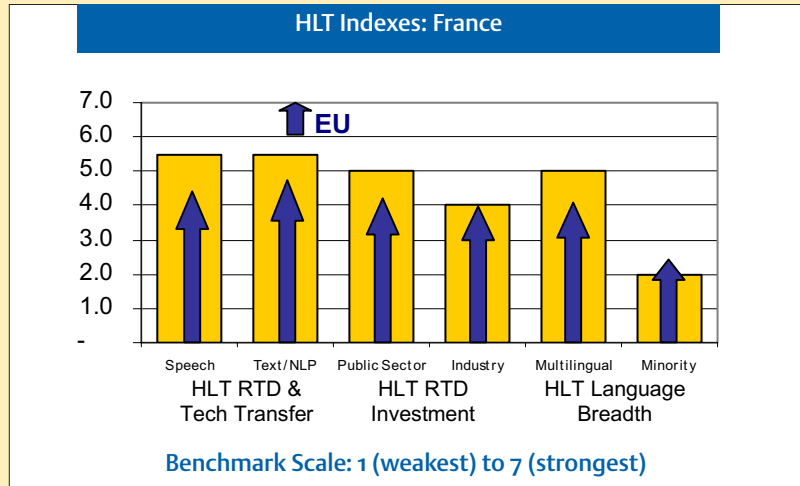
France is a leading player in EU language technology, with a long research tradition, world-class laboratories and coverage of all main domains of activity. It has also nurtured a respectable community of commercial suppliers, some of European and global scale. Research has benefited from consistent public sector support, and France has been a key player in EU collaborative research projects.

Around 87% of the population of France speak French as a native tongue. France also has a large base of regional languages (a total of nearly 30) spoken by around 8% of the population; another 5% are native speakers of immigrant languages, the largest group speaking Arabic. HLT in France benefits from the fact that the French language is a significant, high-density language of considerable commercial importance, and French research also covers other major commercial languages. In addition there is some research on regional and immigrant languages and recently work on Arabic has gained increasing importance, due to historical ties between France and North Africa. There is as yet, however, little operational language technology support for less dense languages.

France participates largely in Francophone language R&D networks, and is committed to preserving and enhancing its language as a vector of knowledge and competitiveness in the concert of nations.

HLT Benchmark: 4.7

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. France scores well above average for all major language technology indictors, testifying to a highly robust R&D environment, and a solid performance in all branches of language technology. It has some 25 research centres, including world-class laboratories in speech technology and in core and advanced text components.

France also has a long tradition of machine translation research, and is home to two of Europe's commercial translation suppliers. Available modules extend beyond the French-English pair, and include such major languages as German and Russian, In addition there is an active national association of language technology players which brings together both academic and industry researchers and developers.

All key language technology components for French, and often for other languages as well, have been developed, and France is also home to the Evaluations and Language Resources Distribution Agency (ELDA) dedicated to assembling and adding value to language resources in all languages.

Work on other offshore languages is widespread in the research base, including Arabic, Russian and Japanese. However, France has a policy of using French for all national affairs, with the effect of reducing active focus on the country's minority languages (Breton, Basque, Occitan, Alsatian, etc). Immigrant languages, although the subject of research, do not yet appear to benefit from specific attention in the marketplace.

Technology Transfer

France's track record in creating language technology companies goes back at least twenty years. In the speech field there have been some notable success in the transfer of public research to the private sector, and in the last decade France has seen numerous new companies either spun off from large public and private industrial concerns or from scratch with venture capital.

Today there are at around thirty commercial suppliers across all segments of the market, certain of them aggressively export-driven with affiliates opening up in other countries, though many of them relatively small with a limited customer base.

Technology transfer support is available through ANVAR, the French National Agency for the Valorising Innovation, which promotes and funds innovative projects especially for SMEs.

HLT Policy

France has always had a strong policy of national language support to underpin its cultural and ideological objectives. More recently, this strategic position has translated into more sustained support for French language technology development.

The High Council for the French Language includes a committee dedicated to Human Language Technology, which produced an influential report in 1999 that eventually led to the launch of the *Technolangue* programme in 2002. *Technolangue* is funded by the French Ministries of Research & Technology, Culture & Communication, and

Economy, Finances & Industry. This three-year programme aims to build a strong infrastructure to feed existing HLT-related development projects and meet the marketplace need for industrial-strength linguistic data.

France also has a number of other national networks and organisations that support HLT development in various ways. These include The French National Network for Research in Telecommunications (RNRT), the French National Network for Technologies Software (RNTL) and the French Network for Audio-visual end Multimedia Research (RIAM). In addition, the Francophone Agency for the French Language (AUPELF), which has international scope, provides support for language technology activities.

HLT Scorecard: 4.7

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). France scores above the EU average on all language-technology Benchmark indicators, making it one of the advanced players in this sector. Its Opportunity score coincides with the European median, pulling France out of the 'leaders' group due to the less competitive status of its economic and infrastructure environment indicators.

The challenge for France is to ensure that its strong potential for exploiting its language technology assets is matched by further support for technology transfer, to give it greater competitive advantage in the marketplace.

HLT Suppliers

France has around 30 suppliers of HLT products and services - for example: Telisma, Temis, Systran, Vecsys, Sinequa, Lingway, Elan, Auralog, Noematics, Semantia, Xerox Research Centre Europe.

HLT Labs

France has around 25 research labs working in the

HLT field, including LIMSI, LORIA, LADL, TaLaNa, CRIM-INALCO, ENST, CLIPS, GRESEC.

HLT Initiatives
Technolangue.

Opportunity Snapshot - France

Economy and Society - France	
△ Total Population	60,000,000
△ Languages (number of native speakers)	
French	52,000,000
Regional languages (circa 28)	5,000,000
Immigrant languages (circa 35)	3,000,000
△ % of citizens who can speak a language in addition to mother tongue	47 %
△ French-speaking Internet users in France	17,400,000
△ Total GDP (€ millions)	€ 1,310,000 M
GDP per capita	€ 21,900
RTD and Innovation – France	
△ Annual RTD Expenditure (€ millions)	€ 30,300 M
Total RTD as % of GDP	2.16 %
Public RTD Expenditure (€ millions)	€ 11,200 M
Business RTD Expenditure (€ millions)	€ 19,100 M
△ High-Tech Patents per 1M population	
European Patent Office	20.2
US Patent Office	13.3
△ Language Technology R&D	
Number of HLT Research Centres	25
Number of Active HLT Suppliers	30
ICT Infrastructure – France	
△ Number per 100 inhabitants:	
PCs	30.5
Internet Users	18.3
Mobile telephone subscriptions	58
Telephone lines	58
△ Computers with an Internet connection	6.3 %
△ ICT Spending	
ICT Expenditure as % of GDP	6.2 %
Per capita ICT expenditure	€ 1,446

HLT Indexes: Greece

Benchmark Scale: 1 (weakest) to 7 (strongest)
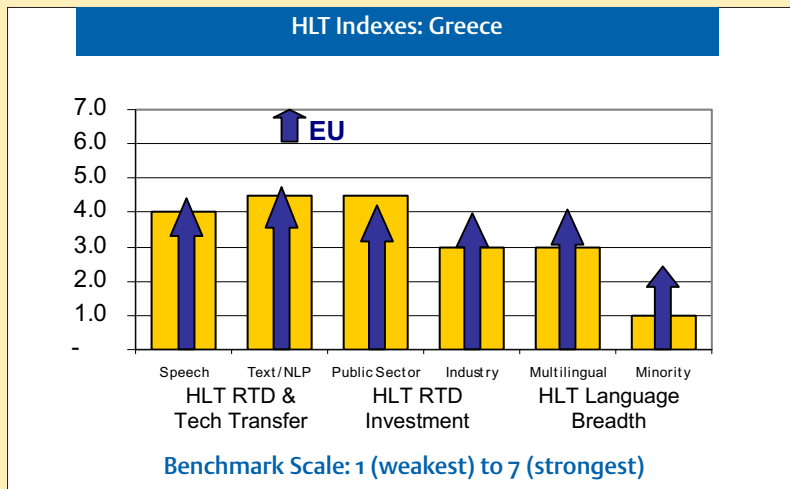
Greece has a relatively long tradition of national R&D in language and speech technology, and participates in a broad range of advanced EU projects in this field. The first results of this work are now finding their way onto the market as core language components.

Greece is strategically positioned as a bridgehead to the Balkans, Turkey and the Middle East, and is currently the only EU Member State to use a non-Latin script. Greece could, thus, benefit from developing a multilingual technology policy towards this region. The 2004 Olympics will also provide a real-world test bed for language technologies in general. The R&D community enjoys strong public support, but over the long term will benefit from more private investment.

HLT Benchmark: 3.4

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Greece scores just below the EU average on most measures in HLT research. The country has a relatively strong base in text and NLP applications, with in addition a practical focus on developing translation solutions for the Greek language.

There is also a fairly well developed speech research community, with the first signs of commercial text-to-speech and speech recognition technology development now emerging.

The country has a pool of ten research centres, with the government-funded ILSP (Institute for Language and Speech Processing) in Athens acting as the country's main centre of excellence for language technology. Greek researchers have been active in EU- funded projects for many years and have built up a substantial track record in collaborative research.

Most core language technology components have been developed for Greek, including lexica, parsers and syntax checkers. There is also ongoing work on collecting and preparing resources. As yet, however, multi-language technology research is less developed. ILSP has engineered the translation system in use at the European Commission to provide an on-line resource for Greek in-country civil servants.

Another area of national importance and a government priority for Greece is applying language technology to language learning contexts, particularly aiming at the large diasporic population of Greeks.

Technology Transfer

Greece has made progress in supporting technology transfer through government-funded programmes. As a result there are a number of language technology companies in operation in Greece, certain of which have benefited from this support. Several

companies are producing speech interface technology (including local start-ups as well as IBM Hellas); all commercial speech activity appears to be limited to the Greek language. There are, in addition, several organisations developing cross-language applications, one of which is Swedish but carries out R&D in Athens. Most, if not all, multi-language HLT tools and products are offered in Greece by re-sellers and integrators who source technology outside the country. ILSP, the state-funded institute, is the only organisation that works in both the speech and text/NLP domains, and the only one developing more advanced, knowledge-oriented technologies.

HLT Policy

The Greek government has supported language technology since the 1980s, through a series of projects within more general ICT research programmes. These began with the LOGOS programme in 1991 to prime the research infrastructure, and continued with the DIALOGOS project on improving man-machine communications through language technology up to 1998. A Greek HLT programme was launched in 1999 which included 12 large projects and 20 smaller research projects. A new programme involving language technology along with image and sound processing was planned to launch in May 2003.

Under the country's second Operational Programme for Research and Technology, the General Secretariat for Research and Technology (Ministry of Development) operates a network of national technological centres and technological parks. At the same time, Liaison Offices linking research and industrial communities have been set up to ensure that effective utilisation of research results.

The Ministry of Education also supports the Centre for the Greek language that fosters and promotes the language both inside and outside Greece.

HLT Scorecard: 3.3

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Greece scores just below the EU average on most "readiness" measures, indicating that there is still room for greater effort to create the potential for exploiting HLT research in the marketplace.

Transfer to market of Greek language technology will depend to a large degree on the more widespread penetration of ICT applications throughout the population, and more convincing demonstrations on the added value that such tools bring to business, government and education.

It is clear that greater public awareness of the benefits of language technology in a digital society and economy would help stimulate the market and thereby intensify the need for dynamic technology transfer, supported by new sources of capital.

HLT Suppliers

Greece has around 15 suppliers of HLT products and services - for example: Knowledge, Dialogos, Sena, ILSP, Neurosoft, Altec, Exodus, Voice-In.

HLT Labs

Greece has around ten research labs working in the HLT field, including: ILSP, SLT Group - University of Patras, Educational and Language Technology Laboratory-University of Athens, NCSR "Demokritos".

HLT Initiatives

LOGOS, DIALOGOS, Language Technology Network, Terminology Co-ordination, SOUND, IMAGE and LANGUAGE PROCESSING.

Opportunity Snapshot - Greece

Economy and Society – Greece	
△ Total Population	10,900,000
△ Languages (number of native speakers)	
Greek	10,400,000
Regional languages (circa 11)	400,000
Immigrant languages (circa 9)	500,000
△ % of citizens who can speak a language in addition to mother tongue	44 %
△ Greek-speaking Internet users in Greece	1,400,000
△ Total GDP (€ millions)	€ 116,800 M
GDP per capita	€ 11,000
RTD and Innovation - Greece	
△ Annual RTD Expenditure (€ millions)	€ 627 M
Total RTD as % of GDP	0.51 %
Public RTD Expenditure (€ millions)	€ 467 M
Business RTD Expenditure (€ millions)	€ 160 M
△ High-Tech Patents per 1M population	
European Patent Office	0.5
US Patent Office	0.5
△ Language Technology R&D	
Number of HLT Research Centres	9
Number of Active HLT Suppliers	15
ICT Infrastructure - Greece	
△ Number per 100 inhabitants:	
PCs	7.1
Internet Users	12.3
Mobile telephone subscriptions	55.9
Telephone lines	53.2
△ Computers with an Internet connection	14.8 %
△ ICT Spending	
ICT Expenditure as % of GDP	6.08 %
Per capita ICT expenditure	€ 691

Ireland

HLT Indexes: Ireland

Speech	Text/NLP	Public Sector	Industry	Multilingual	Minority
HLT RTD & Tech Transfer		HLT RTD Investment		HLT Language Breadth	

EU

7.0
6.0
5.0
4.0
3.0
2.0
1.0
-

Benchmark Scale: 1 (weakest) to 7 (strongest)
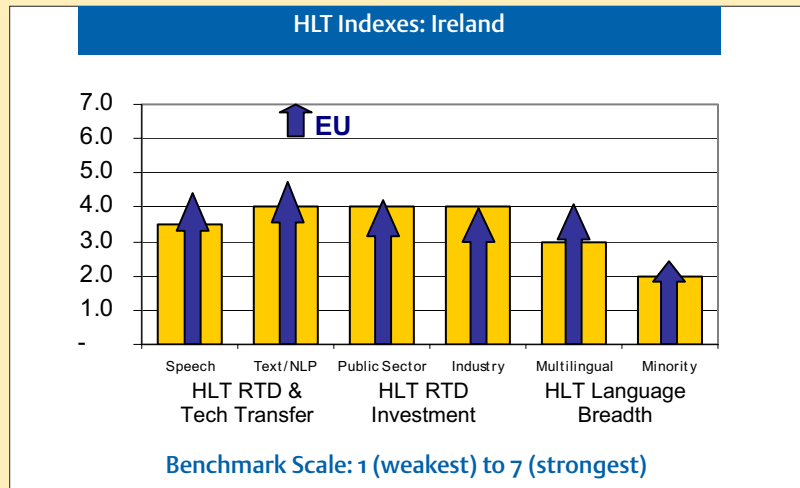
Ireland is a middle rank player in EU language technology, with an atypical profile. Sharing a language with the UK and the USA, the country has chosen to focus its resources on building a powerful IT service industry rather than on developing a national language research base. It hosts one of the world's largest concentrations of language localisation expertise and research, and acts as a key bridge between the European and North American software industry.

Although the Irish language in Ireland plays a substantially smaller role than English in business, government and education, this bilingual country has developed expertise in low-density language research and core technology.

HLT Benchmark: 3.5

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Ireland scores below the EU average on all measures of robustness in HLT research. This appears be due to national strategic choices concerning the development of a viable information society economy. Spared the need to devote a substantial effort to developing core English language tools, due to existing work in the UK and the USA, Ireland has focused most of its language technology efforts on enhancing the specific field of software localisation, which became a major source of national revenue in the 1990s. This has led to a high level of collaboration with industrial players in the vanguard of the IT revolution, and what amounts to world-class excellence in developing, evaluating and training in localisation productivity tools.

Ireland has ten research centres that focus in various ways on language technology, four of them industrial facilities with a close interest in applied research in translation and localisation activities. There is a small community of speech researchers, as well as the presence of MediaLab Europe (associated with the Massachusetts Institute of Technology), which could potentially act as a centre of attraction for more advanced research agendas in the interaction/-interface field.

In spite of the clear focus on translation software tools, however, the country scores substantially below the EU average for 'HLT language breadth' since the research effort is devoted more to software and management processes than to developing linguistic resources or multiple languages as such.

Technology Transfer

Given Ireland's preferred focus on the localisation sector - an application area rather than a source of broad-based language technologies – there has been minimal transfer of language technology to the

marketplace. Several localisation suppliers have developed tools that have been commercialised; in addition, the largest global localisation supplier has its European headquarters in Dublin, hosting a range of research activities related to translation automation.

Through its Enterprise Ireland scheme and Innovation Relay Centres, now managed by a new Technology Transfer & Business Partnerships service, Ireland offers a number of mechanisms by which technology transfer can be achieved, and can claim relative success in other disciplines.

The strong presence of industrial research centres has, however, meant that various forms of existing language technology have been successfully adapted and integrated into proprietary systems in the field of software localisation, and to a smaller degree in call centre dialogue management.

HLT Policy

Ireland has not established a dedicated national HLT programme, but has funded small academic projects that cover English language and Irish language core technologies through the government's Research, Technological Development and Innovation activities. More generally, the country's strategy of attracting new inward investment in the international services sector has had the effect encouraging foreign enterprises into Ireland, among them the substantial localisation sector.

The government launched a R&D Capability Scheme in 2000 to support larger firms to make new investments in capital and human resources. This may have the effect of encouraging greater localisation R&D uptake among those same IT companies who were attracted to Ireland in the first place.

Despite extensive use of English in everyday life, the Irish language is constitutionally recognised as the nation's first official language, and government policy states that every citizen has the right to conduct business with the public service through Irish. There is some funding provided for research into the processing of Irish.

The National Centre for Language Technology (NCLT) is hosted at the School of Computer Applications at Dublin City University (DCU). The NCLT was set up (in 1988) when DCU was nominated as the Irish research base for the European Commission's Eurotra Machine Translation R&D project. The HLT research community has now grown substantially in Ireland, but principally with funding from European, rather than Irish, programmes.

Ireland is also home to the Localisation Research Centre (LRC), based at the University of Limerick, an information, educational, and research centre for the localisation community. The LRC provides a comprehensive information service to the localisation industry, and while its focus is on Irish-based companies, it is also active in wider European initiatives. The centre conducts research and development in localisation and related areas, organises regular conferences and meetings, produces a range of publications, and oversees a number of education and training programs. The LRC maintains a library and showcase of localisation tools that can be used for evaluation. The LRC is largely funded by industry and by participation in EU programmes, rather than directly by Ireland.

HLT Scorecard: 4.4

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Ireland scores relatively highly in terms of national 'readiness' due to its above EU average commitment to enabling businesses and industry as well as the education and training sector to invest in information society technologies.

Ireland's strategic decision to support the IT localisation industry in the 1990s may well need to be revised if this sector is to remain competitive. By taking advantage of more advanced language technologies in multilingual content management, it should be

able to establish expertise in emerging niche markets in this sector.

HLT Suppliers

Ireland has a small number of suppliers of HLT products and services including: Voice Logics, Bowne Global Solutions, Lotus (IBM), and Alchemy.

HLT Labs

Ireland has around 10 research labs working in the HLT field, including: Computational Linguistics Lab, TCD, National Centre for Language Technology (Dublin City University) Bowne Global Solutions, Dublin, DSP Group, University College Dublin, MediaLab Adaptive Technologies group (with UCD), Sun Microsystems, Ireland Research Lab, Cognitive Language Modelling, National Univ of Maynooth, Language and Intelligence, Dublin City University, CLSC, Trinity College Dublin, IBM / Lotus Research, Dublin, Linguistics Institute of Ireland.

HLT Initiatives

[not funded directly in Irish programmes]: Localisation Research Center (LRC), National Centre for Language Technology (NCLT).

Opportunity Snapshot - Ireland

Economy and Society - Ireland	
△ Total Population	3,800,000
△ Languages (number of native speakers)	
English	3,540,000
Irish	260,000
△ % of citizens who can speak a language in addition to mother tongue	33 %
△ Internet users in Ireland	1,070,000
△ Total GDP (€ millions)	€ 103,275 M
GDP per capita	€ 25,825
RTD and Innovation – Ireland	
△ Annual RTD Expenditure (€ millions)	€ 1,428 M
Total RTD as % of GDP	1.38 %
Public RTD Expenditure (€ millions)	€ 362 M
Business RTD Expenditure (€ millions)	€ 1,066 M
△ High-Tech Patents per 1M population	
European Patent Office	13.3
US Patent Office	3.8
△ Language Technology R&D	
Number of HLT Research Centres	10
Number of Active HLT Suppliers	3
ICT Infrastructure – Ireland	
△ Number per 100 inhabitants:	
PCs	36.5
Internet Users	25.0
Mobile telephone subscriptions	66.8
Telephone lines	42.6
△ Computers with an Internet connection	8.1 %
△ ICT Spending	
ICT Expenditure as % of GDP	5.35 %
Per capita ICT expenditure	€ 1,290

Italy

HLT Indexes: Italy

Benchmark Scale: 1 (weakest) to 7 (strongest)
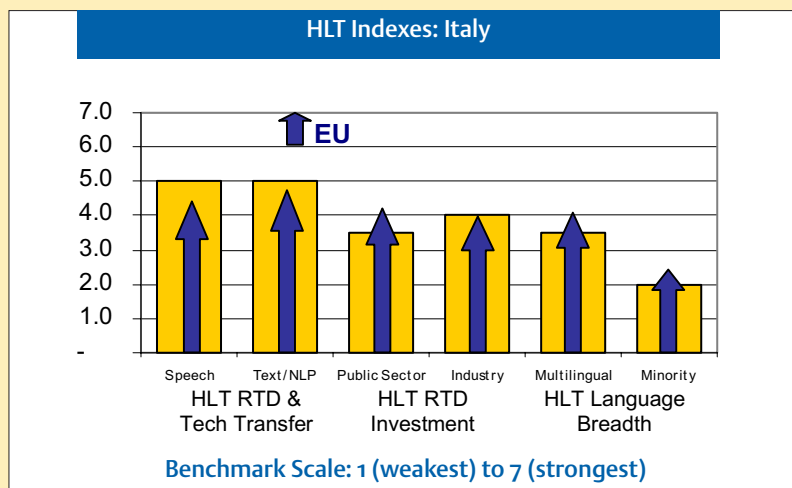
Italy has one of Europe's longest traditions of research in HLT, and has recently developed a substantial supplier base. National support for HLT has, however, been somewhat inconsistent, and Italy has only recently begun to plan for a national research policy in the language technology field. In addition, market factors such as innovation potential and ICT take-up are relatively weak in Italy, posing additional challenges to HLT exploitation nationally. Italy has actively participated in EC research projects, and has been instrumental in initiating actions focused on industry standards and evaluation. Recently renewed government support for affirmative language technology action could improve the research and technology transfer potential of the excellent research base in Italy.

Italian is one of Europe's larger language populations, and Italy has very small cross-border speaker communities. Linguistic diversity in Italy is driven by regional languages, and more than half the population (55%) use one of the eight major regional languages (Emiliano, Lombard, Ligurian, Napoletano, Piemontese, Sard, Sicilian, and Venetian). Another 20+ regional languages have very small speaker populations.

HLT Benchmark: 3.9

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Italy scores just below the EU average on measures of robustness in HLT research. There has been a consistent tradition of text and NLP applications, and speech technology research has led to the creation of a world-class supplier. The country has around 18 HLT research centres, one of which has been active since the 1970s. However, public-sector investment in HLT has been sporadic, and Italy has had no major national programmes for development of HLT applications, though smaller initiatives have been funded. Italy has not devoted significant resources to processing languages other than standard Italian. Italian researchers have an excellent record of cross-border collaboration, and participation in EU-funded programmes.

Industry involvement in HLT research stands at the European norm. With around 25 suppliers, Italy has an unusually large commercial language technology population, even though not all of these are hardcore HLT players. Research in cross-lingual applications is relatively weak in Italy, compared to other comparably sized Member States.

Technology Transfer

Italy has developed a network of regional technology transfer centres that support the process of bringing high technology research results to the marketplace. Although there have been numerous suc-

cesses, Italy's record in new business formation is substantially lower than most other EU countries, which theoretically has the effect of reducing opportunities open to potential language technology companies.

At the same time, Italy scores well on channel access, with the existence of a number of large-scale national IT industries and a developed telecommunications culture, both factors likely to encourage the transfer of technologies to the market.

HLT Policy

Italy was not among the larger EU countries that launched large-scale HLT programmes during the late 1980s and early 1990s when the field began to attract public funding, even though the country's HLT research has always benefited from national and especially from EU information technology programmes. In 1997, HLT was designated a national research policy, with the launch of two three-year projects: TAL – a national framework for developing language resources, and LRCMM, devoted to mono- and multilingual research in computational linguistics, with a view to strengthening innovation in this field.

In 2002, a further public commitment was made to HLT with a plan to create an official forum on language technologies as part of the Ministry of Telecommunications strategic plan. The forum, which is still in the process of approval, will include a network of industry and research actors to provide policy guidance in the HLT field. Driving this initiative was the recognition that the Italian language needs to be maintained within the global language economy as the preferred medium for the country's citizens.

Additional support for HLT research is available through general funding mechanisms of the Ministry for Higher Education, Training & Research (MIUR) and the Italian National Research Council (CNR). The HLT Network is a 30-member association of representatives from ministries, public administration, industry, universities and research groups.

The network is a discussion group on issues related to HLT.

HLT Scorecard: 4.0

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Italy scores below the European average on many "readiness" measures, indicating only a medium potential for exploitation of HLT research. Only some of the current supplier base has been directly resourced from technology transfer from the national research base.

Italy will benefit from its recent decision to develop a stronger national framework of standards-compliant resource and tool developers, working in tandem with funding agencies, to ensure a more centralised, cost-effective language technology fabric. The key challenge will be to overcome the uneven impact of the private sector economic environment on its capacity to develop an aggressive market-centric policy, and to gain language market share beyond its national frontiers.

HLT Suppliers

Italy has more than 25 suppliers of HLT products and services - for example: Abla, ACP, Advanced Computer Systems, Alceo, CEDAT85, CELI, Cirte Manifatturiera, D'Agostini Organizzazione, DIDAEL, Eptamedia, Eulogos, Expert System, Giunti-multimedia, GST, Hi-Flier, Itaca, Loquendo, Mediavoice, Necsy, Omega, Quinary, Rigel, Synthema, Thamus, Yana Research, YourVoice.

HLT Labs

Italy has more than 15 research labs working in the HLT field, including: CNR–Istituto di Linguistica Computazionale – Pisa, CNR – Istituto per le scienze della cognizione (Institute for Cognitive Sciences) – Sezione di Padova (phonetics, speech technologies), Eulogos, Scuola Normale Superiore di Pisa – Centro per la fonetica sperimentale (Centre for Experimental Phonetics), Istituto Trentino di Cultura - Centro per la Ricerca Scientifica e Tecnologica ITC-

IRST, Centro Ricerche Fiat, Fondazione Ugo Bordoni, Università degli Studi di Ancona, Università di Bari - Sistemi di Elab. dell'Informazione, Universita degli Studi di Firenze, Università degli Studi di Genova, Università degli Studi di Napoli – CIRASS, Universitòdegli Studi di Roma 3 Tor Vergata, Università degli Studi di Torino - Dipartimento di informatica, Università degli Studi di Udine, Universita degli Studi di Venezia Cà Foscari- Laboratorio di Linguistica Computazionale, Università degli studi di Verona.

HLT Initiatives
Forum for HLT in Italy, HLT Network, The National Project in Natural Language Processing.

Opportunity Snapshot - Italy

Economy and Society - Italy	
△ Total Population	58,000,000
△ Languages (number of native speakers)	
Italian	55,000,000
Major regional languages (Emiliano, Lombard, Ligurian, Napoletano, Piemontese, Sard, Sicilian, Venetian)	32,400,000
Minor regional languages (circa 24)	1,500,000
△ % of citizens who can speak a language in addition to mother tongue	46 %
△ Internet users in Italy	22,600,000
△ Total GDP (€ millions)	€ 1,100,000 M
GDP per capita	€ 18,900
RTD and Innovation - Italy	
△ Annual RTD Expenditure (€ millions)	€ 12.123 M
Total RTD as % of GDP	1 %
Public RTD Expenditure (€ millions)	€ 5,595 M
Business RTD Expenditure (€ millions)	€ 6,528 M
△ High-Tech Patents per 1M population	
European Patent Office	4.8
US Patent Office	4.2
△ Language Technology R&D	
Number of HLT Research Centres	17
Number of Active HLT Suppliers	26
ICT Infrastructure - Italy	
△ Number per 100 inhabitants:	
PCs	20.9
Internet Users	19.1
Mobile telephone subscriptions	73.7
Telephone lines	47.4
△ Computers with an Internet connection	12.8 %
△ ICT Spending	
ICT Expenditure as % of GDP	5.5 %
Per capita ICT expenditure	€ 1,065

HLT Indexes: The Netherlands

Benchmark Scale: 1 (weakest) to 7 (strongest)
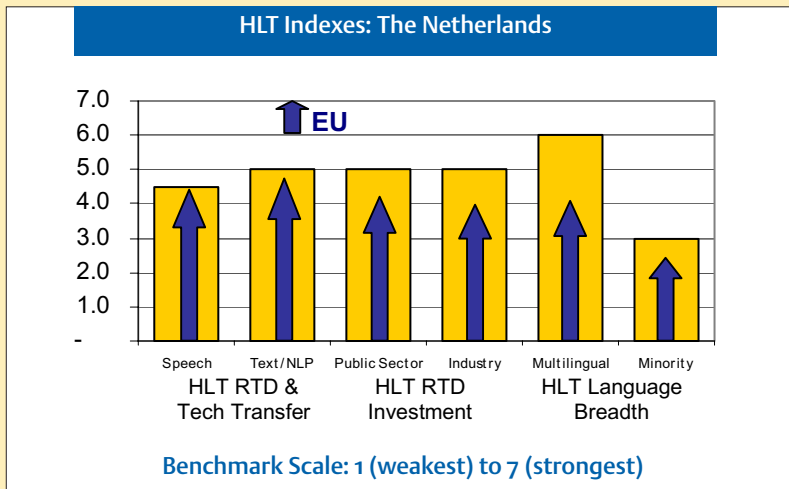
The Netherlands

HLT in The Netherlands benefits from a healthy research tradition, significant support for innovation, and a robust economic environment to support technology development. With one of the more advanced ICT infrastructures in Europe, citizens and companies in The Netherlands are readily able to accept and absorb advanced products and services that incorporate HLT.

The Dutch Language Union provides a structure for collaboration between Europe's two Dutch-speaking regions (The Netherlands and Flanders), and is in the forefront of the definition and development of standards for core HLT components.

Dutch is the native language of the overwhelming majority (95%) of citizens of The Netherlands. National policy strongly supports HLT for Dutch, but there is little research attention either to minority languages and dialects (such as Frisian), or to immigrant languages. Dutch is a relatively low-density language, and most speakers live in Europe. This limits the market opportunity for Dutch-language HLT, and is no doubt responsible for the strong multi-language focus of much HLT research in the Netherlands. In spite of this, however, the number of cross-language products is still low.

HLT Benchmark: 4.9

The Netherlands is a leader in European HLT, and ranks near the top in overall potential, and in RTD investment policy. The country ranks in the middle tier for research and technology transfer, largely because the number of start-ups and examples of commercial innovation are low relative to the strength of the research base.

Netherlands HLT research is biased toward public-sector-funded programmes. Research is carried out in 17 departments at 10 Universities, with an additional 10 Research Institutes involved in research relevant to HLT.

The development of core HLT components for the Dutch language is extremely well advanced, and there is every reason to suppose that next-generation HLT products and services will push into more advanced application areas. The Netherlands HLT research agenda encompasses the full range of relevant disciplines to make this happen, including speech recognition, NLP using both knowledge-based and probabilistic methodologies, dialogue management and output generation (in both NLP and speech), and human factors research. The Netherlands has an outstanding track record in multilingual HLT. There is less evidence, however, of strong research in cross-language capabilities.

Technology Transfer

There is a small but growing number of companies in The Netherlands developing and supplying products and services based on HLT. EUROMAP has identified 15 companies in the field. Most commercial activity is in knowledge applications based on text, or interface applications based on speech. Over half these companies supply basic speech or language components, as well as applications. Almost all companies are focused on either speech or text, with few examples integrating different HLT technologies.

There is little commercial activity in cross-language applications, and no national supplier of machine translation for the Dutch language - a significant gap in coverage probably due, in part, to the low density of the Dutch language. No free Dutch gisting engine is available on the Web. Weakness in the cross-language focus may also reflect a shift in the centre of gravity of the localisation industry from The Netherlands (which was originally a leader in this field) to Ireland, which occurred in the 1990s. Revitalising the development of cross-language products and services is a notable opportunity. Language services are a natural market for The Netherlands, which has the most multilingual citizency in the EU; 75% of Dutch people speak English, over half also speak German, and 87% can speak at least one additional language.

Despite the lack of cross-language tools, many of the products developed in The Netherlands are available in multiple languages, reflecting the strong tradition in addressing many languages in HLT research and applications. Only one or two suppliers has confined its product development exclusively to the Dutch language. This multi-language approach suggests that Netherlands-based suppliers are in a strong position to reach and service a pan-European market, though few products extend beyond European languages.

HLT Policy

There is strong policy support for HLT in The Netherlands, and nothing exemplifies this more than the Platform for Dutch HLT, an initiative of the Dutch Language Union (NTU, Nederlandse Taalunie). The programme has support from all relevant actors, including the Ministry of Education, Culture and Science (OC&W), the Ministry of Economic Affairs (EZ), the Netherlands Organisation for Scientific Research NWO and Senter/EG-Liaison (the body that promotes participation by academics and companies in funded R&D programmes).

The Dutch HLT Platform is a collaboration between institutions in The Netherlands and Flanders. It promotes networking between researchers and companies to encourage participation in European-level projects. At a tactical level, the platform has established priorities for further development of basic Dutch-language HLT components, and determined the cost of doing so. It has set criteria for creating core components as well as a blueprint for managing, maintaining, making available and distributing the basic Dutch-language resources that can be used in education and research and for developing HLT tools and applications.

HLT Scorecard: 5.2

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). The Netherlands scores above the EU average on every opportunity measure that supports healthy development of the HLT sector, and with Finland, Germany and the UK has the highest HLT Scorecard in the EU.

EUROMAP research suggests that The Netherlands is in a strong position to be a leader in the HLT field, and has the potential to develop world-class, advanced applications, products and services based on HLT research. This potential has not yet been achieved, however. The biggest challenge for the HLT community in The Netherlands will be to enhance the ability to transfer HLT research to market. This will entail exploitation of the strong multilingual focus of HLT research, and opportunities to develop HLT functions with a potential for pan-European distribution.

HLT Suppliers

The Netherlands has approximately 15 suppliers of HLT products and services - for example: Fluency/ Van Dale Data, Human Inference, Knowledge Concepts, Linguistic Systems, Polder land Language & Speech Technology, *TALO, Comsys, Compuleer, HuQ, Sentient Machine Research.

HLT Labs

The Netherlands has more than 25 research labs working in the HLT field, including: University of Twente (CTIT, LE Group), University of Nijmegen (NIII, NICI), Tilberg U. (ITK, Centre for Language Studies), TNO.

HLT Initiatives

Nederlandse Taalunie (Dutch Language Union), Dutch HLT Platform, Technology Radar/Informe (initiatives of Ministry of Economic Affairs), Spoken Dutch Corpus.

Opportunity Snapshot - The Netherlands

Economy and Society - The Netherlands	
△ Total Population	16,000,000
△ Languages (number of native speakers)	
Dutch	15,000,000
German/Frisian, plus dialects, and immigrant languages	1,000,000
△ % of citizens who can speak a language in addition to mother tongue	87%
△ Dutch-speaking Internet users in Netherlands	9,700,000
△ Gross Domestic Product	
Total GDP (€ millions)	€ 380,000 M
GDP per capita	€ 24,000
RTD and Innovation - The Netherlands	
△ Annual RTD Expenditure (€ millions)	€ 7,700 M
Total RTD as % of GDP	1.9%
Public RTD Expenditure (€ millions)	€ 3,489 M
Business RTD Expenditure (€ millions)	€ 4,211 M
△ High-Tech Patents per 1M population	
European Patent Office	35.8
US Patent Office	19.6
△ Language Technology R&D	
Number of HLT Research Centres	27
Number of Active HLT Suppliers	15
ICT Infrastructure - The Netherlands	
△ Number per 100 inhabitants:	
PCs	39.5
Internet Users	42.5
Mobile telephone subscriptions	67.1
Telephone lines	60.7
△ Computers with an Internet connection	25.8%
△ ICT Spending	
ICT Expenditure as % of GDP	6.9%
Per capita ICT expenditure	€ 1,657

HLT Indexes: Portugal

Benchmark Scale: 1 (weakest) to 7 (strongest)
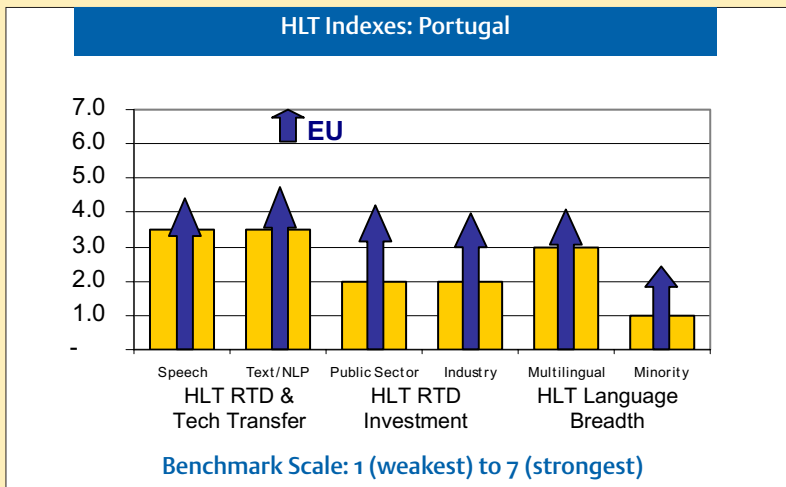
Portugal has a small but reputable language and speech technology research base, equipped with a special resources centre, but has so far been unable to develop sufficient critical technology mass to transfer results to market-ready applications. Under-performance in the country's economy and business investment community is partly responsible for this. Investment from both the public sector and from industry has been lacking for HLT in Portugal.

Portugal has a national language population – and therefore market - that is smaller than its cognate language population in Brazil, which is more active commercially in this sector. More than 10% of the population has a native language other than Portuguese, the vast majority from Africa (excluding Angola and Mozambique) and North Africa. There are regional language populations but they are extremely small (less than 1% of population). Due to a long tradition of emigration, many Portuguese are relatively multilingual in outlook, which ironically could act as a brake on developing real-world language technology solutions for their own requirements.

HLT Benchmark: 2.6

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Portugal scores below the EU average on all measures of robustness in HLT research. There is an active speech research community, but so far there has been little attempt to transfer results to the marketplace, with the result that cross-border suppliers are benefiting from first-mover status among the well-developed telecommunications companies.

Basic NLP and text research in Portuguese was carried out during the 1990s and there is continuing work on text technologies, but there is little evidence that this initiative has developed, for example, industrial strength translation modules to serve growing information society demands.

Portugal has one of the lowest R&D intensities in the EU, with similarly low levels of public and private investment in language technology. Although historically Portugal has been an 'outward facing' country, with a tradition of emigration and overseas trade - and hence multilingual in practice - strong multi-language research is not a priority.

Technology Transfer

Portugal has a very low rate of venture capital investment in general, and a traditional focus on low-R&D-intensity industry and manufacturing. Despite the active work of the AITEC incubator associated with INESC, a major academic centre, to spin-off new ICT companies, venture capital support appears not to have been forthcoming in the field of language and speech technology. The government is trying to
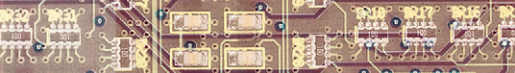
remedy this situation with new patenting-incitement and other initiatives, but these will take some time to come into effect.

Currently, there appears to be only one viable language technology product vendor in Portugal today, supplying first generation core technology dictionary and proofing tools. This means that Portuguese businesses in general do not have access to language technologies vital to knowledge management and competitive intelligence.

Brazilian Portuguese technology and service suppliers, on the other hand, appear to be more proactive and present in the marketplace, which will make it harder for Portuguese HLT companies to develop and expand their market share.

HLT Policy
Portugal first funded language technology research through a general IT programme in the 1990s but has not yet launched a dedicated HLT programme. A framework contract established between the National Board for Science and Technology (JNICT) and the Institute of Theoretical and Computational Language (ILTEC) forms the basis for research funding.

Recent government efforts to boost ICT spending and promote greater home internet access – both vital 'readiness' factors in language technology take-up - are to be welcomed.

HLT Scorecard: 3.3
The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Portugal scores some way below average on most "readiness" measures, indicating that potential for exploitation of HLT research is not yet acceptable. Combined with the low HLT Benchmark scores, Portugal emerges as the country that has the most catching up to do of all EU countries in terms of technology transfer support.

To meet this challenge, a policy of more ambitious cross-border technology partnerships in this sector may be worth considering if Portugal wishes to achieve a market presence for Portuguese language technologies. More generally, development and exploitation of Portuguese knowledge assets will depend on remediation in the HLT domain. This will ensure that Portugal's well-established research capabilities can translate into appreciable benefits for Portuguese-speaking citizens.

HLT Suppliers
Portugal appears to have few commercial HLT suppliers; the only one identified by EUROMAP is Porteditores.

HLT Labs
Portugal has around seven research labs working in the HLT field, including: National Institute of Theoretical and Computational Language – ILTEC, L^2F, Spoken Language Systems Lab, Univ Lisbon, Neural Network Group, INESC CSTC, Lisbon Technical University, Linguateca Foundation and Projects, gEPL, University of Minho Braga.

HLT Initiatives
None identified.

Opportunity Snapshot - Portugal

Economy and Society - Portugal	
△ Total Population	10,000,000
△ Languages (number of native speakers)	
Portuguese	8,950,000
Regional languages (circa 5)	50,000
Immigrant languages (circa 10)	1,000,000
△ % of citizens who can speak a language in addition to mother tongue	33 %
△ Internet users in Portugal	4,400,000
△ Total GDP (€millions)	€110,000 M
GDP per capita	€ 10,900
RTD and Innovation - Portugal	
△ Annual RTD Expenditure (€ millions)	€ 622 M
Total RTD as % of GDP	0.54 %
Public RTD Expenditure (€ millions)	€ 461 M
Business RTD Expenditure (€ millions)	€ 161 M
△ High-Tech Patents per 1M population	
European Patent Office	0.4
US Patent Office	0.1
△ Language Technology R&D	
Number of HLT Research Centres	7
Number of Active HLT Suppliers	1
ICT Infrastructure - Portugal	
△ Number per 100 inhabitants:	
PCs	10.5
Internet Users	30.2
Mobile telephone subscriptions	66.5
Telephone lines	43.1
△ Computers with an Internet connection	5.9 %
△ ICT Spending	
ICT Expenditure as % of GDP	7.01 %
Per capita ICT expenditure	€ 747

Sweden

EU

	Speech	Text/NLP	Public Sector	Industry	Multilingual	Minority

HLT RTD &
Tech Transfer

HLT RTD
Investment

HLT Language
Breadth

Benchmark Scale: 1 (weakest) to 7 (strongest)
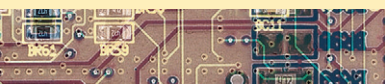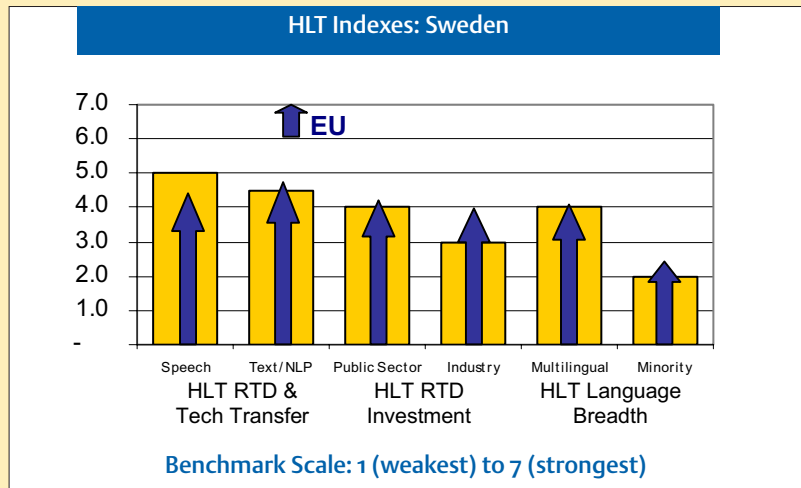
Sweden has one of the most advanced knowledge economies in the world, and is currently devoting considerable level of resources to ensuring that its language technology is ready to meet the new generation of communicative challenges in the most appropriate way. The country enjoys a strong research community, particularly in speech technology, but commercial language technology has not developed to a corresponding degree.

Sweden is relatively linguistically diverse; 75% of the population is native Swedish speaking, and 15% are Skåne speakers. Another 8% speak either minority regional languages (4%) or immigrant languages (4%). In addition, English plays a key role in the country's international communications. The Swedish government is taking action to ensure that language technology can be used to preserve and improve communication in Swedish for all, and that other languages are supported where necessary.

HLT Benchmark: 3.9

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. Sweden scores close to the EU average for measures involving language technology research and development. Although it has a long tradition in speech science and linguistics, the development of a robust language and speech technology community has taken longer than in certain comparable countries.

Today there is a renewed thrust to boost the training of a new generation of language technology specialists, focused around the Graduate School for Language Technology, co-ordinated by the Faculty of Arts in the University of Göteborg, a close collaboration between all language technology and computational linguistic research departments in the country's universities.

There are around ten university research centres dedicated to computational linguistic or language and speech technology, and in addition to building core technologies for the language, there is advanced work in cognitive linguistics, human interface and dialogue design, and advanced speech processing.

Technology Transfer

Sweden scores well above average on all indicators for supply-side readiness, and offers a potentially advantageous business formation environment for language technology suppliers. The government has created a range of agencies and programmes to promote new businesses, transfer technology to SMEs, and foster closer R&D-industry collaboration.
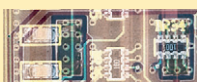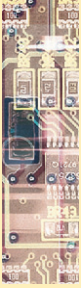
Despite this generally nurturing environment and the availability of capital, the number of successful Swedish language technology companies is relatively limited – five in the highly competitive speech area and three in text and NLP applications. This is partly due to Sweden's relatively recent decision to promote language engineering (as opposed to theoretical or computational linguistics) as a discipline of national relevance.

HLT Policy

Today, Sweden enjoys national level support for language technology through the creation in 2001 of VINNOVA (the Swedish Agency for Innovation Systems) whose role is to fund research, encourage university-industry collaboration, and boost innovation in the ICT sector in general.

VINNOVA is responsible for its own national scale programmes, one of which is a Human Language Technologies action line due to run until 2006. This initiative aims to develop generic technologies for Swedish and other languages, and expand knowledge of how language systems can boost the effectiveness of ICT systems.

In part this entails a strong training effort, a domain to which Sweden devotes considerable resources, with extensive use of online learning environments.

Sweden is also extremely sensitive to the quality and sustainability of its own language. The Committee on the Swedish Language is committed to the promoting standardised 'plain Swedish' in the administration, and ensuring that the language is 'accessible to all citizens'. It acknowledges the role of language technology in achieving these goals, and certain of its recommendations - the creation of a language technology secretariat, and the development of machine translation for Swedish – have been taken up in VINNOVA's action line.

HLT Scorecard: 5.2

The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). Sweden enjoys the highest HLT Opportunity scores in the EU, due to its highly competitive knowledge society infrastructure. This should favour the rapid development of market-ready technologies when it effectively consolidates its language technology expertise.

Sweden is faced with the challenge of both delivering a broad range of first generation language tools and technologies to the market at the earliest opportunity, while at the same time sustaining its longer-term ambitions of conducting next generation research.

HLT Suppliers

Sweden has around 10 suppliers of HLT products and services - for example: ESTeam, Euroling AB, Comintell, HT Speech Systems, Hapax, Icepeak, Pipebeach, Telia, Promotor, Voxi.

HLT Labs

Sweden has more than 10 research labs working in the HLT field, including: departments and centres at Göteborg University, Linköping University, Lund University, University of Skövde, Stockholm University, Uppsala University, Chalmers University of Technology, Royal Institute of Technology, and SICS, Swedish Institute of Computer Science AB.

HLT Initiatives

VINNOVA HLT Programme.
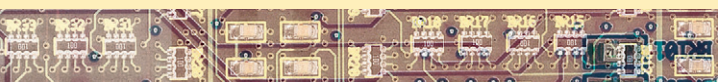
Opportunity Snapshot - Sweden

Economy and Society - Sweden	
△ Total Population	8,900,000
△ Languages (number of native speakers)	
Swedish	6,670,000
Skåne	1,500,000
Other regional languages (circa 12)	346,000
Immigrant languages (circa 20)	365,000
△ % of citizens who can speak a language in addition to mother tongue	81 %
△ Internet users in Sweden	6,000,000
△ Total GDP (€ millions)	€ 210,000 M
GDP per capita	€ 23,600
RTD and Innovation – Sweden	
Annual RTD Expenditure (€ millions)	€ 9,150 M
Total RTD as % of GDP	3.71 %
Public RTD Expenditure (€ millions)	€ 2,121 M
Business RTD Expenditure (€ millions)	€ 7,029 M
△ High-Tech Patents per 1M population	
European Patent Office	22.9
US Patent Office	29.5
△ Language Technology R&D	
Number of HLT Research Centres	11
Number of Active HLT Suppliers	9
ICT Infrastructure – Sweden	
△ Number per 100 inhabitants:	
PCs	50.7
Internet Users	50.6
Mobile telephone subscriptions	71.4
Telephone lines	68.2
△ Computers with an Internet connection	13.2 %
△ ICT Spending	
ICT Expenditure as % of GDP	8.27 %
Per capita ICT expenditure	€ 2,060

HLT Indexes: United Kingdom

| | Speech | Text/NLP | Public Sector | Industry | Multilingual | Minority |

HLT RTD & Tech Transfer   HLT RTD Investment   HLT Language Breadth

Benchmark Scale: 1 (weakest) to 7 (strongest)
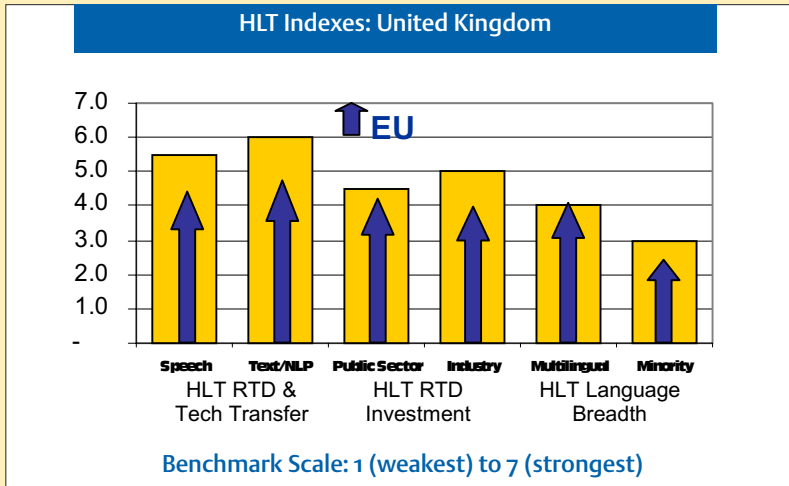
United Kingdom

The United Kingdom has been one of the leading European language technology players for some two decades. Although facilitated by sharing a global language with the USA, the UK has built an advanced research community and been highly successful in transferring innovative results to the market. Largely driven by a dynamic venture capital culture and favoured by its business environment, the UK today fields a number of world-class supplier companies targeting cross-border markets.

Although speaking the world's premier international language, the UK hosts minority languages such as Welsh, and substantial populations of immigrant languages from the Indian subcontinent. Programme support came early to the UK, but government investment has been less active in recent years.

HLT Benchmark: 4.7

The HLT Benchmark measures the relative maturity of language technology research and development in the EU. The UK scores well above average on all measures of robustness in HLT research, apart from multilinguality, where it ranks average. The global range of the national language may have initially inhibited research in multiple languages, but recently there has been more focus beyond English, both for high-density commercial languages (such as Japanese) and increasingly for immigrant languages (such as those from the sub-continent). Both speech and text/NLP research tracks have focused not only on developing individual components to a high degree of maturity, but also on providing the architectures and platforms that play a key role in enabling language technology to integrate with software engineering and IT standards generally.

There are some 20 HLT research centres in the UK, at least three of them funded by industry, working right across the language/speech technology spectrum, from fundamental research in computational linguistics paradigms to experimental multidisciplinary agendas such as multimodality, cognitive interfaces and usability issues. As well as being one of the very first sites to develop strong expertise in corpus linguistics, the UK also has a track record in licensing research-driven language technology systems. This has meant that core components have been available for a relatively long time, which in turn has resourced greater experimentation in research programmes.

Technology Transfer

The United Kingdom scores relatively well for supply-side readiness, and for innovation potential, which has in part facilitated the path to market for language technology research results. A dynamic venture

capital culture has also largely contributed to the availability of start-up funding for new businesses, the UK being the EU's top scorer in terms of new business formation in recent years.

The UK also provides technology developers with a high level of access to channels. It experienced early liberalisation of its telecommunications market, hosts the largest financial market in Europe, as well as many first-adopter IT companies capable of exploring the possibilities of integrating new technologies.

As a result, the UK currently has some 35 technology suppliers, ranging across all categories, from knowledge management tools and language-driven taxonomy systems to various speech technology and dialogue system applications. A number of these companies are export focused and increasingly capable of adapting their products to multiple languages.

Many of these commercial undertakings are spin-offs from university research, which in turn testifies to a fairly healthy technology transfer environment. And the large majority are now dedicated to providing more comprehensive embedded technology solutions, rather than core components.

HLT Policy
UK support for language technology research began in the 1990s with the Department of Trade and Industry's four-year Speech and Language Technology (SALT) programme, which funded a wide range of small collaborative projects in the field. Since that period, there has been no major programme specifically involving language technology support, although there are numerous public sources of research, development and technology transfer funding.

The EPSRC is the UK's main agency for funding research in engineering and physical sciences, including information technology. The Information Technology and Computer Science Programme (IT&CS) is most relevant to HLT; of particular interest to HLT researchers is funding for Human Com-

munication and the Human Computer Interaction research. The EPSRC funds a number of HLT projects in the UK. In addition, the UK Research Council's Basic Technology Research Programme is designed to create fundamental new capabilities that will underpin industries of the future, seeking to transcend research council boundaries and give researchers the chance to explore radical new ideas. This programme supports co-operation and collaboration across disciplines, and is particularly relevant to HLT, which spans traditional research boundaries.

To a large degree, the UK research and technology development community acts a lobbying body dedicated to prompting funding policy actions. SALT, for example, evolved into a self-perpetuating research constituency of more than 300 members from academia and industry. Today, CLUK (Computational Linguistics UK), a collaborative group of research institutions, has taken over this role of representing the views of the UK computational linguistics community in UK funding bodies.

HLT Scorecard: 5.2
The HLT Scorecard compares the HLT benchmark with neutral, third-party measures of the business environment and infrastructure that promote the take-up of HLT (the Opportunity Index). The UK is among the very top scorers on both "readiness" measures, and on HLT research potential. This score reflects the relative maturity and critical mass of the UK's market-ready technology, its openness to potential commercial opportunities, and the research base's capacity to push towards more advanced, multidisciplinary agendas.

However, the lack of any coherent language technology policy capable of mobilising resources around ambitious new national projects may tend to exacerbate the competition for funding, and thereby dilute the UK's specifically national effort to maintain its leading position in this domain.

HLT Suppliers

The UK has more than 30 suppliers of HLT products and services - for example: 20/20 Speech, 3F Ltd, Aculab, AllVoice, Autonomy, BNC, BTexact Technologies, BlueChip Technologies, Collins, Dremedia, Fluency, Fourth Person, Gabrielle, GATE, Infogistics, Linguamatics, Lobal Technologies, Novauris Labs, Oxford University Press, Psytechnics, Rhetorical Systems, SDL plc, Softsound, Solcara, Speech Point, SPSS, SRC, Sysmedia, Telephonetics, Tisento, Transversal, Vocalis, Vox Generation, Wordmap.

HLT Labs

The UK has around 20 research labs working in the HLT field, including: U. of Aberdeen Computing Science, U. of Brighton Information Technology Research Institute (ITRI), Btexact Technologies, U. of Cambridge Natural Language Group & Speech Vision and Robotics Group, Canon Research Centre Europe, Cardiff U. Computational Linguistics Unit, UCL Department of Phonetics and Linguistics, U. of Edinburgh Centre for Speech Technology Research (CSTR), U. of Edinburgh and U. of Glasgow Human Communication Research Centre (HCRC), Keele U. Human Machine Perception Group, U. of Lancaster Unit for Computer Research on the English Language, U. of Leeds Centre for Computer Analysis of Language And Speech, Sharp Lab Europe, U. of Sheffield Speech and Hearing Research Group & Natural Language Processing Group, U. of Sunderland Natural Language Engineering Group, U. of Sussex Natural Language Processing and Computational Linguistics at COGS, UMIST Department of Language Engineering.

HLT Initiatives

SALT, CLUK

Opportunity Snapshot - UK

Economy and Society – UK

△	Total Population	60,000,000
△	Languages (number of native speakers)	
	English	55,000,000
	Welsh	510,000
	Other regional languages (Cornish, French, Gaelic, Romani, Scots)	300,000
	Immigrant languages (50 or more)	4,190,000
△	% of citizens who can speak a language in addition to mother tongue	27%
△	Number of Internet users	33,000,000
△	Gross Domestic Product	
	Total GDP (€ millions)	€ 1,425,000
	GDP per capita	€ 24,000

RTD and Innovation – UK

△	Annual RTD Expenditure (€ millions)	€ 27,700 M
	Total RTD as % of GDP	1.8 %
	Public RTD Expenditure (€ millions)	€ 9,133 M
	Business RTD Expenditure (€ millions)	€ 18,575 M
△	High-Tech Patents per 1M population	
	European Patent Office	18.9
	US Patent Office	14.4
△	Language Technology R&D	
	Number of HLT Research Centres	19
	Number of Active HLT Suppliers	33

ICT Infrastructure – UK

△	Number per 100 inhabitants:	
	PCs	33.5
	Internet Users	55.4
	Mobile telephone subscriptions	72.7
	Telephone lines	56.7
△	Computers with an Internet connection	8.3 %
△	ICT Spending	
	ICT Expenditure as % of GDP	7.4 %
	Per capita ICT expenditure	€ 1,709

Opportunity Snapshot - Consolidated Data

Measure/Factor	Austria	Belgium	Denmark	Finland	France	Germany	Greece	Ireland	Italy	Neth.	Portugal	Spain	Sweden	UK
ECONOMY & SOCIETY														
Total Population (M)	8.2	10.3	5.3	5.2	60.0	83.0	10.6	3.8	58.0	16.0	10.0	40.0	8.9	60.0
% of citizens who can speak a language in addition to mother tongue	61%	61%	85%	58%	47%	53%	44%	33%	46%	87%	33%	32%	81%	27%
Number of Internet users	3.7	3.5	3.4	2.1	17.4	37.1	1.4	1.1	22.6	9.7	4.4	11.3	6.0	33.0
Total GDP (€B)	188.5	230,000	162,000	121,000	1,310,000	1,850,000	116,800	103,275	1,100,000	380,000	110,000	582,000	210,000	1,425,000
GDP per capita (€)	23,000	23,000	30,000	23,250	21,900	22,250	11,000	25,825	18,900	24,000	10,900	14,545	23,600	24,000
RTD & INNOVATION														
Annual RTD Expenditure (€M)	3,000	4,420	3,500	4,108	30,300	48,200	627	1,428	12,123	7,700	622	5,500	9,150	27,700
Total RTD as % of GDP	1.5%	1.8%	2.2%	3.1%	2.2%	2.4%	0.5%	1.4%	1.0%	1.9%	0.5%	0.9%	3.7%	1.8%
Public RTD Expenditure (€M)	1,330	1,242	1,253	1,290	11,200	15,200	467	362	5,595	3,489	461	2,600	2,121	9,133
Business RTD Expenditure (€M)	1,721	3,179	2,224	2,818	19,100	33,000	160	1,066	6,528	4,211	161	2,800	7,029	18,575
High-Tech Patents per 1M population														
European Patent Office	9.8	17.6	21.5	80.4	20.2	29.3	0.5	13.3	4.8	35.8	0.4	2.5	22.9	18.9
US Patent Office	5.6	12.8	17.3	35.9	13.3	14.4	0.5	3.8	4.2	19.6	0.1	1.0	29.5	14.4
Language Technology R&D														
Number of HLT Research Centres	7	18	5	21	25	85	9	10	17	27	7	30	11	19
Number of Active HLT Suppliers	4	20	7	14	30	63	15	3	26	15	1	15	9	33
ICT INFRASTRUCTURE														
Number per 100 inhabitants:														
PCs	7.7	34.5	43.2	39.6	30.5	33.6	7.1	36.5	20.9	39.5	10.5	14.3	50.7	33.5
Internet Users	32.9	26.2	29.6	38.5	18.3	31.3	12.3	25	19.1	42.5	30.2	17.5	50.6	55.4
Mobile telephone subscriptions	78.5	54.9	61.0	72.6	49.4	58.6	55.9	66.8	73.7	67.1	66.5	60.9	71.4	72.7
Telephone lines	47.4	49.9	75.3	54.7	58.0	60.1	53.2	42.6	47.4	60.7	43.1	42.1	68.2	56.7
Computers with an Internet connection	21.3%	8.6%	14.5%	25.8%	6.3%	7.4%	14.8%	8.1%	12.8%	25.8%	5.9%	7.9%	13.2%	8.3%
ICT Spending														
ICT Expenditure as % of GDP	5.9%	5.8%	6.2%	6.4%	6.2%	5.7%	6.1%	5.4%	5.5%	6.9%	7.0%	6.8%	8.3%	7.4%
Per capita ICT expenditure	1,482	1,381	€2,000	1,534	1,446	1,400	691	1,290	1,065	1,657	747	973	2,060	1,709

Conclusions & Recommendations

Conclusions: the state of HLT in Europe

HLT research and development

For obvious reasons, HLT research has historically evolved with a national R&D bias towards the native language(s) of the national research communities. While this was essential in early HLT research, it is increasingly common to find a multi-language focus, especially in the more successful research departments and labs. This is a healthy development, and should help overcome inappropriate biases about 'ownership' of HLT for a particular language. As the HLT research community in Europe becomes ever more integrated, language expertise migrates across the whole of the EU, while naturally retaining its roots in national language communities. It is essential that language technology expertise and linguistic expertise be free to migrate and integrate across the EU research community.

HLT research and development is a long, complex process that needs substantial public support. The necessary training, resource, tool and technology development cannot be assured by market forces alone. Europe's success in the HLT field has been built on public funding, in the universities, national research institutes, and in funded projects. It is unlikely that the field can advance effectively - especially to bring all languages to the same level of sophistication, and incorporating the new languages of the expanded Union, without continued public investment on a significant scale.

Consistent and long-term funding of HLT research at the national level has paid off handsomely, and has contributed significantly to the strong national research base in Germany, France and the UK. It is unlikely, however, that all Member States, especially in the expanded Union, will be able to support programmes at the level of the more technologically

advanced members (including the Netherlands and Finland, as well as other Nordic countries). Consequently, the structure of EU funding will need to accommodate variations in the level of national support.

While national programmes in key Member States have been crucial in building core capabilities in HLT (as a complement to EU programmes), they have by no means been 'one size fits all'. National approaches to HLT research have mirrored local priorities and structures. In Germany, for example, large comprehensive programmes with a single focus (e.g. Verbmobil) linked industry to the research community in a very structured way. In France, HLT research was closely linked to (then) national laboratories (e.g. France Telecom). In the UK a relatively early Speech and Language Technology Programme solidified a strong network of national researchers, kick-starting market transfer at around the beginning of deregulation of the telecoms industry. This suggests that a truly 'European' approach to future HLT research will need to be adaptable, variable, and able to adjust to the different environmental conditions of Member States.

While research activities funded under the HLT-specific actions of the EU Framework Research Programmes are relatively visible both inside and outside the research community, the resulting picture is nevertheless incomplete. For example, there is as yet no coherent, transparent view of the considerable language-related R&D in other areas of IST research (for example in the area of Digital Libraries at ERCIM, or of Fraud Prevention at the JRC), nor of the important if structurally quite varied national programmes (for example in France, Italy, Lithuania and Estonia), nor commercially funded research (especially in the field of in-vehicle speech technology applications in Sweden's Telematics Valley, for example, or in controlled language applications in the aeronautics and vehicle documentation sectors). This lack of a coherent and comprehensive overview is likely to become even more extreme as the Sixth Framework Programme gradually implements a policy of

embedding previously 'stand-alone' HLT activities into its more mainstream IST research. Without a clear map with which to identify the patterns of on-going R&D actions, there is a risk not only of unnecessary duplication of effort, but also of making it harder for the investment community to contribute effectively to the process of transferring technology to the marketplace.

The research status of the languages of Europe is highly variable. A few languages (English, German, and French) are well served, enabling the emergence of more advanced research topics and market applications. Some of the less 'dense' languages (measured in numbers of speakers) are not even fully enabled for full exploitation of first-generation HLT applications. This means that there must be further public investment to bring all languages to a relatively equal status, at a baseline level, since this is an absolute prerequisite for future development of advanced ISTs capable of serving all European citizens equally.

At the same time, HLT research has, for several years, been moving steadily toward 'engineering' and away from theoretical research. Even apparent theoretical shifts (e.g. statistical and data-driven, as opposed to rule-based, NLP) are more like natural hybrids than true changes of paradigm. While this is a natural cycle, it is likely that the field will be refreshed by substantial re-thinking of its basic assumptions; accommodating this level of basic theoretical work (as opposed to back-filling baseline R&D for 'new' EU languages) should therefore be on the agenda for next-generation HLT research. It seems quite plausible that new theoretical approaches will arise from cross-fertilisation with other technical computing and engineering disciplines, which further emphasises the benefits of incorporating advanced HLT components into the mainstream FP6 research agenda.

But to ensure that Europe does not evolve into a two-speed culture for language technology, with one well-funded half of the HLT R&D agenda focused on embedding advanced systems for just a few of the more 'strategic' languages, while the other half attempts to ensure baseline coverage for the lower density or less 'strategic' languages, it is imperative that there be some form of autonomous 'language technology agency' whose task would be to sustain an appropriate degree of autonomy for the HLT field (especially in the critical area of baseline language components and resources), independently of HLT's ultimate technological destiny of becoming an embedded component of the information society infrastructure.

Market transfer

So far, there has been no direct link between robustness of the HLT research effort in any particular language community, and actual effectiveness of transfer to market. There is of course a clear split between examples of successful language technology transfer for high-density languages (especially English, German and French), and transfer for low-density languages, which is clearly due to the commercial potential of larger markets where high-density languages are spoken. There is however a notable exception to this in European Spanish, where the research effort is still quite diffuse, partly because of national support for a number of 'regional' languages, all of which have official status.

Another special case is Italian, globally less commonly spoken than Spanish, but high-density in Europe, but which is comparatively weak in HLT transfer, no doubt due to specific conditions in the business environment and technical infrastructure. Italy has a long and powerful tradition of HLT research, going right back to the beginnings of computational linguistics in the 1950s, and it is clear that its current position is due more to commercial 'timing' than to any inherent technology weakness.

By contrast, the relatively strong research community in both Finland and the Netherlands, where the business environment and infrastructure are among the strongest in Europe, has nevertheless transferred less technology, especially higher-end tools and products, than might have been expected. The conclusion

is that transfer of HLT to market is influenced by three strong factors: size of the linguistic community, business environment & infrastructure, and sharpness of research focus. Since the 'cost' of technologising a language is ultimately the same whether it is spoken by a population of 2 million or 200 million, it is now becoming clear that there need be no *necessary* link between language-specific research, technology development and market-transfer activities and the specific geographies where a language is spoken.

One effect of a truly European-wide, as opposed to country-based, marketplace for research and development, for example, may well be to encourage the creation of centres of best practice in language technology development, so what turn out to be the 'best' HLT architectures are chosen as the optimum development environments for any 'national' language, wherever it may be spoken and written.

What is clearly needed in a truly interactive European information society is language parity at all levels, both *inwards* and *outwards*, to use an analogy from investment. Until now, there has been a natural yet constraining tendency to develop language technologies for transfer of information *into* the national language. In practice, of course, I need access in my language to content and interaction in *your* language, as much as you need *my* content and interaction accessible in *your* language. Enabling such language parity at a technology level and ensuring its commercialisation will only be achieved by setting up a comprehensive multi-language infrastructure, which to a far greater degree than is true today would delink language processing from R&D geographies. Achieving such parity would be a critical item on the agenda of any eventual 'European language technology agency'.

Policy priorities
It is widely acknowledged that if Europe is to become a leading knowledge- and technology-based economy, fulfilling the objectives of eEurope, the momen-tum of innovation and R&D progress must be increased. HLT should continue to be promoted as a key technology advantage for Europe.

Market opportunities for HLT correlate strongly with 'environmental' opportunities such as the state of the ICT infrastructure, the strength of existing local or national ICT markets, readiness to accept new products and services by consumers and businesses, and the availability of channels to market (products as well as services) in which HLT capabilities must be embedded. The close correlation between market opportunities and strong HLT research suggests that in this, as in other areas of IST research, the business environment and technical infrastructure cannot be ignored if Europe is fully to exploit its potential in this important technological domain.

The 'information highway' analogy is a powerful one for HLT research, since language-enabling will literally eliminate barriers to communication across the networks of the Union, permitting the free flow of information, and the services and facilities based on information. From this perspective, the HLT infrastructure should have the same status and priority as the physical infrastructure that permits the free flow of goods and people in the Union.

Recommendations for future support of HLT

HLT in the ERA
Establish a concrete and visible presence for HLT activities in the European Research Area: The goal should be to have a set of robust, rich, stable, multilingual, 'autonomic' HLT modules, capable of being embedded into emerging IST operating environments. This is most likely to be achieved if HLT research is both a priority within the IST components of the ERA (cf. the contribution that research into interfaces, cognitive processes, interaction, knowledge technologies and semantics will make to IST research in FP6) and treated as a companion

technology for an innovative research agenda. A baseline objective should be to have such modules for all present and future EU languages, and to assure that components and resources for low-density languages are back-filled as a matter of priority.

Structures for Visioneering in HLT

Establish a 'language technology agency' to collectively supervise the gradual transition from the national HLT efforts to a truly European technology level of language parity, and to federate and rapidly circulate best practices at all levels of HLT R&D. A plausible first stage would be to create a LangTech Observatory for HLT, which would bring several advantages to the European research community:

- Research tracking to reduce or eliminate duplication of effort, provide more open access to exploitation opportunities, provide guidance in setting the HLT research agenda.
- Promoting the inclusion of HLT in all relevant European research efforts by making the field more visible to researchers in other domains.
- Shifting the focus from purely geographically defined national data, to a more 'language-oriented' observatory function, by transferring European best practices to the national level.
- Providing reliable data for EU innovation tracking, at policy level.

HLT Infrastructural Funds

The equivalent of 'linguistic infrastructural funds' would be an appropriate investment to support languages that lack a strong core of components and resources, and are lagging in the move to next-generation embedded HLT applications.

Research planners should consider disengaging 'language' from 'geography', and support linguistic infrastructure research and development wherever it is most likely to succeed. This should involve cross-border collaborations between strong HLT research locations, and geographical locations where less-technologically-developed languages are spoken (especially in New Accession Countries). This would

also benefit language communities with relatively strong HLT research, but weaker local opportunities for exploitation.

Digital Language Infrastructure

While language 'ownership' should ultimately be operationally disengaged from geography as a matter of funding principle, this will clearly take some time. Meanwhile, there will still be a major role for national HLT 'agencies' or sponsors, such as the 'digital language infrastructure' being developed by Nederlandse Taalunie for Dutch-Flemish (in joint programmes between Netherlands and Flanders), and in similar initiatives under the French Technolangue programme. This could form one mechanism to assure that all European languages are adequately supplied with core resources and components - or at least that missing elements are identified.

The role of the EC (via the proposed 'language technology agency') should be to support the collective definition of what constitutes a core 'language kit' without which HLT development cannot advance, promote the development of open-source platforms for developing and implementing such kits, as well as initiating the process of setting standards for interoperability between language components, and between language components and application environments. This would include defining and agreeing on requirements of formal and content quality, availability (free of ownership rights or under certain conditions), multi-functionality and reusability.

During an initial phase, the Taalunie experience should be considered as a model that can be expanded to all European languages, initiating a process that could result in a pan-European network of structures to sponsor HLT for specific languages, with concrete benefits for technology transfer. Taalunie has estimated the cost of the agency to be in the range of €500,000 per year for Dutch-Flemish. At this order of magnitude, the Union could fund ongoing support for core HLT 'language kits' for 20 languages at a
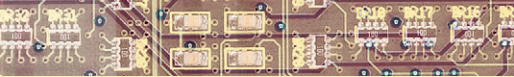
cost of €10M/year - a relatively modest sum in relation to current spending on language services in Europe, in what would be in effect a market-priming programme. The goal should be an 'open source' approach to the evolution of a digital language infrastructure for Europe; this could converge with other open-source software initiatives within the e-government agenda. It could also have a significant impact on near-term development and launch of HLT-based products and services in a much larger set of European languages than currently exists.

All of these infrastructural measures would be supervised by the proposed 'language technology agency', whose justification, status and composition would be subject to the broadest possible consultation. This would enable Europe's fundamental language technology agenda to gain progressive independence from the specific foci of Framework Programmes as such, and achieve continuity of action and impact over and above the specifically project-based approach favoured until now.

Abbreviations

ERA	European Research Area
ERCIM	European Research Consortium for Informatics and Mathematics
FP6	The Sixth EU Framework Programme for Research and Technological Development
HLT	Human Language Technologies
ICT	Information and Communication Technologies
IST	Information Society Technologies, part of FP5 and FP6
JRC	Joint Research Centre
NAC	New Accession Countries
R&D	Research and Development

Sources and Methodology

Research sources

The following publications and studies were used as sources of data for the HLT benchmarking study.

Cyberatlas, Jupitermedia Corporation, 2002, http://cyberatlas.internet.com/

European Information Technology Observatory 2001 (EITO), published by European Economic Interest Grouping (EEIG)

The European Innovation Scoreboard (EIS), published by DG Enterprise, http://trendchart.cordis.lu/Scoreboard2002/index.html

Eurostat Yearbook 2002, OECD

Flash Eurobarometer 125, "Internet and the Public at Large", Gallup Europe, May, 2002.

The Global Competitiveness Report 2001-2002, World Economic Forum, published by the Center for International Development, Harvard University, and Oxford University Press.

The Global Information Technology Report: Readiness for the Networked World, World Economic Forum, published by the Center for International Development, Harvard University, and Oxford University Press.

Global Internet Statistics by Language, Global Reach, 2002, http://www.glreach.com/globstats/

Standard Eurobarometer 55, "Analysis of public opinion towards the European Union", published on http://europa.eu

Opportunity Snapshot - sources

Economy & Society

Measure	Source of data
Total Population (M)	Economics Intelligence Unit and national sources, published in Global Competitiveness Report
% of citizens who can speak a language in addition to mother tongue	Eurobarometer
Number of Internet users by language	Global Reach
Total GDP	Eurostat
GDP per capita	Eurostat

RTD & Innovation

Measure	Source of data
Annual RTD Expenditure	Eurostat. Innovation Scorecard
Total RTD as % of GDP	Eurostat. Innovation Scorecard
Public RTD Expenditure	Eurostat. Innovation Scorecard
Business RTD Expenditure	Eurostat. Innovation Scorecard
European Patents	Innovation Scorecard
US Patents	Innovation Scorecard

ICT Infrastructure

Measure	Source of data
Number per 100 inhabitants:	
PCs	ITU, published in Global Competitiveness Report
Internet Users	Cyberatlas
Mobile telephone subscriptions	ITU, published in Global IT Report
Telephone lines	ITU, published in Global Competitiveness Report
Computers with an Internet connection	ITU, published in Global IT Report
ICT Spending	
ICT Expenditure as % of GDP	EITO
Per capita ICT expenditure	EITO

Language Technology R&D

Number of HLT Research Centres and Number of Active HLT Suppliers: EUROMAP fieldwork.

HLT Scorecard: methodology overview

The HLT Scorecard was calculated using a combination of data from the sources referenced above, and fieldwork from the EUROMAP project. The scorecard is made up of a set of indexes for factors that promote effective results in the HLT domain: HLT research maturity, breadth of language coverage, maturity of the general R& D environment, good access to market channels, ease of new business formation, a favourable environment for high-tech start-ups, trade competitiveness, an advanced ICT infrastructure, and a national market capable of absorbing the kinds of products and services that rely on HLT (what we have called "innovation potential").

The indexes were all normalised to a standard scale, to make it possible to integrate and compare information from different sources. Survey data from the Global Competitiveness Report is reported on a 1-to-7 scale (where 1 is judged to be the most negative or least mature, and 7 is the most positive or most

mature ranking), and we elected to use this scale as a standard way to compare indexes between Member States. Data points used to calculate a score for each index are shown below. Factors were given different weights, based on our assessment of their potential impact on HLT success:

Factor	Weight
HLT Research	4
Language Breadth	2
R&D Environment	2
New Business Formation	2
Access to Channels	2
Supply-side Readiness	2
Trade Competitiveness	1
ICT Infrastructure	4
Innovation Potential	2

The four components of the HLT Research index (i.e. research maturity for speech and text, strength of investment by public and private sector) were all given equal weight within that factor. The components of Language Breadth were given different weights within the factor: multilinguality was weighted double that of "minority/regional" language focus; this enables the index to capture strengths in minority language research (which we consider an important issue) without giving it undue weight in the overall score, since multilinguality is the key measure for commercial success. In turn, the Language Breadth factor was given half the weight of the research score, as illustrated in the table above.

HLT Scorecard: components and sources

HLT Benchmark factors

HLT research maturity
These indexes are based on the authors' assessment (using EUROMAP fieldwork) of four aspects of the HLT research scene in each Member State, and like all factors in the study uses a rating on the scale of 1-7.

We asked the following questions. HLT RTD: How significant is HLT research? How much of it goes on? Is it current? Is it increasing? Is it generally successful in creating new intellectual property? Public-sector investment in HLT RTD: How strong is the commitment of the public sector in funding and generally supporting HLT research? Do publicly-funded programmes focussed on language and HLT exist? Are they current? Have they been sustained over time? Is HLT included in programmes that support ICT generally? Is there strong support for HLT and computational linguistics in the university system? Is there strong support for HLT R&D in not-for-profit institutes? Private-sector HLT R&D: How strong is the commitment of industry to HLT R&D? Are there examples of significant commercial or industrial research centres active in the country? Does industry undertake development of the HLT technologies they own, or partner with local research institutes for development, rather than sourcing technology or development from abroad? Answers were based on a wide range of qualitative and quantitative data and information. EUROMAP assigned the following scores to Member States on the four HLT research indexes.

	HLT R&D & Technology Transfer		HLT R&D Investment	
	Speech	Text/ NLP	Public Sector	Private Sector
EU Average	4.6	4.7	4.2	4.0
Austria	4.0	3.0	3.5	4.0
Belgium	4.5	4.5	4.0	5.0
Denmark	3.8	4.5	4.5	3.0
Finland	4.5	5.0	4.0	5.0
France	5.5	5.5	5.0	4.0
Germany	6.0	6.0	6.5	5.5
Greece	4.0	4.5	4.5	3.0
Ireland	3.5	4.0	4.0	4.0
Italy	5.0	5.0	3.5	4.0
Netherlands	4.5	5.0	5.0	5.0
Portugal	3.5	3.5	2.0	2.0
Spain	4.5	5.0	4.0	3.0
Sweden	5.0	4.5	4.0	3.0
UK	5.5	6.0	4.5	5.0

Language Breadth

These indexes measure how the HLT community addresses the complex and often "political" decisions about what languages are the subject of R&D. These factors are important for several reasons. Multilinguality (and/or cross-linguality) is a significant market differentiator and driver in the HLT field, as well as being a core capability for the European agenda. When non-European languages are included in the agenda, multilinguality unlocks potential in global markets. Attention to minority and/or regional languages (which may encompass a "multi-language focus") can make a significant contribution to accessibility for small language communities and immigrants. We asked the following questions. Multilingual Focus: Is multilinguality (research in multiple languages, or research addressing cross-language issues) perceived as an important aspect of HLT R&D? Does HLT research address languages other than the national language(s)? Are non-national languages included in the research agenda? Do HLT researchers have strong links with individuals or programmes in other countries that broaden the linguistic coverage of the national programmes? Are exchanges - of human and technical resources - with specialists in languages other than the national language(s) common? Work in minority and/or regional languages: Are "minority" languages included in the HLT research agenda? Do smaller language communities (native and/or immigrant) participate in HLT R&D focussing on their languages? Are the languages of neighbouring or immigrant communities considered relevant to the research agenda? EURO-MAP assigned the following scores to Member States on HLT language breadth indexes.

	Multilingual Focus	Minority/Regional Languages
EU Average	4.1	2.4
Austria	3.0	1.0
Belgium	5.0	3.0
Denmark	4.5	4.0
Finland	5.0	3.0
France	5.0	2.0
Germany	4.0	2.0
Greece	3.0	1.0
Ireland	3.0	2.0
Italy	3.5	2.0
Netherlands	6.0	3.0
Portugal	3.0	1.0
Spain	4.0	5.0
Sweden	4.0	2.0
UK	4.0	3.0

Opportunity factors

R&D Environment
World Economic Forum survey questions

Technological Sophistication	Your country's position in technology (1=generally lags behind most countries, 7= is among the world's leaders)
FDI and Technology Transfer	Foreign direct investment in your country (1=brings little new technology, 7=is an important source of new technology)
Quality of Scientific Research Institutions	Scientific research institutions in your country, such as university and government laboratories, are (1=non-existent, 7=the best in their fields)
Company Spending on Research and Development	Companies' spending on research and development in your country (1=is non-existent, 7=is heavy relative to international peers)
Subsidies for Firm-Level Research and Development	Direct government subsidies for firms conducting research and development in your country (1=never occur, 7=are widespread and large)
Tax Credits for Firm-Level Research and Development	Government tax credits for firms conducting research and development in your country (1=never occur, 7=are widespread and large)
University/Industry Research Collaboration	In its R&D activity, business collaboration with local universities is (1=minimal or non-existent, 7=intensive and ongoing)
Availability of Scientists and Engineers	Scientists and engineers in your country are (1=non-existent or rare, 7=widely available)
Brain Drain	Scientists and engineers in your country (1=normally leave to pursue opportunities elsewhere, 7=almost always remain in the country)

Data Measures - Innovation Scorecard

Public R&D investment as a % of GDP
Private R&D investment as a % of GDP
European patents per 1M population
US patents per 1M population

New Business Formation
World Economic Forum survey questions

Administrative Burden for Start-Ups	Starting a new business in your country is generally (1=extremely difficult and time consuming, 7=easy)
State of Cluster Development	How common are clusters in your country? (1=clusters are limited and shallow, 7=clusters are common and deep)
Permits to Start a Firm	Approximately how many permits would you need to start a new firm? (median response listed for each country)
Days to Start a Firm	Considering license and permit requirements, what is the typical number of days required to start a new firm in your country? (median response listed for each country)

Access to Key Channels

This measures the strength of the opportunity for HLT transfer through channel players (such as telecoms companies, manufacturers, service companies, etc.) that are based in the country, and have exhibited an appetite for HLT take-up. We asked the following questions for each Member State. Are there significant (actual or potential) channel players located in the country? Are national telecommunications services open to and capable of adoption of HLT technologies? Are there industry clusters with good HLT take-up potential? Is there evidence that channel players have incorporated, or intend to incorporate, HLT into their products and services? EUROMAP assigned the following scores to Member States on access to key channels for sale/distribution of HLT.

	Access to Key Channels
EU Average	4.7
Austria	4.0
Belgium	4.0
Denmark	3.0
Finland	5.0
France	5.0
Germany	6.5
Greece	2.0
Ireland	5.0
Italy	6.0
Netherlands	5.5
Portugal	2.0
Spain	5.5
Sweden	6.0
UK	6.5

Supply-side Readiness
World Economic Forum survey questions

Venture Capital Availability	Entrepreneurs with innovative but risky projects can generally find venture capital in your country (1=not true, 7=true)
Access to Foreign Capital Markets	Citizens of your country who wish to invest in stocks and bonds and open bank accounts in other country (1=are prohibited from doing so, 7=are free to do so)
Foreign Access to Local Capital Markets	Foreign investors (1=are prohibited from investing in stocks and bonds in your country, 7=are free to invest in stocks and bonds)
Financial Regulation and Supervision	Regulations and supervision of financial institutions are (1=inadequate for financial stability, 7=among the world's most stringent)
Access to Bond Markets	Your company could borrow on the international bond market if necessary (1=not true, 7=true)
Local Equity Market Access	Raising money by issuing shares on the local stock market is (1=nearly impossible, 7=quite possible for a good company)
Sources of Investment Finance	When financing investments, your company typically (1=relies on its own retained earnings, 7=raises funds from banks or the bond markets)
Government Prioritization of ICT	Information and communications technologies are an overall government priority (1=strongly disagree, 7=strongly agree)
Government Success in ICT Promotion	Government programs promoting the use of ICT are (1=not very successful, 7=highly successful)
Laws Relating to ICT Use	Laws relating to electronic commerce, digital signatures, and consumer protection are (1=non-existent, 7=well-developed and enforced)
Legal Framework for ICT Development	The legal framework in your country supports the development of IT businesses (1=no, strongly impedes, 7=yes, significantly promotes)
Intellectual Property Protection	Intellectual property protection in your country is (1=weak or non-existent, 7=equal to the world's most stringent)
Presence of Demanding Regulatory Standards	Regulatory standards -- e.g. for products, energy, safety, environment -- in your country are (1=lax or non-existent, 7=among the world's most stringent)
Local Supplier Quantity	Local suppliers in your country are (1=largely non-existent, 7=numerous and include the most important materials, components, equipment and services)
Local Supplier Quality	Local suppliers in your country are (1=inefficient and have little technological capability, 7=internationally competitive and assist in new product and process development)

Data Measures - Innovation Scorecard

Venture capital invested as % of GDP
New capital available as % of GDP
% of new-to-market products
% of value-add in high-tech products

Trade Competitiveness
World Economic Forum survey questions

Hidden Trade Barriers	In your country, hidden import barriers other than published tariffs and quotas are (1=an important problem, 7=not an important problem)
Extent of Locally Based Competitors	Competition in the local market comes primarily from (1=imports, 7=local firms or local subsidiaries of multinationals)
Entry into Local Markets	Entry of new competitors (1=almost never occurs in the local market, 7=is common in the local market)
Control of International Distribution place	International distribution and marketing from your country (1=takes through foreign companies, 7=is owned and controlled by local companies)
Extent of Regional Sales	Exports from your country to surrounding regions are (1=limited, 7=substantial and growing)
Breadth of International Markets	Exporting companies from your country sell (1=primarily in a few foreign markets, 7= in virtually all international markets)
Nature of Competitive Advantage	Competitive advantage of your nation's companies in international markets is due to (1=low cost labor or natural resources, 7=unique products and processes)
Value Chain Presence	Exporting companies in your country (1=are involved primarily in production, 7=conduct not just in production but also product development, distribution and marketing)

ICT Infrastructure

World Economic Forum survey questions

Telephone/Fax Infrastructure Quality	New telephone lines for your business are (1=scarce and difficult to obtain, 7=widely available and highly reliable)
Speed and Cost of Internet Access	Lease-line or dial-up access to the Internet in your country is (1=slow and expensive, 7=as fast and cheap as anywhere in the world)
Public Access to Internet	Public access to the Internet through libraries, post offices etc. is (1=very limited, 7=pervasive -- most people have frequent access)
Internet Access in Schools	Internet access in schools is (1=very limited, 7=pervasive -- most children have frequent access)
Quality of Competition in Telecommunication Sector	Is competition in your country's telecommunications sector sufficient to ensure high quality, infrequent interruptions and low prices? (1=no, 7=yes, equal to world's best)
Quality of Competition in ISP Sector	Is competition among your country's Internet Service Providers sufficient to ensure high quality, infrequent interruptions and low prices? (1=no, 7=yes, equal to world's best)
Local Availability of Information Technology Services	In your industry, specialized IT services are (1=not available in the country, 7=available from world-class local institutions)

Data Measures - EITO

Per-capita ICT expenditure
ICT % of GDP
PCs per White Collar Worker
Telephones lines per population
Mobile users per population
% of households with PCs

Data Measures - Eurobarometer Survey
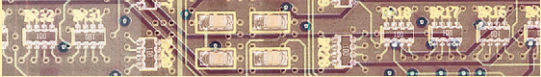
Availability of the Internet in homes	% of households with Internet access
High-speed Internet access in homes	% of households with high-speed Internet access
Internet Access through Mobile Phone	% of people who access the Internet using a mobile phone
Frequency of Internet use at home	% of home users who access the Internet every day
Use of electronic signatures	% of home users who use electronic signatures
Internet purchasing from home	% of home users who frequently or occasionally buy products or services over the Internet
International Internet purchasing from home	% of home users who have purchased goods or services from Websites located in other countries
Contact with public administrations through the Internet	% of home users who have at some time contacted a public administration through the Internet
Companies with Internet access	% of companies with more than 10 employees which have Internet connection
Companies with Web sites	% of companies with more than 10 employees which have a Web site

Innovation Potential
World Economic Forum survey questions

High Skilled IT Job Market	Highly skilled information technology workers in your industry (1=must leave the country to find good jobs, 7=have their pick of well-paid, desirable jobs within the country)
IT Training and Education	Your country's IT training and educational programs (1=lag far behind most countries, 7=are among the world's best)
Extent of Staff Training	In your country, companies' general approach to human resources is to invest (1=little in training and development, 7=heavily to attract, train and retain staff)
Quality of Management Schools	Management schools in your country are (1=limited and of poor quality, 7=among the world's best)
Firm-Level Innovation	In your business, continuous innovation plays a major role in generating revenue (1=not true, 7=true)
Firm-Level Technology Absorption	Companies in your country are (1=not interested in absorbing new technology, 7=aggressive in absorbing new technology)
Extent of Product and Process Collaboration	Product and process development in your country is conducted (1=within companies or with foreign suppliers, 7=in collaboration with local suppliers, customers & research institutions)
Local Availability of Specialized Research and Training Services	In your industry, specialized research and training services are (1=not available in the country, 7=available from world-class local institutions)
Capacity for Innovation	Companies obtain technology (1=exclusively from foreign companies, 7=by pioneering their own new products or processes)
Internet Effects on Business	To what extent has the Internet improved your firm's ability to coordinate with customers and suppliers to reduce inventory costs (1=no change, 7=huge improvement)
Decentralization of Corporate Activity	Corporate activity in your country is (1=dominated by a few business groups, 7=spread among many firms)
Government Procurement of Advanced Technology Products	Government decisions on the procurement of advanced technology products are based on (1=price alone, 7=technology and encouraging innovation)
Government On-line Services	On-line government services -- e.g. downloadable permit applications, tax payments -- in your country are (1=not available, 7=commonly available)
Buyer Sophistication	Buyers in your country are (1=unsophisticated and choose based on the lowest price, 7=knowledgeable and demanding and buy innovative products)
Degree of Customer Orientation	Firms in your country (1=generally treat their customers badly, 7=pay close attention to customer satisfaction)

Data Measures: Innovation Scorecard

Science & Engineering graduates
Levels of tertiary education
Levels of LifeLong Learning
Workforce employed in high-tech manufacturing
Workforce employed in high-tech services
SMEs with in-house innovation activities
SMEs innovating through co-operative programmes
Expenditure on innovation as % of sales

Data Measures - Eurobarometer survey

% of companies that can take orders on the Internet
% of company sales over the Internet
% of companies that purchases goods and/or services on the Internet
% of goods and/or services purchased on the Internet