www.meta-net.eu
office@meta-net.eu
Tel: +49 30 3949 1833
Fax: +49 30 3949 1810

META-NET
White Paper Series
Languages in the European Information Society – Sweden

Preface

This series of language white papers is for journalists, politicians, language communities, language teachers and others, who want to establish a truly multilingual Europe.

This series promotes knowledge about language technology (LT) and its potential. The coverage and use of language technology in Europe varies from language to language. Consequently, required actions to support research and development vary, and the necessary steps depend on many factors, such as the complexity of the language or the size of its community.

META-NET has faced this challenge by initiating an analysis of the current state of affairs for language resources and technologies. The analysis focuses on the 23 official European languages and several important regional languages. The results of the analysis suggests that there are many significant gaps for each language. Detailed expert analysis and assessment of the situation for each language will help maximise the impact of language technology and minimize any associated risks.

META-NET is a European Commission Network of Excellence that consists of 44 research centres from 31 countries. META-NET is working with stakeholders from many areas of society, industry and research to generate strategic visions and produce a strategic research agenda that shows how language technology applications can address any gaps by 2020.

Imprint

Authors/Editors of text common to all languages:

Dr. Aljoscha Burchardt, DFKI
Prof. Dr. Markus Egg, Humboldt-Universität zu Berlin
Kathrin Eichler, DFKI
Dr. Georg Rehm, DFKI
Prof. Dr. Manfred Stede, Universität Potsdam
Prof. Dr. Hans Uszkoreit, Universität des Saarlandes and DFKI

Authors/Editors of text regarding Swedish:

Lars Borin, University of Gothenburg
Martha D. Brandt, University of Gothenburg
Jens Edlund, KTH Royal Institute of Technology
Jonas Lindh, University of Gothenburg
Mikael Parkvall, Stockholm University

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME(grant agreement no.: 249119), CESAR (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893), and META-NORD (grant agreement no.: 270899).

Table of Contents

Executive Summary

Many European languages run the risk of becoming victims of the digital age because they are underrepresented and under-resourced online. Huge regional market opportunities remain untapped today because of language barriers. If we do not take action now, many European citizens will become socially and economically disadvantaged because they speak their native language.

Innovative, language technology (LT) is an intermediary that will enable European citizens to participate in an egalitarian, inclusive and economically successful knowledge and information society. Multilingual language technology will be a gateway for instantaneous, cheap and effortless communication and interaction across language boundaries.

Today, language services are primarily offered by commercial providers from the US. Google Translate, a free service, is just one example. The recent success of Watson, an IBM computer system that won an episode of the Jeopardy game show against human candidates, illustrates the immense potential of language technology. As Europeans, we have to ask ourselves several urgent questions:

- Should our communications and knowledge infrastructure be dependent upon monopolistic companies?
- Can we truly rely on language-related services that can be immediately switched off by others?
- Are we actively competing in the global market for research and development in language technology?
- Are third parties from other continents willing to address our translation problems and other issues that relate to European multilingualism?
- Can our European cultural background help shape the knowledge society by offering better, more secure, more precise, more innovative and more robust high-quality technology?

In terms of number of speakers, Swedish is among the top 2% of languages in the world, and enjoys a relatively secure position as the national and majority language of a Western European nation, the result of a continuous tradition of writing and standardisation going back to the middle ages. Modern information and communication technologies (ICT) have permeated all aspects of Swedish society: 85% of the population have a broadband connection and Swedish is among the best represented languages on the web. However, at present this general ICT maturity does not come with an equally impressive set of software applications for processing linguistic content in Swedish, or in the other languages used in the both de jure and de facto multilingual environment that characterizes modern Swedish society. Despite a history of research and development in Language Technology reaching back to the 1960s, Swedish still has a long way to go before the language will be as richly endowed in this respect as English.

META-NET contributes to building a strong, multilingual European digital information space. By realising this goal, a multicultural union of nations can prosper and become a role model for peaceful and egalitarian international cooperation. If this goal cannot be achieved, Europe will have to choose between sacrificing its cultural identities or suffering economic defeat.

A Risk for our Languages and a Challenge for Language Technology

We are currently witnessing a digital revolution that is comparable to Gutenberg's invention of modern printing.

As recent events in North Africa illustrate, we are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in digitised and network communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

After Gutenberg's invention, real breakthroughs in communication and knowledge exchange were accomplished by efforts like Luther's translation of the Bible into common language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled an exchange across languages;
- the creation of journalistic and bibliographic guidelines assured the quality and availability of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology helped to automate and facilitate many of the processes:

- desktop publishing software replaces typewriting and typesetting;
- Microsoft PowerPoint replaces overhead projector transparencies;
- e-mail sends and receives documents often faster than with a fax machine;
- Skype makes Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- search engines provide keyword-based access to web pages;
- online services like Google Translate produce quick and approximate translations;
- social media platforms facilitate collaboration and information sharing.

Although such tools and applications are helpful, can they sufficiently implement a sustainable, multilingual European information society, a modern and inclusive society where information and goods can flow freely?

Language Borders Hinder the European Information Society

We cannot precisely know what the future information society will look like. When it comes to discussing a common European energy strategy or foreign policy, we might want to listen to European foreign minsters speak in their native language. We might want a platform where people, who speak many

different languages and who have varying language proficiency, can discuss a particular subject while technology automatically gathers their opinions and generates brief summaries. We also might want to speak with a health insurance help desk that is located in a foreign country.

A global economy and information space confronts us with more languages, speakers and content.

It is clear that communicative needs have a different quality as compared to a few years ago. In a global economy and information space, more languages, speakers and content confront us and require us to quickly interact with new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter and YouTube) is only the tip of the iceberg.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognize that it is in a language we do not understand. According to a recent report requested by the European Commission, 57% of Internet users in Europe purchase goods and services in languages that are not their native language. (English is the most common foreign language followed by French, German and Spanish.) 55% of user read content in a foreign language while only 35% use another language to write e-mails or post comments on the web.[1] A few years ago, English might have been the lingua franca of the web—the vast majority of content on the web was in English. The situation has now changed drastically. The amount of online content in other languages (particularly Asian and Arabic languages) has exploded.

Which European languages will thrive and persist in the networked information and knowledge society?

An ubiquitous digital divide that is caused by language borders has surprisingly not gained much attention in the public discourse; yet, it raises a very pressing question, "Which European languages will thrive and persist in the networked information and knowledge society?"

Our Languages at Risk

The printing press contributed to an invaluable exchange of information in Europe, but it also lead to the extinction of many European languages. Regional and minority languages were rarely printed. As a result, many languages like Cornish or Dalmatian were often limited to oral forms of transmission, which limited their continued adoption, spread and use.

The wide variety of languages in Europe is one of its most important cultural assets and an essential part of Europe's success.

The approximately 60 languages of Europe are one of its richest and most important cultural assets. Europe's multitude of languages is also a vital part of its social success.[2] While popular languages like English or Spanish will certainly maintain their presence in the emerging digital society and market, many European languages could be cut off by digital communications and become irrelevant for the Internet society. Such developments would certainly be unwelcome. On one hand, a strategic opportunity would be lost that would weaken Europe's global standing. On the other hand, such developments would conflict with the goal of equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights,

[1] European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

[2] European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

such as political expression, education and participation in society.[3]

Language Technology is a Key Enabling Technology

In the past, investment efforts have focused on language education and translation. For example, According to some estimates, the European market for translation, interpretation, software localisation and website globalisation was € 8.4 billion in 2008 and was expected to grow by 10% per annum.[4] Yet, this existing capacity is not enough to satisfy current and future needs.

Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates across different languages.

Language technology is a key enabling technology that can protect and foster European languages. Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates regardless of language barriers or computer skills. Language technology already assists everyday tasks, such as writing e-mails, searching for information online or booking a flight. We benefit from language technology when we:
searching for and translating web pages,

- use the spelling and grammar checking features in a word processor;
- view product recommendations at an online shop;
- hear the verbal instructions of a synthetic voice in a navigation system;
- translate web pages with an online service.

The language technologies detailed in this paper are an essential part of innovative future applications. Language technology is typically an enabling technology within a larger application framework like a navigation system or a search engine. These white papers focus on the readiness of core technologies in the each language.

One can think of language technology as the operating system for the content and user interaction.

In the near future, we need language technology for all European languages that is available, affordable and tightly integrated within larger software environments. An interactive, multimedia and multilingual user experience is not possible without language technology.

Opportunities for Language Technology

Language technology can make automatic translation, content production, information processing and knowledge management possible for all European languages. Language technology can also further the development of intuitive language-based interfaces for household electronics, machinery, vehicles, computers and robots. Although many prototypes already exist, commercial and industrial applications are still in the early stages of development. The current rate of progress creates a genuine window of opportunity with research steadily progressing during the last few years. For example, machine translation (MT) already delivers a reasonable amount of accuracy within specific domains, and experimental applications provide multilingual information and knowledge management as well as content production in many European languages.

---

[3] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[4] European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

A Risk for our Languages and a Challenge for Language Technology

Language applications, voice-based user interfaces and dialogue systems are traditionally found in highly specialised domains, and they often exhibit limited performance. One active field of research is the use of language technology for rescue operations in disaster areas. In such high-risk environments, translation accuracy can be a matter of life or death. The same reasoning applies to the use of language technology in the health care industry. Intelligent robots with cross-lingual language capabilities have the potential to save lives.

There are huge market opportunities in the education and entertainment industries for the integration of language technologies in games, edutainment offerings, simulation environments or training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just a few more examples where language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a further need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology represents a tremendous opportunity for the European Union that makes both economic and cultural sense. Multilingualism in Europe has become the rule. European businesses, organisations and schools are also multinational and diverse. Citizens want to communicate across the language borders that still exist in the European Common Market. Language technology can help overcome such remaining barriers while supporting the free and open use of language. Furthermore, innovative, multilingual language technology for European can also help us communicate with our global partners and their multilingual communities. Language technologies support a wealth of international economic opportunities.

Multilingualism is the rule, not an exception.

Challenges Facing Language Technology

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. We cannot wait ten or twenty years for significant improvements to be made that can further communication and productivity in our multilingual environment.

The current pace of technological progress is too slow to arrive at substantial software products within the next ten to twenty years.

Language technologies with broad use, such as the grammar and spell checking features in word processors, are typically monolingual, and they are only available for a handful of languages. Applications for multilingual communication require a certain level of sophistication. Machine translation and online services like Google Translate or Bing Translator are excellent at creating a good approximation of a document's contents. But such online services and professional MT applications are fraught with various difficulties when highly accurate and complete translations are required. There are many well-known examples of funny sounding mistranslations, for example, literal translations of the names *Bush* or *Kohl*, that illustrate the challenges language technology must still face.

Language Acquisition

To illustrate how computers handle language and why language acquisition is

a very difficult task, we take a brief look at the way humans acquire first and second languages, and then we sketch how machine translation systems work—there's a reason why the field of language technology is closely linked to the field of artificial intelligence.

Humans acquire language skills in two different ways: learning examples and learning the underlying language rules.

Humans acquire language skills in two different ways. First, a baby learns its native language via examples. Exposure to concrete, linguistic specimens by language users, such as parents, siblings and other family members, helps babies from the age of about two or so produce their first words and short phrases. This is only possible because of a special genetic disposition humans have for learning their first language.

Learning a second language usually requires much more effort. At school age, foreign languages are usually acquired by learning their grammatical structure, vocabulary and orthography from books and educational materials that describe linguistic knowledge in terms of abstract rules, tables and example texts. Learning a foreign language takes a lot of time and effort, and it gets more difficult with age.

The two main types of language technology systems acquire language in a similar manner as humans.

The two main types of language technology systems acquire language capabilities in a similar manner as humans. Statistical approaches obtain linguistic knowledge from vast collections of concrete example texts in a single language or in so-called parallel texts that are available in two or more languages. Machine learning algorithms model some kind of language faculty that can derive patterns of how words, short phrases and complete sentences are correctly used in a single language or translated from one language to another. The sheer number of sentences that statistical approaches require is huge. Performance quality increases as the number of analyzed texts increases. It is not uncommon to train such systems on texts that comprise millions of sentences. This is one of the reasons why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, available online information, and translation services such as Google Search and Google Translate rely on a statistical (data-driven) approach.

Rule-based systems are the second major type of language technology. Experts from linguistics, computational linguistics and computer science encode grammatical analysis (translation rules) and compile vocabulary lists (lexicons). The establishment of a rule-based system is very time consuming and labour intensive. Rule-based systems also require highly specialised experts. Some of the leading rule-based machine translation systems have been under constant development for more than twenty years. The advantage of rule-based systems is that the experts can more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. Due to financial constraints, rule-based language technology is only feasible for major languages.

Swedish in the European Information Society

General Facts

According to the estimation of Parkvall (2009), the number of monolingual native speakers of Swedish[5] is about 85% of Sweden's population, which corresponds to approximately 7.7 million people. Of the remaining 15% of the population (approximately 1.35 million people), those who have grown up in Sweden can be assumed to have acquired Swedish as one of their native languages, whether as an addition to an immigrant language or to an indigenous minority tongue.

Additionally, a similar number (1.35 million) of Sweden's residents are born abroad, according to *Statistics Sweden*[6] in 2010. The foreign-born population includes adopted children, some individuals born abroad to Swedish parents, and members of Swedish-speaking ethnic groups in Finland, Estonia and the Ukraine (see further information regarding these ethnic groups below). Together, these ethnic groups total just over 100,000.

The following table shows the proportion of languages (mother tongue figures) of Sweden as of 2006 (Parkvall, 2009):

Official "majority language"	
Swedish	85.2%
Non-official indigenous language	
Swedish Sign Language	0.1%
Non-official indigenous languages ("dialects")	
Elfdalian	0.02%
Överkalixmål	0.02%
Official minority languages	
Finnish (incl. Meänkieli)	2.5%
Romani	0.1%
Saami	0.05%
Yiddish	0.01%
Non-official immigrant languages	
Serbian, Croatian and Bosnian	1.2%
Arabic	1.0%
Kurdish	0.7%
Spanish	0.7%
German	0.7%
Persian	0.6%
Norwegian	0.6%
Danish	0.6%
Polish	0.5%
Albanian	0.5%
English	0.5%
Aramaic	0.4%
Turkish	0.4%
Somali	0.3%
Hungarian	0.2%

5 i.e. have Swedish as their only mother tongue.

6 http://www.scb.se

Russian	0.2%
Thai	0,2%
Cantonese	0,1%
Greek	0,1%
Estonian	0.1%
Other immigrant languages	2.3%

Parkvall (2009) estimates about 185,000 native speakers of highly divergent Swedish dialects, of whom 5–10,000 use varieties divergent enough from the standard language to merit being considered languages in their own right.

In general, however, the regional differences in Sweden are moderately marked, and – as in most other industrialized countries – people born after the Second World War generally speak the standard with only phonological clues betraying their approximate geographical origin. Some lexical peculiarities can of course also be noticed, but the differences in morphology and syntax are, generally speaking, no longer more noticeable between different geographical areas than they are between generations. Swedish-speakers in Finland have in general followed the same path, although the local dialects are in somewhat better health there than they are in Sweden. However, east of the Baltic, words and constructions denoting concepts regarding modern society are frequently borrowed or calqued from Finnish.

The geographical differences that do exist are virtually exclusive to the spoken language, and for a newspaper text, it would be well-nigh impossible to determine the area in which it was produced, and even for a newspaper from Finland, this would be difficult, save for a small number of words and expressions denoting concepts relating specifically to Finnish society.

The number of daily newspapers in Sweden was 168 in 2008, according to Statistics Sweden, a number that seems reasonably stable despite falling circulation.[7] 26,182 "books and pamphlets" were published in Sweden in 2008, a number which increased consistently over the last decade. The total includes 86% original works and 14% translations. Interestingly, about one fourth of the original works were published in languages other than Swedish. However, only approximately 3% of these publications were in any of the indigenous minority languages or major immigrant languages. An overwhelming 22% of all original works published in Sweden in 2008 were in English.

Additionally, UNESCO's *Index translationum* database[8] features 31,474 translations into Swedish, and 31,358 with Swedish as the source language. Given that Statistics Sweden counts about 3,000 annual translations into Swedish in Sweden alone, it would seem that the two sources differ in scope. However, since 2005, the *Index translationum* does include about 2,500 cases yearly of Swedish as a target language of translations, which is compatible with the figures already cited.

According to Statistics Finland,[9] about 500 original Swedish-language titles

[7] The definition of a "daily" newspaper is one which is published at least three times a week.

[8] http://www.unesco.org/xtrans/

[9] http://www.stat.fi

are published yearly in Finland and about an additional 100 publications are translated into Swedish.

Among the 50 songs most frequently played on P3 (the public service radio music channel[10]) in 2010, 88% were performed in English (five songs were in Swedish and one in French; note that many of the English-language songs were sung by Swedish performers). In other popular music charts, however, Swedish tends to fare somewhat better.

As for television, 74% of the programs on the public service channel SVT were of domestic origin in 1999, which implies the use of Swedish or – more rarely – one of the national minority languages. In the commercial TV channels TV3, TV4 and TV5, this proportion was between 12% and 49% (Falk 2001:79). Again, a language other than Swedish almost invariably implies English, especially in the commercial channels.

In Finland, the national public broadcasting offers two radio channels in Swedish, and almost 20 hours of televised material, in addition to which a similar amount of Swedish TV programming is available exclusively on the web.[11]

At the cinemas, Swedish films were responsible for about one fourth of the attendance around the turn of the millennium (Falk 2001:85), with – again – the remainder being almost exclusively in English.

Particularities of the Swedish Language

In general, Swedish is a relatively normal representative of European languages in general, and Germanic languages in particular. The most 'exotic' aspects of the language are found in the domain of phonology, with notable features being

- a phonemic pitch accent system;
- presence of the cross-linguistically unusual phoneme /ɧ/;
- an unusually large vowel system, including front rounded vowels; and
- rather liberal phonotactics with CCC onsets, and CCCC codas, yielding half a million potential syllables.

Structurally, Swedish generally follows the patterns typical of Germanic languages, including V2 word order. More unusual traits that might deserve mention include negation placement before the tensed verb in subordinate clauses, and the presence of a "reflexive possessive" in the third person (i.e., a special possessive form used if and only if the possessor is co-referential with the subject).

In line with e.g. German, the language features plenty of compounding, which may yield rather long words. While any native speaker phonologically marks these as compounds, and while they are written as one word in the prescriptive tradition, many writers produce a blank in-between the constituent words, something that might be relevant for language technology purposes. A compound word such as *långhårig* 'long-haired' might thus be written *lång hårig*, which, in a more normative vein would be interpreted as 'tall (and) hairy'.

10 http://sverigesradio.se/sida/artikel.aspx?programid=3040&artikel=4262315

11 http://svenska.yle.fi

Recent Developments

Language Legislation

Language legislation in Sweden was virtually non-existent until 1999, when a law on minority languages was passed by the parliament. It promoted five languages (Finnish, Saami, Romani, Yiddish and Meänkieli [or Torne Valley Finnish]) to the status of "official minority languages". Simultaneously Sweden ratified the *European Charter on Regional or Minority Languages* for these languages. In practice, however, the concrete effects of thisese measures were limited, and seemingly cosmetic in nature.

After the passing of the minority-language bill, some people found it odd that the country only had minority languages, but not an official majority language. As is the case in countries such as Britain and the United States, the majority language was of course *de facto* official, but lacked *de jure* recognition. Therefore, a new language law became effective in 2009, which stipulated that Swedish is the "main language" (*huvudspråk*) of the country[12].

It is difficult to deny that the text of this law is rather vacuous. Loosely translated, it states the obvious fact that "Swedish is the main language of Sweden", and that "every inhabitant of Sweden should have access to it". Speakers of any language (the "main" one, the five "minority" ones, and any other language) should be allowed to "use and develop" their mother tongue. The authorities have a "special responsibility" for protecting Swedish, the minority languages and Swedish Sign Language.

The closest that the new law gets to regulating actual behaviour would seem to be Section 10, which states that the language of "courts, authorities, and other administrative bodies performing public services" should be Swedish. A couple of complaints have been filed against authorities since, by individuals and organizations who have observed what they perceive as an excessive use of English, complaints which have met with varying degrees of success. They usually deal with symbolic issues such as the email addresses of the government ministries, which used the English name of the ministry in question, rather than the Swedish one.

For a convenient overview (in French) of language legislation issues with regard to Sweden (and indeed any other country in the world), the Canadian site *L'aménagement linguistique dans le monde*[13] can be recommended, it being as accurate as one can reasonably expect from a work that aspires to cover the entire planet.

Language Cultivation in Sweden

As mentioned above, the Swedish language has until recently not had any official recognition whatsoever in Sweden, and while it has been recognised as such in Finland, authorities have in general not interfered with the development and makeup of the language as such.

---

[12] The full text can be found in *Svensk författningssamling*, No. 2009:600 (http://www.riksdagen.se/webbnav/index.aspx?nid=3911&bet=2009:600).

[13] http://www.tlfq.ulaval.ca/axl

Some official or semi-official bodies, such as *Klarspråksgruppen* (the governmental committee 'Clear Language Group'), the Swedish Academy and *Svenska Språknämnden* ('Swedish language board') have engaged in language cultivation, and are or were seen as having a normative mandate. In Finland, the *Forskningscentralen för de inhemska språken* ('The Research Institute for the Languages of Finland') fulfils a similar role. In 2006, the *Språkrådet* ('Language Council of Sweden'), was formed by the government, an organization billing itself as the "official language cultivation body of Sweden". Its mission is to "monitor the development of spoken and written Swedish and also to monitor the use and status of all other languages spoken in Sweden [and to] strengthen Nordic language unity". However, their homepage[14] explicitly states that "all other languages spoken in Sweden" refers only to Swedish, the five official minority languages and Swedish Sign Language.

There are also a number of private initiatives, which usually combat anglicisms and the use of English at the expense of Swedish, with the most vocal being *Språkförsvaret* ('The language defence'), which enjoys a relatively limited following and a moderate degree of public awareness.

Language in Education

Education in Sweden (and in Swedish-speaking parts of Finland) is generally in Swedish, but there is concern in some circles about English encroaching on Swedish. University-level education in English is not rare, and at some departments, most of the teaching is done in English, regardless of whether or not foreigners are present (Falk 2001:25, 29f). In 1999, 2–3% of the children attending public schools (primary and secondary levels) were taught in a language other than Swedish, which in three fourths of the cases meant English (Falk 2001:18f). This phenomenon appears not to have been investigated since, but Falk noted that the proportion was rising steadily. She also referred to studies (Falk 2001:19) demonstrating that these children were less proficient in Swedish than their Swedish-educated peers.

There also exist a limited number of schools using other languages (German, French, Finnish…) as their main medium of instruction. Specific classes using both Finnish and Swedish have existed, and to some extent still do, in public schools. The use of languages other than Swedish in public education has, however, generally been reduced to schools being obliged to offer mother tongue education outside of normal school hours, provided that it is required by a certain number of students. Here, the language does not have to be an officially recognised one, but can be any language, provided it is actively used in the home environment (though this proviso does not apply to the official minority languages).

In Finland, education in Swedish is offered from kindergarten to university level (in localities where there is a Swedish-speaking presence in the first place). The majority of the students are of course Swedish-speaking Finns, but some schools also have sizeable proportions of Finnish returnee migrants from Sweden, and sometimes also pupils with a purely Finnish background. In the latter case, the parents have taken the advantage of giving their children another language 'for free', but concerns have been expressed that the lack of

[14] http://www.sprakradet.se/international

prior knowledge among these children risks turning them into a 'Trojan horse', and that their presence might turn the classroom (or at least the school playground) into a Finnish-dominated language environment.

International Aspects

Outside Sweden, Swedish also enjoys official standing in Finland, whose statistic authorities claim 290,000 native speakers (about 5.5% of the nation's total population). Their number has been declining since the Second World War, and in terms of their proportion of the population in Finland, the Swedish Finns have been decreasing since the 17th century (when the percentage was about 16.5%).

While occasionally questioned, the status of Swedish in Finland is remarkably strong, given the small size of the minority (which, legally speaking, is not even considered a minority, but one of the two "domestic languages") and the relative lack of international currency of Swedish. All Finns are required to study Swedish, which of course does not guarantee that they leave school with any proficiency in it. Most in fact do not, but when questioned in a survey administered by the European Union,[15] 38% of those with Finnish as their mother tongue did claim capability of conversing in Swedish.

Indigenous Swedish-speaking communities[16] have also existed in four other (contemporary) countries: Russia (small enclaves in the Petersburg and Karelian areas, which were mainly offshoots of Finland's Swedish-speaking population), the United States (where the language of the 17th century colony of New Sweden survived until the early 1800s), Estonia and later the Ukraine. In Estonia, the vast majority of the Swedish-speaking population (present there since at least the 13th century) of about 8,000 fled to Sweden in the wake of the Second World War, and the remaining individuals are probably to be counted in dozens (at most) rather than hundreds or thousands. The Ukrainian group descended from Estonian Swedes deported in the late 18th century. Most immigrated to Sweden and North America in 1929, and only a handful of survivors remain today.

Apart from these groups, Swedish-speakers outside of Sweden and Finland consist of immigrants and temporary expatriates from these two countries. The number is likely to be around 300,000 (Parkvall 2010), mainly in the other Nordic countries, in western Europe, the United States, Canada and Australia. In none of these countries, however, do they represent more than a negligible proportion of the recipient countries' total population.

Looking at Swedish international relations w.r.t breaking through the communication barrier, we see that the vast majority of Swedish-speakers in Finland have a decent (and often impeccable) command of Finnish. For Sweden, EU statistics[17] indicate that about 90% of the Swedish population claim to be capable of conversing in English, 28% in German, and 10% in French. During the entire post-war era, English has been a compulsory school subject, and most school children have studied either German or French (but rarely both).

---

15 Eurobarometer 243

16 Arbitrarily defined as groups where the language survives more than three generational changes among a sizeable proportion.

17 Eurobarometer 237, 243

A recent survey[18] shows that Swedes are not only quantitatively more Anglophone than other nationalities, but that their English is also qualitatively impressive. Continuous media exposure is of course partly responsible for the high level of competence in English, but this does little to improve the knowledge of German or French. In 1994, Spanish was promoted to the same status in the school system as German and French, and it rapidly rose to become the most popular foreign language after English – mostly at the expense of German.

As of 2011, Sweden's foremost trading partner (according to *Statistics Sweden*[19]) is Germany, followed by (in order) Norway, Denmark, Britain, the Netherlands, Finland, the United States, France, Belgium, China and Russia.

Swedes travel extensively, but are not likely to use anything other than English on their trips abroad. Similarly, tourists travelling to Sweden will probably have a hard time being understood by Swedes if they use another language than English (or, of course, Swedish).

In short, the linguistic reality for the average Swedish native speaker in Sweden is such that only two languages co-exist: Swedish and English. The Swedes are proud of their knowledge of English – most of them do speak English and they speak it relatively well. Sweden is unusual, however, also because it relies to such an extent on one single *lingua franca*, where EU statistics[20] indicate that other Europeans are more likely to speak a variety of foreign languages. Indeed, respondents were asked whether they favoured (a) the current EU policy that every EU citizen should learn a language other than their mother tongue; and (b) whether they would favour a policy requiring the learning of *two* additional languages. The Swedes were resoundingly in favour of the first proposal, but were opposed to the second one to a higher degree than *any other* nationality.

Globally speaking, Swedish has a large number of native speakers (over 98% of the world's 6–7,000 languages have smaller native speaker communities). Additionally, its presence in public life is even larger than this number alone would suggest. It is very much a healthy language, with a secure position in Sweden (if not in Finland) in the short- to medium-term perspective. However, even though the only competition in the local linguistic ecology stems from English, it must not be ignored for it is not negligible - as can be seen from the already strong position of English in the daily lives of many Swedes, which continues to strengthen.

Swedish on the Internet

Swedish is conspicuous on the web, and in some surveys that have been carried out in this regard, it is consistently featured among the 15 or so best represented languages in the world (see, e.g., Parkvall 2006:63). At the time of writing, Swedish ranks as number 11 among the languages used on Wikipedia. In other similar measures of media presence (film industry, economic power, etc.), Swedish is typically among the top 20 among the world's 6,000

18 http://www.ef.se/epi/
19 http://www.scb.se
20 Eurobarometer 243

or so languages, although in terms of native speakers, it only ranks about 85th (Parkvall 2006:55–64). Swedish is also the dominant language in broadcasting in Sweden, including the nationwide public service networks. It should be kept in mind, however, that much of the material broadcast is of foreign origin, which in the overwhelming majority of cases means Anglo-American.

Swedes are in general keener on using the internet than most other nationalities, and more than two thirds of the adult population use it daily.[21] 85% of the population have access to a broadband connection, and more than half of the Swedes are internet users before the age of four.

Selected Further Reading

Bandle, Oskar (ed) (2002). *The Nordic languages. An international handbook of the history of the North Germanic languages. Vol*. 1. De Gruyter.

Bergman, Gösta (1973) *A short history of the Swedish language*. Swedish institute for cultural relations with foreign countries.

Bergman, Gösta (2001) *Ord med historia*. Prisma.

Bolander, Maria (2005) *Funktionell svensk grammatik*. Liber.

Falk, Maria (2001) *Domänförluster i svenskan*. Utredning för Nordiska Ministerrådets språkpolitiska referensgrupp.

Haugen, Einar (1976) *The Scandinavian languages. An introduction to their history*. Faber.

Hedelin, per (1997) *Norstedts svenska uttalslexikon*. Norstedts ordbok.

Holmes, Philip & Hinchliffe, Ian (2003) *Swedish. A comprehensive grammar*. 2nd edition. Routledge.

Hultman, Tor (2003) *Svenska Akademiens språklära*. Norstedts ordbok.

Josefsson, Gunlög (2001) *Svensk universitetsgrammatik för nybörjare*. Studentlitteratur.

Lewis, M. Paul (ed.) (2009) *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com

Malmgren, Sven-Göran (1994) *Svensk lexikologi. Ord, ordbildning, ordböcker och orddatabaser*. Studentlitteratur.

Nationalencyklopedins ordbok, band 1–3. Bra Böcker, 1996.

Nyord i svenskan från 40-tal till 80-tal. Norstedts ordbok, 1986.

Parkvall, Mikael (2006) *Limits of language*. London: Battlebridge.

[21] http://www.iis.se/docs/SOI2010_web_v1.pdf

Parkvall, Mikael (2009) *Sveriges språk – vem talar vad och var?* Stockholm: Stockholms universitet.

Parkvall, Mikael (2010) Invandrarspråk. In: Dahl, Östen & Lars-Erik Edlund (eds.), *Språken i Sverige*, pp 142–147. Stockholm: Sveriges Nationalatlas.
Garlén, Claes (2003) *Svenska språknämndens uttalsordbok*. Norstedts ordbok.

Pettersson, Getrud (2005). *Svenska språket under sjuhundra år*. Studentlitteratur.

Rosenqvist, Håkan (2007) *Uttalsboken*. Natur & Kultur.

SAOB – Ordbok över svenska språket utgiven av Svenska Akademien. Gleerups, 1898–.

Svenska Akademiens grammatik. Norstedts ordbok, 1999.

Vikør, Lars (2001) *The Nordic languages: their status and interrelations*. Novus,.

Wessén, Elias (1973) *Våra ord, deras uttal och ursprung*. 2 utg. Norstedts ordbok.

Language Technology Support for Swedish

Language Technologies

Language technologies are information technologies that are specialised for dealing with human language. Therefore these technologies are also often subsumed under the term Human Language Technology. Human language occurs in spoken, signed and written form. While speech and sign are the oldest and most natural modes of language communication, complex information and the bulk of human knowledge is recorded and transmitted in written texts, which in turn tend to be derived from, and have close correspondences to, spoken language, rather than sign language. Because of this and for other, practical reasons, sign language has so far received relatively little attention in Language Technology research; for all practical intents and purposes the two language modalities that have been investigated in any depth in this connection are spoken and written language. Speech and text technologies process or produce language in these two forms. Figure 1 illustrates the Language Technology landscape. But language also has aspects common to both forms such as dictionaries, most of the grammar, and the meaning of sentences. Thus, large parts of Language Technology cannot be subsumed under either speech or text technologies. Knowledge technologies include technologies that link language to knowledge.

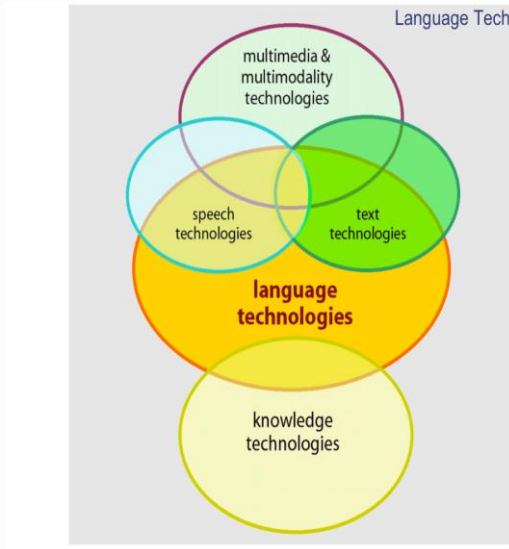

Figure 1: The Language Technology Landscape

In our communication, we mix language with other modes of communication and other information media. We combine speech with gesture and facial expressions. Texts can be combined with pictures and sounds. Movies may contain language in spoken and written form. Thus, speech and text technologies overlap and interact with many other technologies that facilitate the processing of multimodal communication and multimedia documents. Thus, large parts of Language Technology cannot be subsumed under either speech or text technologies. Knowledge technologies include technologies that link language to knowledge.

Language Technology Application Architectures

Typical software applications for language processing consist of several components that mirror different aspects of language and of the task they implement. A typical text processing pipeline will contain modules that deal with successive aspects of the structure and meaning of the text input. displays a highly simplified architecture that can be found in a text processing system. The first three modules deal with the following aspects of the input:

(1) **Pre-processing**: cleaning up the data, removing formatting, detecting the input language, etc.
(2) **Grammatical analysis** (including **lexical analysis**): reducing text words and word sequences to lexical entries and their associated grammatical information, variously complex depending on the language and the concrete software application; finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
(3) **Semantic analysis**: disambiguation (Which meaning of *bank* is the right one in a given context?), resolving anaphora and referring expressions like *she*, *the car*, etc.; representing the meaning of the sentence in a machine-readable way.

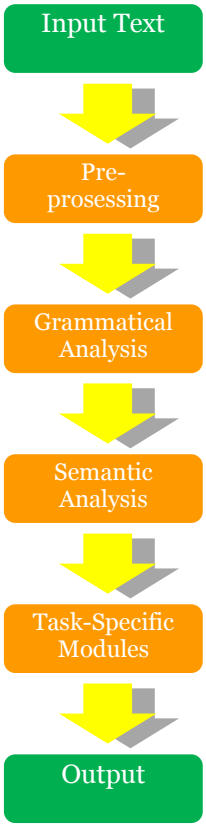

Figure 2: A Typical Text Processing Application Architecture

Task-specific modules then perform many different operations such as auto-

matic summarisation of an input text, database look-ups and many others. Below, we will illustrate **core application areas** and highlight their core modules. Again, the architectures of the applications are highly simplified and idealised, to illustrate the complexity of Language Technology applications in a generally understandable way. The most important tools and resources involved are underlined in the text and can also be found in the table at the end of the chapter. The sections discussing the core application areas also contain an overview of the industries active in the respective field in Sweden.

After introducing the core application areas, we will give a short overview of the situation in Language Technology research and education, concluding with an overview of past and on-going research programs. At the end of this section, we will present an expert estimation on the situation regarding core Language Technology tools and resources on a number of dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of Language Technology for Swedish.

Core Application Areas

Language Checking

Anyone using a word processing tool such as Microsoft Word has come across a spell-checking component that indicates spelling mistakes and proposes corrections. Four decades after the first spelling correction program was made by Ralph Gorin, language checkers nowadays do not simply compare the list of extracted words against a dictionary of correctly spelled words, but have become increasingly sophisticated. In addition to language-dependent algorithms for handling morphology (e.g. plural formation), some are now capable of recognising syntax-related errors, such as a missing verb or a verb that does not agree with its subject in person and number, e.g. in *She* ****write*** *a letter*. However, most available spell checkers (including Microsoft Word) will find no errors in the following first verse of a poem by Jerrold H. Zar (1992):

> *Eye have a spelling chequer,*
> *It came with my Pea Sea.*
> *It plane lee marks four my revue*
> *Miss Steaks I can knot sea.*

For handling this type of errors, analysis of the context is needed in many cases, as in:

> ***Faxen*** *blev tydligen* ***skickad*** *förra veckan, men jag har inte sett* ***den***.
>
> '**The fax machine** was supposedly **sent** last week, but I have not seen **it**.'
>
> ***Faxen*** *blev tydligen* ***skickade*** *förra veckan, men jag har inte sett* ***dem***.
>
> '**The fax messages** were supposedly **sent** last week, but I have not seen **them**.'

This either requires the formulation of language-specific grammar rules, i.e. a high degree of expertise and manual labour, or the use of a so-called statistical language model. Such models calculate the probability of a particular word occurring in a specific environment (i.e., the preceding and following words). For example, *sölig bardisk* 'soiled bar' (literally 'soiled bar counter') is a much more probable word sequence than *sölig bar disk* 'soiled naked counter' (with the parts of the compound written separately). A statistical language model can

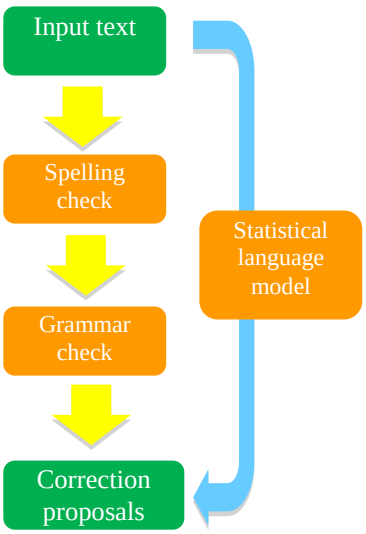

Figure 3: Language Checking (left: rule-based; right: statistical)

be automatically derived using a large amount of (correct) language data (i.e. a corpus). Up until now, these approaches have mostly been developed and evaluated on English language data. However, they do not necessarily transfer straightforwardly to Swedish with its more flexible word order and compound word building.

The use of Language Checking is not limited to word processing tools, but it is also applied in authoring support systems. Accompanying the rising number of technical products, the amount of technical documentation has rapidly increased over the last decades. Fearing customer complaints about incorrect usage and damage resulting from bad or poorly understood instructions, companies have begun to place an increasing focus on the quality of this technical documentation, at the same time targeting the international market. Advances in natural language processing lead to the development of authoring support software, which assists the writer of technical documentation to use vocabulary and sentence structures consistent with certain rules and (corporate) terminology restrictions.

Besides spell checkers and authoring support, Language Checking is also important in the field of computer-assisted language learning and is applied to automatically correct queries sent to Web Search engines, e.g. Google's *Did you mean…* suggestions.

Only a few Swedish companies and Language Service Providers offer products in this area, e.g. Scania and some SMEs.

Web Search

Search on the web, on intranets, or in digital libraries is probably the most widely used and yet underdeveloped Language Technology today.

The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries worldwide.[22] The verb *googla* 'to google' even has an entry in the Swedish modern dictionaries. Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, incorporated basic semantic search capabilities into their algorithmic mix,[23] which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a lot of data at hand and efficient techniques for indexing these data, a mainly statistically based approach can lead to satisfactory results.

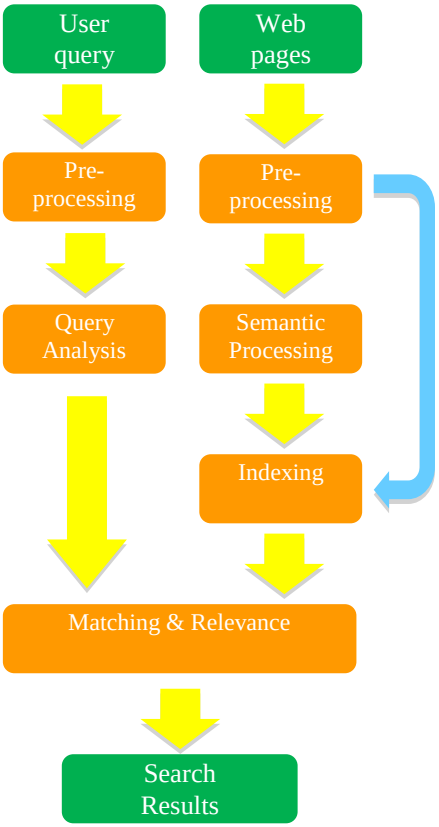

Figure 4: A Web Search Architecture

However, for a more sophisticated information need, integrating deeper linguistic knowledge is essential. In particular, if a search query consists of a question or a complete sentence rather than a list of keywords, retrieving relevant answers to this query requires an analysis of this question or sentence on a syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of relevant documents.

However, for a more sophisticated request for information, integrating deeper

[22] http://www.linksandlaw.com/technicalbackground-searchengines.htm

[23] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html

linguistic knowledge is essential. In the research labs, experiments using machine-readable thesauri and lexical-semantic resources like a wordnet (or the Swedish SALDO – Swedish Associative Thesaurus version 2[24]), can show improvements by allowing the user to find a page on the basis of synonyms of the search terms, or even more loosely related terms.

The next generation of search engines will have to include much more sophisticated Language Technology. If a search query consists of a question or another type of sentence rather than a list of keywords, retrieving relevant answers to this query requires an analysis of this sentence on a syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of the relevant documents. For example, imagine a user inputs the query *Give me a list of all companies that were taken over by other companies in the last five years*. For a satisfactory answer, syntactic parsing needs to be applied to analyse the grammatical structure of the sentence and determine that the user is looking for companies that have been taken over and not companies that took over others. Also, the expression *last five years* needs to be processed in order to find out which years it refers to.

Finally, the processed query needs to be matched against a huge amount of unstructured data in order to find the piece or pieces of information the user is looking for. This is commonly referred to as information retrieval and involves the search for and ranking of relevant documents. In addition, generating a list of companies, we also need to extract the information that a particular string of words in a document refers to a company name. This kind of information is made available by so-called named-entity recognisers.

Even more demanding is the attempt to match a query to documents written in a different language. For cross-lingual information retrieval, we have to automatically translate the query to all possible source languages and transfer the retrieved information back to the target language. The increasing percentage of data available in non-textual formats drives the demand for services enabling multimedia information retrieval, i.e. information search on images, audio, and video data. For audio and video files, this involves a speech recognition module to convert speech content into text or a phonetic representation, to which user queries can be matched.

Open source based technologies like Lucene and SOLr are often used by search-focused companies to provide the basic search infrastructure. Other search-based companies rely on international search technologies like, e.g. FAST or Exalead.

Focus on development for companies lies on providing add-ons and advanced search engines for special-interest portals by exploiting topic-relevant semantics. Due to the still high demands in processing power, such search engines are only economically usable on relatively small text corpora. Processing time easily exceeds that of a common statistical search engine, such as e.g. provided by Google, by a several orders of magnitude. These search engines also have high demand in topic-specific domain modelling, making it not feasible to use these mechanisms on web scale.
Hapax[25] has spent a great amount of resources around 2000–2005 and are now OpenAmplify.

[24] http://spraakbanken.gu.se/eng/saldo/
[25] http://www.hapax.com

Speech Interaction

Speech Interaction technology is the basis for the creation of interfaces that allow a user to interact with machines using spoken language rather than, e.g. a graphical display, a keyboard, and a mouse. Today, such voice user interfaces (VUIs) are usually employed for partially or fully automating service offerings provided by companies to their customers, employees, or partners via the telephone. Business domains that rely heavily on VUIs are banking, logistics, public transportation, telecommunications, customer services and call routing. Other usages of Speech Interaction technology are interfaces to particular devices, e.g. in-car navigation systems, and the employment of spoken language as an alternative to the input/output modalities of graphical user interfaces, e.g. in smartphones.

At its core, most Speech Interaction implementations comprise the following four different technologies, each consisting of a complex set of sub-technologies:

1. Automatic speech recognition (ASR) is responsible for determining which words were actually spoken given a sequence of sounds uttered by a user.
2. Semantic interpretation deals with interpreting the output from the recognizer according to the purpose of the respective system.
3. Dialogue management is required for determining, on the part of the system that the user interacts with, which action shall be taken given the user's input and the functionality of the system.
4. Speech synthesis (Text-to-Speech, TTS) technology is employed for transforming the wording of system-generated utterances into sounds that will be output to the user.

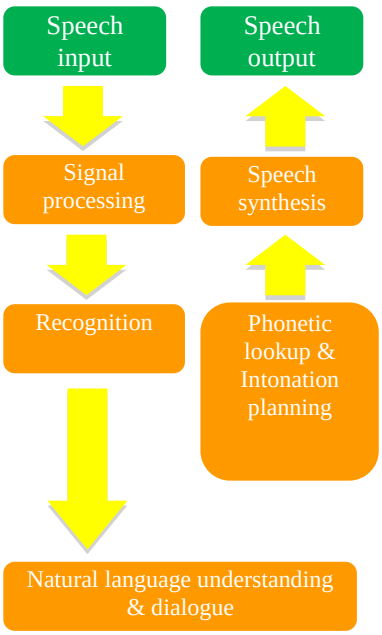

Figure 5: A Simple Speech-based Dialogue Architecture

One of the major challenges is to have an ASR system recognise the words uttered by a user with a sufficient degree of accuracy. This requires either a restriction of the range of possible user utterances to a limited set of keywords, or the creation of language models that cover a large range of natural language user utterances, or individual training by the user. Whereas the former results in a rather rigid and inflexible usage of a VUI and possibly causes a poor user acceptance, the creation, tuning and maintenance of complex language models may increase the costs significantly. However, VUIs that employ complex language models and initially allow a user to flexibly express their intent. In call routing applications, a statistical model is often used to reach an interpretation and a next action directly from the acoustic recognizer, so that the steps of semantic interpretation and dialogue management are merged with the speech recognition. In these applications, open prompts evoking free speech, for example *How may I help you?* greetings – show both a higher automation rate and a higher user acceptance and may therefore be considered as advantageous over a less flexible directed dialogue approach.

For the output part of a VUI, companies tend to use pre-recorded utterances of professional – ideally corporate – speakers. For static utterances, in which the wording does not depend on the particular contexts of use or the personal data of the given user, this will result in a rich user experience. However, the higher the degree of dynamic content in an utterance, the more the output suffers from concatenating single audio files, for example in terms of audible clips and poor prosody. In contrast, today's TTS systems are superior regarding the prosodic naturalness of dynamic utterances, although the overall quality is

often perceived as computer-like by companies purchasing the VUIs.

Regarding the market for Speech Interaction technology, the last decade underwent a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artefacts for a given application. There also has been strong market consolidation within the last ten years, particularly in the field of ASR and TTS. Here, the national markets in the G20 countries – i.e. economically strong countries with a considerable population – are dominated by less than 5 players worldwide, with Nuance and Loquendo being the most prominent ones in Europe.

On the Swedish TTS market, there are voices developed e.g. by Acapela, headquartered in Stockholm and also by the Swedish Library of Talking Books and Braille (TPB). There is also a strong research community mainly based at KTH, Stockholm (who have also developed their own systems).

Regarding dialogue management technology and know-how, markets are strongly dominated by national players, which are usually SMEs. Today's key players in Sweden are Artificial Solutions and SpeechCraft, and among smaller SMEs we can mention Talkamatic,[26] a developer of in-vehicle dialogue systems for the automotive industry. Rather than exclusively relying on a product business based on software licenses, these companies have positioned themselves mostly as full-service providers that offer the creation of VUIs as a system integration service. Finally, within the domain of Speech Interaction, a genuine market for the linguistic core technologies for syntactic and semantic analysis does not exist yet.

As for the actual employment of VUIs, demand in Sweden has strongly increased within the last 10 years. This tendency has been driven by end customers' increasing demand for customer self-service and the considerable cost optimisation aspect of automated telephone services, as well as by a significantly increased acceptance of spoken language as a modality for man-machine interaction. These factors were catalysed by the creation of the Graduate School of Language Technology (GSLT) network, bringing together industry players, research institutes and enterprise customers. In collaboration with others, the school has organised national workshops and invited industry to give talks to the graduate students. As academic partners, the Centre for Language Technology (CLT) at the University of Gothenburg and the department of Speech, Music and Hearing at KTH, Stockholm, were strongly participating in this process of spreading the knowledge about the advantages of Speech Interaction among Swedish enterprises.

Looking beyond today's state of technology, there will be significant changes due to the spread of smartphones as a new platform for managing customer relationships – in addition to the telephone, Internet, and email channels. This tendency will also affect the employment of technology for Speech Interaction. On the one hand, demand for telephony-based VUIs will decrease, in the long run. On the other hand, the usage of spoken language as a user-friendly input modality for smartphones will gain significant importance. This tendency is supported by the observable improvement of speaker-independent speech recognition accuracy for speech dictation services that are already offered as centralised services to smartphone users. Given this 'outsourcing' of the recognition task to the infrastructure of applications, the application-

[26] http://www.talkamatic.se/

specific employment of linguistic core technologies will supposedly gain importance compared to the present situation.

Machine Translation

The idea of using digital computers for translation of natural languages was suggested in the 1940s by A. D. Booth, W. Weaver and others, and was followed by substantial funding for research in this area in the 1950s and beginning again in the 1980s. Nevertheless, Machine Translation still fails to fulfil the high expectations it gave rise to in its early years.

At its basic level, Machine Translation simply substitutes words in one natural language by words in another. This can be useful in subject domains with a very restricted, formulaic language, e.g. weather reports. However, for a good translation of less standardised texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. One major difficulty here lies in the fact that human language is ambiguous, which yields challenges on multiple levels, e.g. word sense disambiguation on the lexical level (*spring* can mean 'a season', 'a source of water in nature', or 'a machine part') or the attachment of prepositional phrases on the syntactic level, as in the Swedish sentences:

> *Polisen tittade på mannen med kikaren.*
> 'The policeman was watching the man with the binoculars.'
>
> *Polisen tittade på mannen med revolvern.*
> 'The policeman was watching the man with the revolver.'

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be feasible in cases like the example above. But often rule-based (or knowledge-driven) systems analyse the input text and create an intermediary, symbolic representation, from which the text in the target language is generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by a skilled linguist.

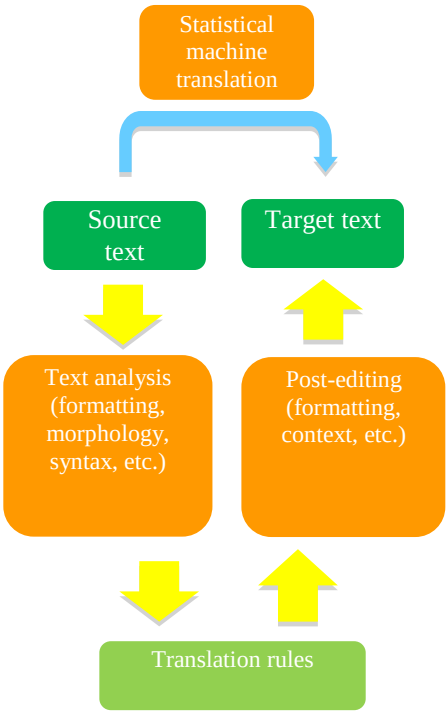

Figure 6: Machine Translation (top: statistical: bottom: rule-based)

Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for Machine Translation. The parameters of these statistical models are derived from the analysis of bilingual or multilingual text corpora, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 11 European languages. Given enough data, statistical Machine Translation works well enough to derive an approximate meaning of a foreign language text. However, unlike knowledge-driven systems, statistical (or data-driven) Machine Translation often generates ungrammatical output. On the other hand, besides the advantage that less human effort is required for grammar writing, data-driven Machine Translation can also cover particularities of the language that are missing from knowledge-driven systems, for example idiomatic expressions.

As the strengths and weaknesses of knowledge- and data-driven Machine Translation are complementary, researchers nowadays unanimously target hybrid approaches combining methodologies of both. This can be done in several ways. One is to use both knowledge- and data-driven systems and have a selection module decide on the best output for each sentence. Howev-

er, for longer sentences, no result will be perfect. A better solution is to combine the best parts of each sentence from multiple outputs, which can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

For Swedish, a challenging aspect of Machine Translation stems from the possibility of creating arbitrary new words by compounding, which makes dictionary analysis and dictionary coverage difficult.

A few Machine Translation systems handle Swedish currently and only a few of the larger commercial actors work on development of Swedish. In addition, there are some SMEs active in the field, e.g. Convertus AB.[27]

Provided that good adaptation is available in terms of user-specific terminology and workflow integration, the use of Machine Translation can increase productivity significantly. Commercial actors have developed special systems for interactive translation support. Language portals provide access to dictionaries and company-specific terminology, translation memory and Machine Translation support. An SME specializing in multilingual terminology mining and terminology management is Fodina AB.[28]

The quality of Machine Translation systems is still considered to have a huge improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories. In addition, most of the current systems are English-centred and therefore only few languages support translations to and from Swedish, which leads to frictions in the total translation workflow, and e.g. forces Machine Translation users to learn different lexicon coding tools for different systems.

Language Technology 'Behind the Scenes'

Building Language Technology applications involve a range of subtasks that do not always surface at the level of interaction with the user, but provide significant service functionalities 'under the hood' of the system. Therefore, they constitute important research issues that have become individual subdisciplines of Computational Linguistics in academia.

Question answering has become an active area of research, for which annotated corpora have been built and scientific competitions have been started. The idea is to move from keyword-based search (to which the engine responds with a whole collection of potentially relevant documents) to the scenario of the user asking a concrete question and the system providing a single answer: *At what age did Neil Armstrong step out on the moon? – 38*. While this is obviously related to the aforementioned core area Web Search, question answering nowadays is primarily an umbrella term for research questions such as (i) what *types* of questions should be distinguished and how should they be handled; (ii) how can a set of documents that potentially contain the answer be analysed and compared (do they give conflicting answers?); and (iii) how can specific information – the answer – be reliably extracted from a document, without unduly ignoring the context.

[27] http://www.convertus.se/home-en.html

[28] http://www.fodina.se/en

This is in turn related to the Information Extraction task, an area that was extremely popular and influential at the time of the 'statistical turn' in Computational Linguistics, in the early 1990s. Information Extraction aims at identifying specific pieces of information in specific classes of documents; this could be e.g. the detection of the key players in company takeovers as reported in newspaper stories. Another scenario that has been worked on is reports on terrorist incidents, where the problem is to map the text to a template specifying the perpetrator, the target, time and location of the incident, and the results of the incident. Domain-specific template-filling is the central characteristic of Information Extraction, which for this reason is another example of a 'behind the scenes' technology that constitutes a well-demarcated research area but for practical purposes then needs to be embedded into a suitable application environment.

Two 'borderline' areas – which sometimes play the role of stand-alone applications and sometimes that of supportive applications – are 'under the hood' components: text summarisation and text generation. Summarisation, obviously, refers to the task of making a long text short, and is offered for instance as a functionality within Microsoft Word. It works largely on a statistical basis, by first identifying 'important' words in a text (that is, for example, words that are highly frequent in this text but markedly less frequent in general language use) and then determining those sentences that contain many important words. These sentences are then marked in the document, or extracted from it, and are taken to constitute the summary. In this scenario, which is by far the most popular one, summarisation equals sentence extraction: the text is reduced to a subset of its sentences. All commercial summarisers make use of this idea. An alternative approach, to which some research is devoted, is to actually synthesize *new* sentences, i.e. to build a summary of sentences that need not show up in that form in the source text. This requires a certain amount of deeper understanding of the text and therefore is much less robust. All in all, a text generator is in most cases not a stand-alone application but embedded into a larger software environment, such as into the clinical information system which collects patient data, stores and processes it, where report generation is just one of many functionalities.

For Swedish, the situation in all these research areas is much less developed than it is for English, where question answering, information extraction, and summarisation have since the 1990s been the subject of numerous open competitions, primarily those organised by DARPA/NIST[29] in the United States. These have significantly improved the state of the art, but the focus has always been on English; some competitions have added multilingual tracks, but Swedish was never prominent. Accordingly, there are hardly any annotated corpora or other resources for these tasks. Summarisation systems, when using purely statistical methods, are often to a good extent language-independent, and thus some research prototypes are available. For text generation, reusable components have traditionally been limited to the surface realisation modules (or "generation grammars"); again, most currently available software is for English.

Language Technology in Education

Language Technology is a highly interdisciplinary field, involving i.a. the ex-

[29] DARPA stands for Defense Advanced Research Projects Agency and NIST for National Institute of Standards and Technology.

pertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists. Research in Language Technology started in Sweden already in the late 1960s, and after a slow but steady progress through the 1970s and 1980s, quite a lot of resources were invested in Language Technology research in the 1990s. The investments have contributed to a relatively well-developed Swedish research community with good organisation. In 2001 the National Graduate School of Language Technology (GSLT) was established by the Swedish government as one of sixteen national graduate schools. The graduate school is hosted by the University of Gothenburg, but is a collaboration between the following centres:

- University of Gothenburg
- University College of Borås
- Chalmers University of Technology (Gothenburg)
- KTH (Royal Institute of Technology; Stockholm)
- Linköping University
- Lund University
- Stockholm University
- Uppsala University

Supervision is also available from SICS (Swedish Institute of Computer Science; Stockholm).[30] Between 2001 and 2010 the University College of Skövde and Linnaeus University (Växjö University) were part of GSLT. At the time of writing, almost 30 PhD degrees have been awarded in the framework of GSLT, in a number of academic subjects, but with a concentration in Linguistics, Computer Science, and Speech Processing. GSLT has contributed significantly to the development of Language Technology in Sweden bringing different research centers and researchers together. It has made it possible to hold national courses and provide high-quality supervision. The PhD courses have also been offered to Nordic and Baltic PhD students through the NGSLT (Nordic Graduate School of Language Technology) network, funded by NorFA in the years 2004–2009. Through its national networking aspect GSLT has also contributed to several new research collaborations and joint proposals to national research funding agencies.

Currently there are two master's programs in Language Technology, one in Gothenburg and one at Uppsala University. Up until recently several universities also had undergraduate programs in computational linguistics (for example Lund University, University of Gothenburg, Uppsala University, Stockholm University) but the number of students has been dropping for several years, which is why new initiatives have been taken with the master's programs.

Language Technology Programmes

The existence of a comparatively lively Language Technology industry in Sweden can in part be traced back to major national Language Technology programs organised in the last decades.

For some years the Swedish Language council and GSLT have cooperated in building and maintaining språkteknologi.se,[31] a web portal for Swedish Lan-

30 http://www.sics.se/

31 http://sprakteknologi.se/

guage Technology with information about activities, resources, products and actors, both academic and commercial.

Sweden has a number of excellent Language Technology research centres:

Speech and Dialogue Technology:

KTH, Royal Institute of Technology

- CTT (Centre for Speech Technology),
- School of Computer Science and Communication (department of Speech, Music and Hearing)

University of Gothenburg

- CLT (Centre for Language Technology),
- Dialogue Lab, mainly at Department of Philosophy, Linguistics and Theory of Science

Text Based Language Technology Research:

University of Gothenburg, CLT, including several departments and units:

- Faculty of Arts
  - The Swedish Language Bank (Språkbanken)
  - Department of Swedish
  - Department of Philosophy, Linguistics and Theory of Science
- IT faculty
  - Department of Applied IT

Chalmers University of Technology

- Department of Computer Science and Engineering (also part of CLT)

University of Borås

- The Swedish School of Library and Information Science

Linköping University

- Department of Computer and Information Science

Lund University

- Department of Linguistics and Phonetics
- Department of Computer Science

Stockholm University

- Department of Computer and Systems Sciences
- Department of Linguistics

Royal Institute of Technology (KTH)

- School of Computer Science and Communication

Uppsala University

- Department of Linguistics and Philology

Research Institutes:

Swedish Institute of Computer Science (SICS)

Language Technology Consortia:

- Centre for Language Technology, University of Gothenburg (CLT)
- The Human Language Technology Center, Uppsala University, Stockholm University, KTH
- Graduate School of Language Technology (GSLT)

Language Council:

- The Swedish Language Council (Språkrådet)

Availability of Tools and Resources for Swedish

The following table provides an overview of the current situation of Language Technology support for Swedish. The rating of existing tools and resources is based on educated estimations by several leading experts using the following criteria (each ranging from 0 to 6).

1. **Quantity**: Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.
    - 0: no tools/resources whatsoever
    - 6: many tools/resources, large variety
2. **Availability**: Are tools/resources accessible, i.e.,are they Open Source, freely usable on any platform or only available for a high price or under very restricted conditions?
    - 0: practically all tools/resources are only available for a high price
    - 6: a large amount of tools/resources is freely, openly available under sensible Open Source or Creative Commons licenses that allow re-use and re-purposing
3. **Quality**: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Are these tools/resources current and also actively maintained?
    - 0: toy resource/tool
    - 6: high-quality tool, human-quality annotations in a resource
4. **Coverage**: To which degree do the best tools meet the respective coverage criteria (styles, genres, text sorts, linguistic phenomena, types of input/output, number languages supported by an MT system etc.)? To which degree are resources representative of the targeted language or sublanguages?
    - 0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases
    - 6: very broad coverage resource, very robust tool, widely applicable, many languages supported
5. **Maturity**: Can the tool/resource be considered mature, stable, ready for the market? Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Is the performance of such a technology adequate and ready for production use or is it only a prototype that cannot be used for production systems? An indicator may be

whether resources/tools are accepted by the community and successfully used in LT systems.
    - 0: preliminary prototype, toy system, proof-of-concept, example resource exercise
    - 6: immediately integratable/applicable component
6. **Sustainability**: How well can the tool/resource be maintained/integrated into current IT systems? Does the tool/resource fulfil a certain level of sustainability concerning documentation/manuals, explanation of use cases, front-ends, GUIs etc.? Does it use/employ standard/best-practice programming environments (such as Java EE)? Do industry/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?
    - 0: completely proprietary, ad hoc data formats and APIs
    - 6: full standard-compliance, fully documented
7. **Adaptability**: How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases etc.?
    - 0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand
    - 6: very high level of adaptability; adaptation also very easy and efficiently possible

Table of Tools and Resources

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	5	4	5	4	5	5	5
Parsing (shallow or deep syntactic analysis)	4	3	5	4	5	5	5
Sentence Semantics (WSD, argument structure, semantic roles)	2	1	2	2	2	1	2
Text Semantics(coreferenceresolution, context, pragmatics, inference)	2	1	3	2	2	1	2
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	1	1	1	1	1	1	1
Information Retrieval(text indexing, multimedia IR, cross-lingual IR)	4	1	4	3	4	3	3
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	4	2	4	4	4	3	4
Language Generation (sentence generation, report generation, text generation)	3	3	3	2	4	3	4
Summarization, Question Answering,advanced Information Access Technologies	2	1	1	1	1	1	1
Machine Translation	4	2	4	2	5	4	4
Speech Recognition	2	1	3	4	5	5	5
Speech Synthesis	3	1	3	3	3	3	3
Dialogue Management (dialogue capabilities and user modelling)	3	2	3	3	4	3	5
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	2	2	4	3	5	5	5
Syntax-Corpora(treebanks, dependency banks)	2	3	3	3	5	5	5
Semantics-Corpora	1	1	1	1	1	1	1
Discourse-Corpora	1	1	1	1	1	1	1
Parallel Corpora, Translation Memories	3	1	5	3	5	5	5
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	4	3	3	3	5	4	4
Multimedia and multimodal data (text data combined with audio/video)	1	1	1	1	1	1	1
Language Models	3	3	4	4	5	3	3
Lexicons, Terminologies	5	1	5	4	3	3	3
Grammars	3	2	3	3	3	4	5
Thesauri, WordNets	3	3	5	4	4	5	5
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	1	1	1	1	1	1	1

Conclusions

As already mentioned, Language Technology research has been pursued in Sweden since the 1960s. Compared to many other languages, Swedish is reasonably well endowed with language tools and resources. However, there is certainly room for improvement, especially in comparison to what is available for English and some other major languages. In many cases, tools and resources exist, but their wider use is hampered by proprietary licenses or arcane data formats, or both.

In this Whitepaper Series, the first effort has been made to assess the overall situation of many European languages with respect to language technology support in a way that allows for high level comparison and identification of gaps and needs.

If we think of the Swedish situation in terms of the BLARK (Basic LAnguage Resource Kit) concept, we may note that there is a conspicuous lack of certain basic resources:

- While there are some – mainly small – specific corpora of high quality, a large balanced corpus (a 'national corpus') does not exist, nor is a large syntactically annotated and manually validated corpus (treebank) available for Swedish. Corpus access is also generally restricted because many copyright issues remain to be resolved.
- No full-scale Swedish wordnet is available to the Language Technology community.
- In the area of multilingual resources, there is a clear focus on Swedish–English resources (and Swedish–English/English–Swedish Machine Translation), and not much in the way of support for other languages, e.g., the national minority languages, other Nordic languages, and other important European and world languages than English.

Many of the tools and resources lack standardization, i.e., even if they exist, sustainability and interoperability are not a given; concerted programs and initiatives are needed to standardise data, information models and interchange formats. As for many other languages, it is clear that the 'lower' levels of linguistic analysis – e.g., morphological and syntactic processing, as well as basic speech processing – are much better catered for than, e.g., semantics, text linguistics and pragmatics.

The most urgent needs of Swedish Language Technology at present are (in order of decreasing feasibility/increasing cost):

(1) Standardisation (for interoperability, of data and content formats, as well as APIs) of existing basic open source/open content tools and resources, in order to make them generally available to the research community and industry.
(2) Negotiations with the aim of improving licensing conditions of other existing basic tools and resources. If negotiations are successful, such tools and resources can then be standardised as in the preceding point.
(3) Creation of missing basic tools and resources in standard formats with maximally open licenses, e.g., a Swedish national corpus (which could include a treebank component and a number of parallel corpora) and a full-scale open Swedish wordnet linked to the English Princeton WordNet.
(4) Basic research on the higher levels of automatic linguistic analysis for

Swedish, and on integration of statistical and rule-based Language Technology, not least in order to aim for a closer interaction between speech and text technology.

META-NET

META – The Multilingual Europe Technology Alliance

META-NET is a Network of Excellence funded by the European Union. It currently consists of 44 members, representing 31 European countries, which are listed below. META-NET is fostering the Multilingual Europe Technology Alliance (META), a growing community of language technology professionals and organisations in Europe.

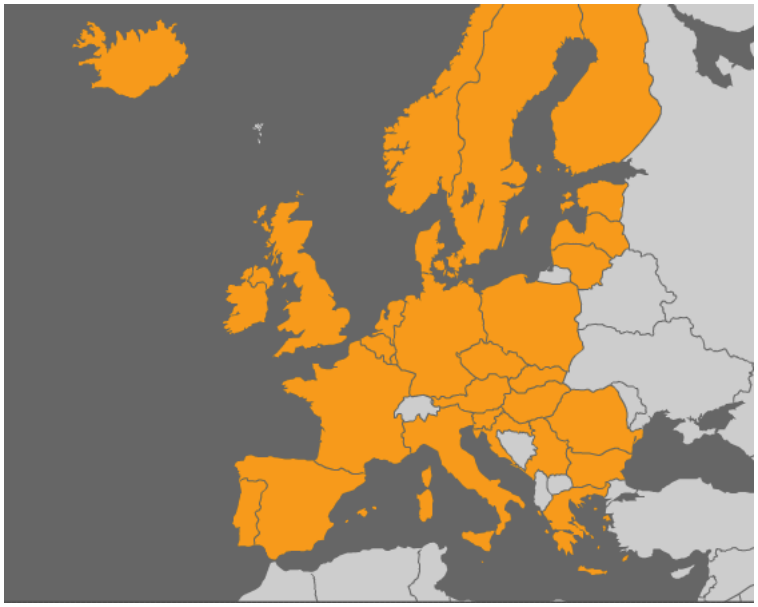

Figure 8: Countries Represented in META-NET

META-NET cooperates with a dozen other large initiatives like CLARIN, which is helping social sciences to establish the field Digital Humanities in Europe. META-NET is dedicated to fostering the technological foundations for establishing and maintaining a truly multilingual European information society that

- makes possible communication and cooperation across languages,
- safeguards equal access to information and knowledge for users of any language,
- offers advanced functionalities of networked information technology to all citizens at affordable costs.

META-NET stimulates and promotes multilingual technologies for all European languages. The technologies enable automatic translation, content production, information processing and knowledge management for a wide variety of applications and subject domains. The network wants to improve current approaches, so better communication and cooperation across languages can take place. Europeans have an equal right to information and knowledge regardless of language.

META-NET's Three Lines of Action

META-NET launched on 1 February 2010 with the goal of advancing research in language technology. The initiative supports a Europe that unites as a single, digital market and information space. META-NET has conducted several activities that further its goals. META-VISION, META-SHARE and META-RESEARCH are the network's three lines of action.

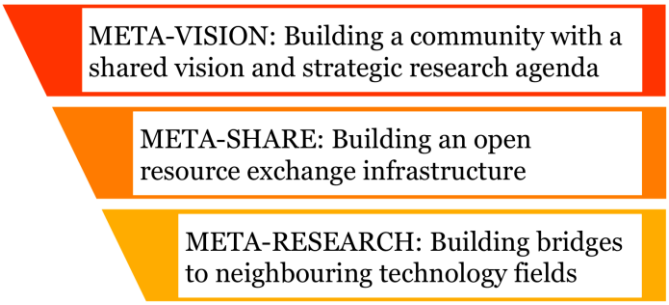

Figure 9: Three Lines of Action in META-NET

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. In META-NET's first year, presentations at the FLaReNet Forum (Spain), language technology Days (Luxembourg), JIAMCATT 2010 (Luxembourg), LREC 2010 (Malta), EAMT 2010 (France) and ICT 2010 (Belgium) centred on public outreach. According to initial estimates, META-NET has already contacted more than 2,500 LT professionals to share its goals and visions with them. At the META-FORUM 2010 event in Brussels, META-NET shared the initial results of its vision building process to more than 250 participants. In a series of interactive sessions, the participants provided feedback on the visions presented by the network.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items. META-SHARE targets existing language data, tools and systems as well as new and emerging products that are required for building and evaluating new technologies, products and services. The reuse, combination, repurposing and re-engineering of language data and tools plays a crucial role. META-SHARE will eventually become a critical part of the LT marketplace for developers, localisation experts, researchers, translators and language professionals from small, mid-sized and large enterprises. META-SHARE addresses the full development cycle of LT—from research to innovative products and services. A key aspect of this activity is establishing META-SHARE as an important and valuable part of a European and global infrastructure for the LT community.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, this activity wants to bring more semantics into machine translation (MT), optimise the division of labour in hybrid MT, exploit context when computing automatic translations and prepare an empirical base for MT. META-RESEARCH is working with other fields and disciplines, such as machine learning and the Semantic Web community. META-RESEARCH focuses on collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community. This activity has already

clearly identified aspects of MT where semantics can impact current best practices. In addition, the activity has created recommendations on how to approach the problem of integrating semantic information in MT. META-RESEARCH is also finalising a new language resource for MT, the Annotated Hybrid Sample MT Corpus, which provides data for English-German, English-Spanish and English-Czech language pairs. META-RESEARCH has also developed software that collects multilingual corpora that are hidden on the web.

Composition of the META-NET Network of Excellence

Country	Member (Affiliation)	Contacts
Austria	Universität Wien	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernolle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	Zagreb University	Marko Tadic
Cyprus	University of Cyprus	Jack Burston
Czech Rep.	Charles University in Prague*	Jan Hajic
Denmark	University of Copenhagen	Bente Maegaard, Bolette Sandford Pedersen
Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University*	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi, Krister Linden
France	CNRS, LIMSI*	Joseph Mariani
	ELDA*	Khalid Choukri
Germany	DFKI*	Hans Uszkoreit, Georg Rehm
	RWTH Aachen*	Hermann Ney
Greece	ILSP, R.C. "Athena"*	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi
	Budapest Technical University	Géza Németh, Gábor Olaszy
Iceland	University of Iceland	Eirikur Rögnvaldsson
Ireland	Dublin City University*	Josef van Genabith
Italy	ConsiglioNazionaleRicerche*	Nicoletta Calzolari
	Fondazione Bruno Kessler*	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiljevs
	Institute of Mathematics and Computer Science, University of Latvia	Inguna Skadina
Lithuania	Institute of the Lithuanian Language	Jolanta Zabarskaitë
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Malta	University of Malta	Mike Rosner
Netherlands	Universiteit Utrecht*	Jan Odijk
Norway	University of Bergen	Koenraad De Smedt
Poland	Polish Academy of Sciences	Adam Przepiórkowski
	University of Łódź	Piotr Pezik
Portugal	University of Lisbon	Antonio Branco
	Inst. for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufis
	University AlexandruIoanCuza	Dan Cristea
Serbia	Belgrade University	Dusko Vitas, Cvetana Krstev, Ivan Obradovic
	Pupin Institute	Sanja Vranes
Slovakia	Slovak Academy of Sciences	Radovan Garabik

Slovenia	Jozef Stefan Institute*	Marko Grobelnik
Spain	Barcelona Media*	Toni Badia
	Technical University of Catalonia	Asunción Moreno
	University Pompeu Fabra	Núria Bel
Sweden	University of Gothenburg	Lars Borin
UK	University of Manchester	Sophia Ananiadou

An * represents the founding members.

How to Participate?

META-NET and META offer many opportunities for participation. Please check out www.meta-net.eu for information on upcoming events and activities.