Information Society
and Media

HUMAN LANGUAGE TECHNOLOGIES FOR EUROPE

Work commissioned by

ITC
irst

Trento, Italy

Work carried out by

ACCIPIO
CONSULTING

Aachen, Germany

This work has been funded by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech-to-Speech Translation. Project Manager: Gianni Lazzari – IST-2002-FP6-506738, Web Site: *http://www.tc-star.org*

April 2006

Legal notice: The content of this report reflects only the author's views. The European Commission and the TC-STAR Consortium are not liable for any use that may be made of the information contained therein.

Preface

The richness of our languages is considered by many to be the distinctive and crowing achievement in human evolution. Language provides us with the means to communicate ideas, emotions and knowledge and express our cultural identity. All human achievements – science and technology, philosophy, art and culture – are enabled and empowered by language.

In Europe different languages are a fact of life. The European Union considers the diversity of tongues as an inalienable component of our cultural heritage; hence the principle of language equality is present in Europe's founding treaties. Preserving linguistic diversity has been at the heart of its policy from the very beginning. However, linguistic diversity demands a sustained and substantial investment, for example the European institutions spend considerable portions of their operational budgets on translation and interpretation services.

For European industry and business language diversity is both a challenge and an asset, as the Commission has recently pointed out in its first Communication ever on a new strategic framework for multilingualism[1]. What is sure is that early investment in multilingual communication technologies can provide rapid access to new and emerging markets anywhere in the world – a factor that is vital for the long-run success of Europe.

The ability to access and use information across languages is vital for citizens, governments, and commerce, and human languages technologies can play an important role in enabling easy communication between people, administrations and businesses. The European Union, in collaboration with the Member States, has sponsored over the last 20 years several R&D actions which have contributed to building expertise, resources and a pan-European language infrastructure.

[1] http://europa.eu.int/comm/education/policies/lang/key/legislation_en.html

Today Europe is one of the most advanced markets for language technologies and machine translation. The European Union is committed to ensuring that the necessary tools and resources are made available for all the EU languages as well as the world's principle commercial languages, paving the way to a pervasive multilingual information society in Europe. Through the deployment of multilingual products and services, such as cross-language information retrieval and machine translation systems, the European Commission aims at achieving its ambitious goal of generalising access to information for all European citizens – a key target of the 2010 initiative.

Multilingualism has become a policy at the European level and will be promoted through a variety of actions in the framework of education, training and research programmes, such as language learning programmes, research in linguistic diversity and human language technologies and digital content programmes.

The present document is a valuable record of the state-of-the-art and the challenges and opportunities waiting Europe in this important research field; it is equally an inspiring document for researchers, industrial players and policy makers, and will certainly contribute to make Europe more multilingual.

Viviane Reding
Member of the European Commission
Responsible for Information
Society and Media

Ján Figel'
Member of the European Commission
Responsible for Education, Training
Culture and Multilingualism

Executive Summary

For our multilingual Europe, cross-lingual communication and information exchange is of fundamental importance. Twenty official European languages, i.e. 190 language pairs or 380 translation directions, put cost and effort on every cross-lingual activity, in government, in business, and in our community. While this effort is comparatively small for some sorts of transaction and communication, it is large enough to prevent others from ever taking place. In the future this situation will be changed dramatically by the availability of translation capability provided by automatic systems – less perfect than professional human translators, but cheaper, faster, available on the spot, and good enough for many purposes. While efficiency gains in traditional human translation are also to be expected, the major uptake of these technologies will be in automated cross-lingual applications. Spoken language translation and machine translation will start in niche markets but rapidly expand in scope, largely independent of the currently existing translation services business. As enabling technologies, they will stimulate Europe's commerce and economy. For Europe it is a strategic necessity to have human language technologies available that facilitate cross-lingual communication and information exchange to the greatest extent possible.

This report begins by illustrating the significance of human language technologies, in particular for Europe, and describing the present state of affairs. It examines the European perspective in a global context with specific reference to the United States, India and East Asia. Current state of the art in research and in business is explored and expectations for future market developments outlined. Interviews with decision makers and specialists from research and business broaden the perspective and provide insight into the topic.

Table of Contents

1. What are Human Language Technologies?

We have long been aware of the ability of technology to fundamentally change the world around us and the way in which we live. Looking back, the world has been dramatically changed by computing, digitalization, and networking. Still, each of these revolutions took place gradually over a number of years, so change, while always remarkable, did not come as a shock.

We are now aware of another computing technology with enormous potential and it is fair to say that we are facing another revolution. It is proceeding rather slowly, and many of the research topics have been addressed for so many years that some have given up hope. It is very hard to teach computers to handle human speech and language – language in both the spoken and the written form – in the various ways that we humans can master: to speak naturally, understand what has been said (and meant), summarize a document or a conversation, find an audio recording given its content, translate from one language to another. We want to be able to interface with machines by voice and language, because we use these communication means, and we want computers to process this form of information in all the ways that we consider useful. The set of technologies which do this are known as *human language technologies* (HLT). Automatic speech recognition, machine translation and text to speech are the more prominent technologies, but there are many more. As with previous advances in computing, networking and digitalization, HLT has the potential to radically alter how we think about and work with information because we will be able to access and process the information encoded in language in fundamentally new ways. This report focuses on one single aspect of HLT, the ability to cross a language barrier, be it in communication between humans or in the treatment of unstructured information in the form of natural language text. Based on the core technologies *machine translation* and *spoken translation*[1], an important set of applications is formed by cross lingual information retrieval summarization, and data assimilation. Another important set comprises cross-lingual communication, i.e. machine translation of written text or spoken language translation or *speech-to-speech translation* (interpretation with written or spoken output). This is so new that we do not even have a term for it yet – would the right term be *cross-lingual information* and *communication technology*?

[1] See chapter 4.3. for more details on these technologies.

2. The Relevance of Human Language Technologies for Europe

2.1. A Critical Barrier for the European Internal Market

Four fundamental freedoms are enshrined in the EC Treaty[2], the *free movement of goods, persons, services and capital*. The internal market, the principal achievement of European integration, was realized at the end of 1992 as an area without internal borders for goods and services. There is no import tax on inter-EU trade, and national taxation systems must respect these four fundamental freedoms.

Concerning the harmonization of regulations and national legal requirements, much has been accomplished towards the support of the internal market. The free movement of goods and services is guaranteed by the *mutual recognition principle* in the single market which eliminates the need for a tedious complete harmonization of Member States' national legislation: Goods which are lawfully produced in one EU member state cannot be banned from sale on the territory of another member state, even if they are produced to technical or quality specifications different from those applied to its own products[3]. The same principle applies to services.

The language barrier is the last trade obstacle rinformation services in Europe.

With taxation and conformance, two important trade barriers have essentially vanished. On the logistic side, distribution of products is still an issue for physical goods. For information, however, a quantum leap occurred with the advent of the internet: Distribution of information has become so much faster and cheaper that one is tempted to say that you can have it for free and on the spot. While these three barriers have been largely overcome, a fourth barrier remains for most inter-country trade, namely the language barrier[4]. Communication across languages and cultures has become vitally important for trade, especially now with the globalization of the economy through the internet. The effort required in presenting a product in the language of the end customer can vary greatly and is dependent on the type of product involved. Generally, this effort will be high in relation to information services and this is reflected in high production

[2] Article 14 of the EC Treaty.

[3] The only exception allowed – overriding general interest such as health, consumer or environment protection – is subject to strict conditions.

[4] This simplification ignores that there are, besides the language issue, e.g. cultural differences that have to be taken into account when placing a good or service into a local market. However, this is in many cases a minor point compared to the effort that it takes to do translation.

costs. Put into one single table, the picture looks roughly like this:

	Goods	Information
Taxation	☺	☺
Conformance	☺	☺
Distribution	😐	☺
Language	😐	☹

Table 1: Trade conditions for international trade within the EU, for physical goods and for information / information services. For information services, the language barrier is a crucial barrier to overcome.

Of the four barriers identified, language remains as the major obstacle to providing information services across different countries. Using the human language technologies that are the theme of this report, this significant obstacle, this last major difference between our common European market and a large domestic market like in the United States, will be eliminated – with significant economic benefits.

2.2. World Languages

On a global scale, there are 6,912 known living languages[5]. Many of these languages are located in Asia / Pacific and in Africa (see fig. 2).

2.2.1. Major World Languages

English is probably the most important language, but in terms of first-language speakers, it is beaten by Chinese. However, if one takes the groups of first-language and second-language English speakers (375 million each) together with the speakers of English as a foreign language (750 million), the result is the significant figure of 1.5 billion people[6] able to speak English.

		Million speakers according to	
Rank	1st Language	source A	source B
1	Chinese	1,113	1,123
2	English	372	322
3	Hindi/Urdu	316	236
4	Spanish	304	266
5	Arabic	201	202
6	Portuguese	165	170
7	Russian	155	288
8	Bengali	125	189
9	Japanese	123	125
10	German	102	98
11	French	70	72
12	Italian	57	63
13	Malay	47	47

Table 2: Major world languages in millions of first-language speakers according to two different sources, (A) The English Company's engco model [Gra] and (B) comparative figures from the Ethnologue ([Gri]; see [Gra]).

Table 2 shows the full list of top ranking languages in terms of first-language speakers. It has been taken from the literature. Quite interestingly, while you would expect a listing of the top 10 or top 20 rankings, the table shows the top 13 instead[7], apparently a number chosen to include French

[5] Source: [Gor]. – It should at least be mentioned that there is a problem of language identification. – A living language is one with at least one speaker for whom this is their first language.

[6] [Cry] and other sources. It must be noted, however, that figures in the literature are inconsistent, and that they largely depend on the supposed level of proficiency in the foreign language.

[7] We have taken the table from [Gra] which shows the top 13.

	Language	Influence
1	English	100
2	German	42
3	French	33
4	Japanese	32
5	Spanish	31
6	Chinese	22
7	Arabic	8
8	Portuguese	5
9	Malay	4
10	Russian	3
11	Hindi/Urdu	0.4
12	Bengali	0.09

Table 3: 'Global influence' of the 12 major languages according to the engco model (see table 2). An index score of 100 represents the position of English in 1995 [Gra].

and Italian. Even in dry statistics, when language is involved, national feelings, cultural issues and a self-centered (or European-centered) perspective are not far away. Language is strongly tied with culture and our sense of familiarity.

Quite obviously, the number of first speakers of a language does not match our intuitive feeling of relevance. As an example, French is perceived as an important language and is the second most commonly taught language in European schools after English. The (perceived) importance of a language also depends on other factors, such as economic and political importance (besides cultural heritage). *Table 3* indicates the global importance of some languages relative to English. It is interesting to note that the importance of a particular language varies according

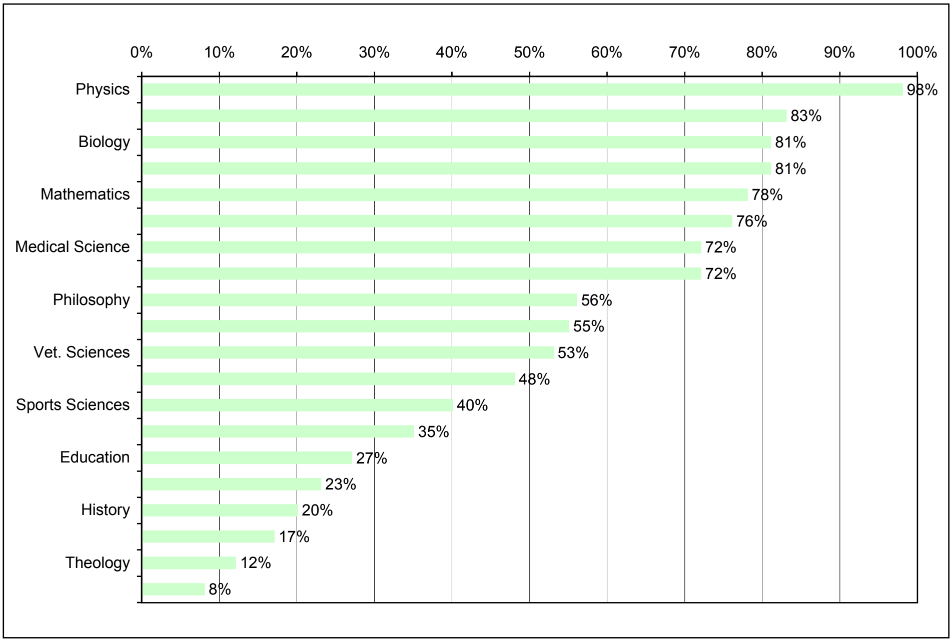


Table 4: Disciplines in which German academics claim English as their working language [Gra].

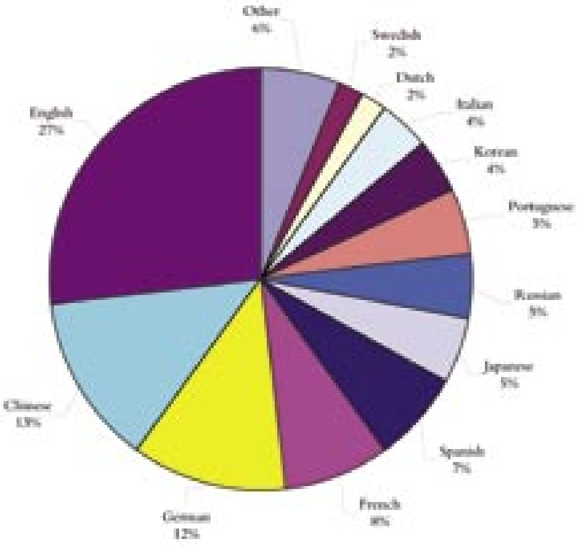


Fig. 1: The proportion of the world's books annually published in each language. English is the most widely used foreign language for book publication: over 60 countries publish titles in English. [Gra]

to the discipline or area of use. A specific example is given in *table 4*.

2.2.2. Endangered Languages

Endangerment of languages occurs in two dimensions: the number of speakers of the language and the number of functions for which the language is used. Typically, bilingual people begin to use only their second language with their children, or use their primary language less and less frequently. About 500 languages are listed in the Ethnologue[8] as nearly extinct[9]. There is a concern

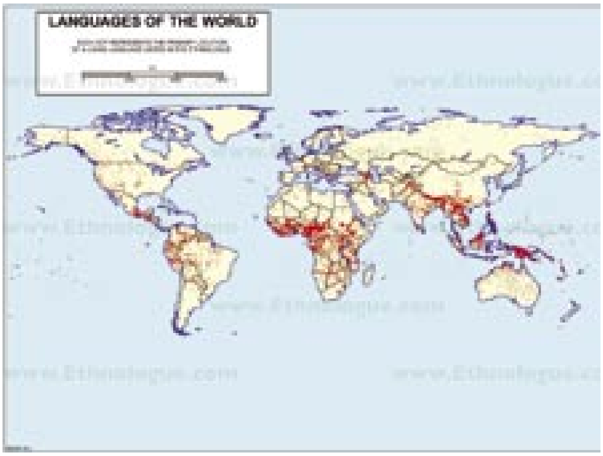


Fig. 2: Languages of the world. Each dot represents the primary location of a living language listed in the Ethnologue.

[8] The Ethnologue (GOR) is a catalogue of more than 6,700 languages spoken in 228 countries.

[9] Nearly extinct: defined by the speaker population being fewer than 50 or being a very small fraction of the ethnic group.

about language endangerment since language is closely linked to culture; this link is such that loss of language is almost always accompanied by social and cultural disruptions. (To the surprise of the author of this report, another concern is the loss to the academic community which studies such languages!)

2.2.3. Size Matters: About Primary, Secondary and Tertiary Languages and Market Forces

A language need not be endangered to be at a disadvantage. Consider a company which is going to extend their business from their local market to an international market. As the cost for localization into a new language does not correlate with the number of speakers but is more or less fixed, there are certainly primary languages – the ones in which one basically must have an offering – and there are secondary and maybe even tertiary languages of minor commercial relevance. While it depends on the circumstances *which* language is considered secondary or tertiary, it can be generally stated that the market forces penalize some languages, typically those with small speaker populations or those associated with weak economies.

A similar rationale even holds when people start learning a foreign language. Would a German rather learn Dutch, which is close to German and thus requires a comparatively small effort, or Spanish, taking into consideration that the Spanish speaking world is exceedingly large compared to the Dutch speaking one? (See *table 2* for typical figures.) And why learn Dutch when most Dutch speak English anyway?[10]. Many considerations influence a choice. Is the language spoken in a neighboring country or in a very distant country? Is there a secondary language in that country which could be used instead?

Many languages are under pressure by market forces.

The attractiveness of a language increases with the likelihood that one will need to speak it. Quite naturally, languages with a small speaking community are once again at a disadvantage.

2.3. A Closer Look at the European Union

Language is strongly related to culture. In many cases, it is a vital component of national identity. No wonder that we Europeans, when creating our European Union, have made the conscious decision not to introduce a primary language but to maintain the various languages and give them equal rights. While the high importance of English as the *lingua franca* of today should not be under-estimated, we certainly

[10] 91% of the Dutch population master a conversation in at least one other language [EB5]. It is, however, pertinent to note that the Netherlands have recently passed a law making knowledge of Dutch a requirement of citizenship. This is indicative of the importance of language to culture and identity.

Czech	CS	eština
Danish	DA	Dansk
Dutch	NL	Nederlands
English	EN	English
Estonian	ET	Eesti
Finnish	FI	Suomi
French	FR	Français
German	DE	Deutsch
Greek	EL	Elinika
Hungarian	HU	Magyar
Italian	IT	Italiano
Latvian	LV	Latviesu valoda
Lithuanian	LT	Lietuviu kalba
Maltese	MT	Malti
Polish	PL	Polski
Portuguese	PT	Português
Slovak	SK	Slovenčina
Slovene	SL	Slovenščina
Spanish	ES	Español
Swedish	SV	Svenska

Table 5: The 20 official languages of the European Union and their abbreviations [ELP]. Erse (Irish) will become the 21st official language of the EU from 1 January 2007.

live in a multi-lingual world, and language matters. The European Union is multilingual by design, and there is even a Commissioner for Education, Training, Culture and Multilingualism.

The various languages are considered to be of equal importance, and certain documents are available in (i.e. have been translated into) all languages, in particular laws or parliamentary debates. In order to be cost-effective and fast on the other hand, there are three working languages which are internally used in the operation of the European Union, namely English, French, and German.

To summarize, we have good reasons to protect our cultural heritage, but maintaining many languages comes at a cost[11], both economically (e.g. translation cost for a product) and in terms of other effort (the effort required of an individual to acquire an additional language). Any technology that reduces this cost supports our cultural heritage.

[11] It goes without saying that not producing translations is not an option: Citizens need to understand the law, products have to be localized into the various languages in order to be sold. Our argument here is that any cost, even if rather low, is a hindrance, and that any substantial reduction in cost and improvement in accessibility will have a beneficial impact.

Interview with Karl-Johan Lönnroth, Director-General, Directorate General for Translation (DGT)

Karl-Johan Lönnroth
Director-General
Directorate General for Translation
European Commission
Luxembourg, Luxembourg and Brussels, Belgium

Since Jan. 2004, Karl-Johan Lönnroth has been the Director-General of the DGT. Earlier posts:
2000-2003, Deputy Director-General, Directorate General for Employment and Social Affairs, CEC (European Commission)
1996-2000, Director of Employment Strategy and the European Social Fund, Directorate-General for Employment and Social Affairs, CEC
1991-1996, Director of Employment Department, International Labour Office, Geneva
1971-1991, in different posts at the Finnish Ministry of Labour as researcher, head of planning, deputy director for labour market services, and special advisor
1973-1977, Secretariat of the Nordic Council of Ministers, officer responsible for employment and migration affairs and tripartite co-operation. Consultancies for the OECD, the Nordic Council of Ministers and the Finnish Embassy in Stockholm.
Education: Master of Political Sciences, University of Helsinki (1970), Master of Arts, University of Wisconsin, USA (1972), Ecole Nationale d'Administration (ENA) (1983).
Language skills: Finnish, Swedish, English, French and German; elementary knowledge of Russian and Spanish.
Mr. Lönnroth has extensive experience in international co-operation: the Nordic countries, OECD, UN, ILO; and bilateral co-operation e.g. Eastern Europe. He has published over 40 articles, publications etc. concerning labor, employment, migration, social and political issues, and societal issues.

The DGT is quite a large organization, isn't it?

I would assume we are the largest in the world. We work with 21 languages. The global figure for the whole linguistic service for the European Union is € 1.1 billion a year or about 1% of the whole EU budget. That includes the interpretation and translation, and it includes not only the Commission but all the other institutions like the European Parliament. It sounds like a lot but all in all it is € 2.55 per citizen.

Doesn't the large number of languages put a heavy burden on the European society and economy?

That question is somewhat value loaded. It assumes that this is a burden and a cost rather than an advantage. Multilingualism is actually part of the European social model; we have this cultural diversity. It is to be considered as a wealth and dynamical element rather than a hindrance, and I would say that of course the issue of having several languages opens up markets. The multilingualism policy is beneficial for our dynamic society. In Europe, we respect fundamental freedoms and cultural diversity, and through this diversity you get also new ideas and new impetus to the economy. It is a burden for those who would like to have the freedom of movement in the single labor market and do only know one language. This is one reason why the Union tries to promote the knowledge of languages and language learning.

How has DGT dealt with the challenges connected with the EU enlargement?

We have recruited more than 500 new staff during the last year. I think we are assuming the new challenge quite well. The difficulty that we have is to develop and coin new terms and improve the quality of the translations, because some of the terms in the European policy do not exist in the new languages. That is the main challenge. We mostly translate legislative text from scratch and for the first time, and that's relativity complicated sometimes to translate. The integration of these new languages has gone rather well.

How do you expect machine translation to develop?

Machine translation as a translation tool and translation memory is something that is still developing, and it has a bright future in my view. However this does not mean that the human element will go down. The development goes more into the direction of a combination which I would call intelligent translation, the translator is then to ensure the quality, coherence, correct terminology, and the machine translation is there to help understand and to improve the productivity.

How do you see the profession of the translators?

The language industry as such is the fastest growing industry in the world. Globalization and the multitude of languages increase the need for multilingual services, which also means that the profession should be better recognized, because it is very important. Given the big challenges, it will also change: A translator is no longer only a transformer of a text but becomes more a linguistic adviser, an editor. So the diversification of the profession will also continue.

We are in contact with universities and their associations to develop the curricula and training, and we are trying to develop what we call the European master of translation which would be a kind of standardized qualification, which everybody would recognize, and also it would make recruitment easier. That's what I mean, with a better professionalization of the profession.

2.4. Communication

Many Europeans speak two or more languages. But about half of European Union citizens speak no language other than their own.

Wouldn't it be a great progress if Europeans who do not speak the same language could talk easily with each other? Any form of communication, even on a reduced level, would be progress with respect to the current situation. The translation would not necessarily need to be perfect or well formed. Such a requirement is very different from the requirements currently placed on professional translation. As we will see later, this has a fundamental effect on market forces and especially on the prospects of new market entrants.

Can we imagine today what a seamless and low-cost translation of human language would mean to us? We have experienced technology doing a great deal for the enhancement of communication between people, bridging both space and time. One of the first major steps in human culture was the development of human writing, which allowed knowledge to be passed on to future generations. Bridging space started very early, for example, the transmission of information along the borders of the Roman or the Chinese empires using optical signals. The introduction of postal services, the introduction of the telephone, the transatlantic telecommunication line, the cellular phone and the Internet were other giant steps.

We are facing another big step: the instant and cheap availability of translation, now bridging cultures rather than time and space. Will this step be considered as relevant as the others in two decades from now? It is difficult for us to imagine a future with the availability of instant speech translation and written text translation

Language	speaking it as a mother tongue	speaking it not as a mother tongue	total speakers
English	13%	34%	47%
German	18%	12%	30%
French	12%	11%	23%
Italian	13%	2%	15%
Spanish	9%	5%	14%
Polish	9%	1%	10%
Dutch	5%	1%	6%
Russian	1%	5%	6%

Table 6: Foreign language skills in the EU: Proportion of European citizens speaking the respective language in the EU (as mother tongue / as secondary or foreign language / either one). The survey was fielded in the European Union of 25 Member States and, in addition, in the accession countries (Bulgaria and Romania), the candidate countries (Croatia and Turkey) and among the Turkish Cypriot community. – Source: [EB5]

at reasonable cost because, at present, such communication simply doesn't take place. It is a fair assumption, though, that the pervasive availability of such a tool will have effects as significant as those that arose from the introduction of the telephone or the Internet.

After bridging time and space, bridging languages and cultures is the next big step.

2.5. The Next Step in Industrialization: Machines that Process the Written or Spoken Word

As we have just stated, the ubiquitous availability of low-cost translation will give rise to fundamental changes. In a revolutionary leap, a process that previously could only be performed by humans can now be performed by machines in a very effective way.

In the nineteen sixties, information technology spread from government, scientific and military applications to commercial usage, first in banking and insurance and later in virtually all areas of human life. It became possible for machines to process information. Much of the work which had previously required skilled humans was now done by faster, cheaper, more efficient machines. Only certain types of information processing were possible or showed reasonable performance. The processing of language, beyond that of structured data such as names and addresses, remained essentially out of scope. While the production of written content, such as newspapers, books, and print material in general, was largely supported by IT, the processing of incoming information remained very limited.

With the advent of the Internet, a lot changed. The necessity of processing language in several respects, such as searching, summarizing, translating, and classifying grew enormously. The broad scale introduction of customer self service on the Internet has lead to a massive increase in end customer communication with both companies and governments. At the same time, customer expectations have risen, and most customers take for granted that an e-mail will be answered within a day. The word-based search of documents on the Internet, mainly by Google, has found its way into virtually every home. However, two major bottlenecks have become apparent: first, searching for words is only a substitute for the search for information, and one would rather be in a position to have a *semantic web* and the ability to search for content rather than words. Second, as the Internet is being adopted by large portions of the population, there is a growing need for localization, especially with respect to languages. This creates a huge demand for large-scale and/or real-time translation, specifically for the production of multilingual web sites, which have become a major driver of the translation industry worldwide, as well as for cross-lingual document searching; the latter is the subject of considerable activities in companies like IBM and SAP, Google and Yahoo.

On a larger scale, translation is driven by globalization and the internet.

2.6. The Market

Concerning the economic side of translation, it is not enough to look at the currently existing translation market. Other important aspects have to be taken into account. For Europe, the cost of translation and its effect on the European economy is of paramount importance in the respect that, besides the direct cost of translation, language barriers can hinder the exploitation of new opportunities.

2.6.1. Translation as a Cost Factor

While we cherish the cultural richness of Europe as reflected in its many languages, the language barriers place a burden on our economy, as they add transaction costs to any activity that crosses a language border. In many typical cases, this cost is only a small fraction of the product cost, maybe in the range of 0.25% to not more than 2%. Some of this cost is direct (even though it might not be easy to determine) and some is indirect in the sense that it inhibits the creation of value.

It is even difficult to give a reliable estimate of the direct cost. In a large corporation, for example, there might be a budget for the localization of the web site. However, all the small translation activities pursued in the many different departments are usually not covered by such a figure, and they might easily add up to a larger sum. On the other hand, starting from the salaries of all translators does not work either, as many of them work part-time or as free-lancers. A quantification difficulty arises if an organization pays a localization company which uses a subcontractor which pays a free-lancer: Should the money be taken into account at all stages or at the first stage only? Thus, available figures differ according to the methodology used[12]. It does, however, seem safe to state that the global translation market is in the range of 8 to 30 billion euro. Here are two figures on government spending from the EU:

- Each year the European Parliament spends € 300 million, or 30% of their budget, on the translation of all parliamentary debates and EU documents into the 20 official European languages.
- The European Union spends € 1.1 billion per year, i.e. 1% of their budget, on all translation and interpretation services.

The EU spends € 1.1 billion per year on their translation and interpetation services.

2.6.2. The Established Markets of Localization and Translation

While the public has some rough ideas about the work of translators, translating books or documents, not so much is generally known about the industry which does the localization of software and of Internet pages, which represents a large part of the language industry. It is important to understand that their customers – companies and institutions – really need 'localization'

[12] Read more about this topic in the interview with Renato Beninatto.

and not just translation. Apart from the translation aspect, localization of software also ensures that the software runs properly, this extends to the online help and documentation, and also ensures that written material is properly formatted. The different lengths of translated texts, both in numbers of words and numbers of characters, have to be taken into account along with different writing directions. Apart from left-to-right and right-to-left, there are bidirectional languages: If Latin words are embedded into Arabic texts, both writing directions have to be supported! There are many character sets to be supported, and double-byte languages like Japanese, Chinese and Korean require specific support by the software. (The Unicode system is a solution to these problems but might not be applicable when too much Legacy content is involved.) In general terms, software that is intended

Rank	Company	HQ Country	Revenue in US$ M	Employees	Offices	Status
1	Lionbridge Technologies	US	377.1	4,000	50	Public
2	Titan Corp.	US	285.4	n/a	n/a	Public
3	SDL International	UK	146.0	1,400	36	Public
4	STAR AG	CH	96.0	750	33	Private
5	RWS Group	UK	63.4	350	7	Public
6	SDI Media Group	US	60.3	200	20	Private
7	Xerox Global Services	UK	60.0	200	4	Public
8	Euroscript S.à.r.l.	LU	54.5	600	9	Private
9	Transperfect/Translations	US	50.2	325	29	Private
10	CLS Communication	CH	36.0	260	11	Private
11	Logos Group	IT	36.0	150	17	Private
12	LCJ EEIG	DE/IT/BE/SP	21.6	140	9	Private
13	Thebigword	UK	20.0	122	7	Private
14	Hewlett-Packard ACG	FR	20.0	65	6	Public
15	Moravia	CZ	19.0	350	11	Private
16	TOIN	JP	19.0	105	5	Private
17	Merrill Brink International	US	18.5	120	4	Private
18	VistaTEC	IE	18.2	123	3	Private
19	Transware	IE	18.0	160	8	Private
20	McNeil Multilingual	US	17.2	105	9	Private

Table 7: Ranking of top 20 language service providers - 2004 revenue. (The two important acquisitions of 2005, the acquisition of Bowne Global Solutions by Lionbridge and the acquisition of TRADOS by SDL, have been consolidated in the figures.)

to be localized should be planned accordingly from the beginning to avoid unnecessary cost afterwards. And last but not least, localization is also about cultural differences. It requires creativity and empathy to transport a concept in a way that it still works in a different language or culture.

The usual method for localization is to separate the product into text components as well as into user interface elements. The translatable text is then translated, and the user interface and documentation reengineered so that they function properly in the target language.

The main segments of the localization market are software and web site localization.

According to one source[13], the language industry generated 8.8 billion US dollars in revenue in 2005. This comprises human language translation as well as the use of tools etc. Two important segments currently drive the growth of the market: the handling of multilingual web sites and software localization. Counting only those with five employees or more, 5,000 companies worldwide contribute to this market.

The major technologies currently exploited in translation are: translation memory (TM), terminology databases, and software and tools to handle multilingual web sites and to do software localization. Translation memory is a well established (though not universally used) supporting technology that both drives down costs and improves quality, especially the consistency of translations, which is important, for example, in the technical and legal fields. In a similar way, machine translation or spoken language translation, used as a tool, can be expected to have a beneficial effect on both price and quality of human translation. As outlined in the forthcoming chapter 2.6.4., which also defines the terminology, machine translation is a sustaining innovation in this usage scenario.

[13] Common Sense Advisory [CSA]. Other sources estimate the global figures to be significantly higher, in the range of € 30 billion. For this report, it is not the exact figures that matter but the fact that the language industry is an established and growing industry, much smaller in size than other industries but acting as an enabler for commercial activities of a much more significant volume.

Interview with Renato Beninatto, COO, Common Sense Advisory

Renato S. Beninatto
Chief Operations Officer and VP of Consulting Practice
Common Sense Advisory, Inc.
Boston, USA

Renato has over 20 years of executive-level experience in the localization industry. He has served on the executive teams for some of the industry's most prominent companies, most recently as Vice President and Director of Alpnet Inc. and Berlitz GlobalNET, respectively. He focuses on strategies that drive growth on a global scale. He specializes in making companies successful in global markets and in starting businesses that span across borders.

Currently he is a partner and a lead research analyst at Common Sense Advisory, Inc. a market research and consulting company specialized in the translation and localization industry, with clients in all continents. Renato focuses on the supply side and metrics practice of the company.

I have seen figures about the volume of the global translation market that do not give a consistent picture. Why is it so difficult to arrive at these figures?

Market sizing is a complex exercise that involves equal doses of logic and skepticism. The ultimate goal is to achieve a credible approximation of the market.

What drove Common Sense Advisory to measure the translation industry was driven in part by my previous experience in selling translation services for two publicly-traded translation companies and personal contacts with the largest translation buyers in the market. When you talk about the translation market to outsiders, they invariably see huge opportunities, as they see the gap between what is published in all languages and what is published in their own. But when you are on the street, selling those services, the reality is much harder. Translation is not strategic, therefore, not budgeted. We have come up with the "translation-to-toilet-paper ratio" image to describe how much translation represents in companies budgets.

The major flaw of market sizing methodologies conducted by outsiders is that they rely on the information of market players. One of the key characteristics of the translations market is subcontracting up to four levels. For example, a multilingual vendor (MLV) outsources its Eastern European languages to a company in Hungary, which in turn outsources it to other companies in the Czech Republic, Poland, and Bulgaria, which in turn outsource it to a free-lance translator. At Common Sense Advisory, we only count the first outsourcing, which represents the money actually disbursed by the end user for the job.

Which methodology have you followed?

In "Beggars at the Globalization Banquet," a report that we published in November 2002, we found out that companies – depending on industry and size – spend between 0.25% and 2% of their xenorevenues* in translation. This a lot less than the 3% of total revenues that organizations tended to use as an assumption.

To arrive to our numbers, we tested other market sizing approaches to see if our numbers were consistent. So we looked at the number and the revenues of translation companies in the market, at the number of translators worldwide, and at documented government expenditures. The cross-tabulation of these data gave us confidence in our numbers.

What are the main findings?

We are constantly doing research on the language industry as a whole and finding details about every aspect of it. Our main findings to date have been that translation is something that gets done at mid-management levels and only reaches the boardroom when something goes very wrong.

We also found that at US$ 9.5 billion in 2006, the market is about the same size as the bicycle market worldwide, and that although there much talk about translation technology, the aggregate revenue of all players in this space is only around US$ 100 million.

So you say that translation and localization has only little commercial relevance?

No. Quite the opposite. Even though translation is very cheap, it enables companies to penetrate new markets and multiply its revenues. Reducing its cost will promote the real globalization of markets.

Can you give example figures for large translation users?

As you know, the DGT spent € 1.1 billion in 2004 with the accession of 10 new countries, but their numbers are more around 800 million now. The Canadian Translation Bureau has revenues of less than US$ 200 million. Among the private companies, Microsoft and Oracle are the big spenders. Microsoft spends around US$ 300 million a year and Oracle about US$ 200 million. Automotive companies spend between US$ 10 and US$ 35 million a year. There are very few companies that spend more than US$ 3 million a year in the market, and everyone wants to sell to them.

* Revenues from outside the home country

2.6.3. From Human Translation to MT: Dramatic Cost Reduction and Accessibility Improvement

Putting the quality issue aside for a moment, two major market drivers push the use of machine translation: cost and accessibility. Driving costs down by a significant order of magnitude will certainly boost the use of translation, as the latent demand is, as yet, far from satisfied. A strong growth in demand is highly likely and plausible and has been observed in other cases, such as in the airline industry with the low-cost carriers and in the telecom industry with the advent of voice-over-IP. It should be stressed that not only the translations as such contribute to the economic effect of the technology: new types of transactions made possible by technological development represent a significant portion of the economic effect. The second driver is accessibility. Most translation today is not executed as (real-time) interpretation but as paperwork in remote offices. Material to be translated is generally sent away, the translation is then provided after hours or days. Real-time interpretation exists but it is costly. Compared to latency times of several hours, the instant availability of a text translation opens up a range of possibilities for new applications that would be impossible if delays were involved.

Two major market drivers pushing MT: Cost and accessibility.

In terms of quality, machine translation will remain inferior to human translation for many years. As a consequence, the various market segments will be dominated by one of two product offerings, either human translation or machine translation. Which product offering dominates in any particular market segment will depend on the unique characteristics and demands of that segment. Human translation will prevail in all areas where high quality is an absolute necessity. In contrast, machine translation will take over the low end[14] of the market, and it will also dominate in new markets or market segments which emerge as a consequence of the availability of low-cost translation technology. With quality and performance improvement over time, machine translation will move up-market.

[14] Low end refers to the quality of the translation. In terms of turn-around time and access, automatic translation is clearly at the high end. It is expected to become very high volume.

Interview with Michael Anobile, Managing Director of LISA

Michael Anobile
Managing Director
The Localization Industry Standards Association (LISA)
Romainmôtier, Switzerland

An international businessman with over 25 years experience in the IT sector, Michael Anobile received a Bachelors of Science degree in communications at Syracuse University, and participated in the Masters Degree program in political communications at the University of Maryland. After relocating his family to Switzerland in 1980 to become European Training Manager for Exxon Office Systems, he subsequently held a number of European and Swiss senior management positions in the IT and language-technology industries, focusing on global business development and marketing.

A founding member of LISA (The Localization Industry Standards Association), and the Managing Director from its inception, he is responsible for the day-to-day management of the Association including outreach programs to other standards organizations (e.g., ISO, Unicode, Openi18N, W3C, OASIS, etc.) and government agencies (e.g., US Department of Commerce, the DoD, FBI, NVTC, NSA along with various Asian, Canadian and European language technology and national standards and trade bodies) as well as its international forums, training programs and industry marketing & public relations projects.

What is LISA about?

We are a membership driven organization focusing on companies, governments and NGO's. In addition to the SME sector, we work on the institutional level – groups like the World Bank, McDonald's, Coca-Cola, IBM, the European Directorate-General for Translation, Industry Canada or the Canadian Bureau of Translation. Some of these groups comprise greater than 1,000 translators. We help them to understand the global perspective that localization offers in terms of business, technology, and workflow issues. This includes internationalization, translation, and how to design products and services for the global market.

What is your major recommendation to do localization the right way?

Best practice in localization requires that a product be designed for international use. Therefore the internationalization effort, the product lifecycle or service, the markets and distribution are crucial issues. This "holistic" approach assumes that information will be translated. Therefore, how the translation process can be automated is fundamental to quality, end user acceptance and cost.

Can you estimate the demand for translation if it was for free?

I have no idea. However, practice shows that if you ask a customer whether or not they would purchase their translation if it were done by a machine they are highly likely to say: No. However, if you make available a translation that is in fact done by a machine, then you offer it to your customer as a free choice, they are highly likely to download and use that document. The dilemma is apparent: There is a certain amount of resistance to what you termed earlier as "disruptive technology".

What contribution has technology made to translation?

Translation memory is one of the most significant contributions that the industry has made to translation. By the way, it is twenty year old technology. Look how long it has taken for it to gain a foothold. I wholeheartedly applaud the effectiveness of the TRADOS, SDL and other TM developers like Atril and Logos, because they really worked hard to implement translation memory on a very wide scale, throughout many business sectors. TM takes the fundamental word processing concept that, if you ever typed something once, then you need never type it again. So when you translate something which you know you will be reused, then you need never translate that phrase again. Of course, you have to control the context of the memories, and ensure their reusability across various platforms and tools. This is how LISA's translation memory standard, TMX©, contributes to the industry.

What is your attitude towards machine translation?

MT is very cost effective and efficient in well-defined translation applications such as knowledge databases, call centers, and technical documentation. The industry best practice is to control the authoring and terminology management processes. Understand how the end user will interact with the application; understand the level of information and the quality required. This will enable you to streamline the automated translation process by defining, building, and maintaining specific terminologies. The result will be greater accuracy and higher receptivity. MT works very well when the right expectations and resources are applied.

One intention of this report is to raise awareness for the automatic translation technologies but also translation as such. What is your message?

It is extremely important that our political and business leaders understand the importance that language plays. Language is an enabler; not an obstacle. It increases understanding and cooperation amongst peoples and cultures. Language provides wider access to important social and political information. I agree with the goal of raising public awareness. Because it's more than just translation: It is about communication and providing access to relevant data that can help people be more economically, politically and socially responsive. A good example is the European Union Directorate of Translation's language policy to support developing markets in Eastern Europe and the new members.

2.6.4. Sustaining and Disruptive Technological Innovations

Isn't it puzzling when huge, dominant and well-managed corporations fail to jump on the next innovation? Well, the puzzle has been solved, now we understand the dynamics of innovation much better than a decade ago. It is certain properties of innovations that largely determine their uptake on the market and their chances of creating a totally new market champion; these properties are utilized to classify innovations as either *sustaining innovations* or *disruptive innovations*, defined as follows.[15]

Most technological innovations foster improved product performance. Such innovations are called sustaining innovations. Whether incremental or discontinuous, these *sustaining innovations* have in common the improvement of the performance of established products according to performance criteria that mainstream customers in major markets have historically valued.

Sometimes, however, innovations result in inferior product performance, at least in the short-term. These *disruptive* innovations typically underperform established products in mainstream markets, but they have other features that are of value to certain customers. These customers are generally new to the market.

Sustaining innovations typically do not change the market landscape, i.e. the market leader develops them (or buys in later) and remains the market leader. As disruptive innovations are inferior to the existing technology or the existing product offering, they are of no value to the current market leader in the present state of the market. Due to certain mechanisms that are both plausible and quantitatively underpinned, this leads to

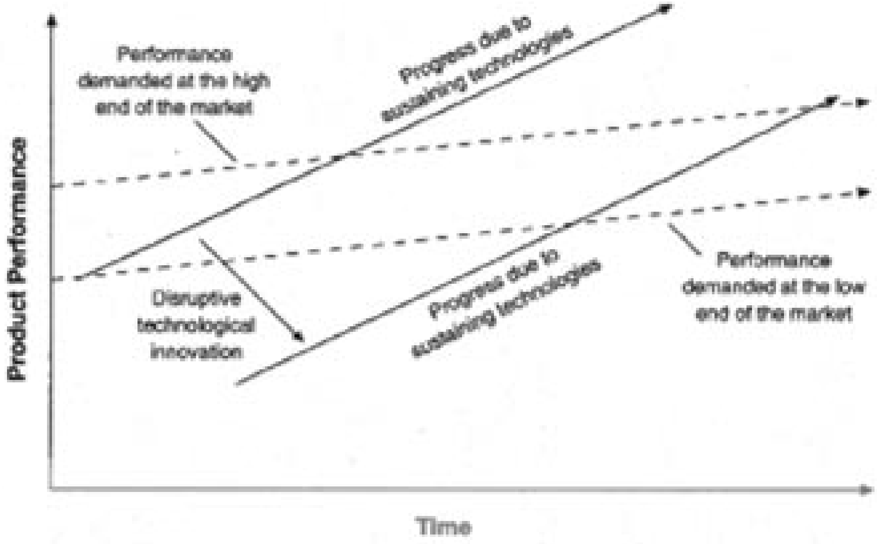

Performance demanded at the high end of the market
Progress due to sustaining technologies
Product Performance
Disruptive technological innovation
Progress due to sustaining technologies
Performance demanded at the low end of the market
Time

Fig. 3: Sustaining and disruptive technological change. – Source: [CRa].

[15] The following two paragraphs are close to [CRa].

Established Technology	**Disruptive Technology**
Minicomputers	Personal computers
Notebook computers	Hand-held digital appliances
Silver halide photographic film	Digital photography
Wireline telephony	Mobile telephony
Microsoft OS and Office	Linux and Open Office
Hard disk drives	Flash memory

Table 8: Examples for disruptive technological innovations, together with the corresponding established technology. – Source: [Chr], [CRa].

the situation where new players arrive and extend a smaller niche market, which is of little interest to the market leader, until it grows to an substantial size. Interestingly, the market leader is prevented from pursuing the new option by a behavior that is generally considered good management. With disruptive innovation, historically, the market leader is typically unable to retain his position and is replaced by another market player.

This poses an interesting question for existing market players and incumbents, but also for politics: *Are machine translation and spoken language translation sustaining or disruptive innovations?*

In the case of spoken language translation, the innovation is disruptive: it is not easy to see how this technology could support human interpreters, and compared to human translation, its performance, whilst improving, remains inferior. It is thus expected that this technology will start in markets very distinct from the existing ones and extend its market share over time.These target markets, on the other hand, are not addressed by human translators (the live translation of television broadcasts is a good example).

Machine translation can be both a sustaining and a disruptive innovation, depending on which application is involved and the market in which it is introduced. Translation memory techniques, as an example, are being employed as a preparatory and supporting tool for human translation, both facilitating translation and improving quality. In this context machine translation can be seen as a sustaining innovation: it improves what exists already.

In sharp contrast, most of the other applications of machine translation would be disruptive. Just as in the case of spoken language translation, machine translation services such as online web site translation, the online translation of chats, and cross-language customer self service would be inferior to human translation in terms of quality. The cost advantage, however, would allow growth of these markets, which would remain unattractive for any traditional human translation service due to the low margins involved.

MT and SLT are essentially disruptive innovations like the PC or digital photography: Starting in a niche and going strong later.

3. Asia, Europe and the United States: Similarities and Differences

The two major singular drivers of translation technology in recent years were 9/11 and the enlargement of the EU.

3.1. European Union

The situation concerning languages and the need for language technologies has to a large extent been described in chapter 2.3.: *A Closer Look at the European Union*. To summarize the main points:

- Europe has a multilingual society and is multilingual by design.
- All official European languages have the same rights.
- Translation from and to these twenty official languages, forming 190 language pairs, requires a substantial amount of effort.
- Any measure facilitating translation would foster communication between Europe's citizens and would amplify inter-EU trade.
- Any company dealing with the EU market has to localize its products to the regional markets by providing for local languages.

The European Commission is committed to HLT research and has over time funded several research projects on machine translation. Quick access is possible via the IST (Information Society Technologies) project search[16]. A non-exhaustive list comprises LC-STAR[17], MATCHPAD[18], METIS and METIS-II[19], NESPOLE![20], TC-STAR_p and TC-STAR[21], TQPRO[22] and TransType2[23]. The European Commission played a decisive role in producing necessary linguistic resources, having financed many projects in this area of study.

In the 6th Framework Program, the European Union spends € 135 million on multi-modal interfaces and language technology, i.e. roughly € 15 million per year on language technology.

The European research community, by nature, is well suited to the multifaceted task of language technology development. Europe has many large public research institutes of international quality, allowing the work to be divided and a variety of approaches to be pursued. The concepts of multilingualism and multiculturalism are generally well understood and embraced in both the academic and business

[16] IST project search: under http://www.cordis.lu/ist/projects/projects.htm

[17] LC-STAR: http://www.lc-star.com/

[18] MATCHPAD: http://www.systransoft.com/R&D/Matchpad/index.html

[19] METIS-II: http://www.ilsp.gr/metis/

[20] NESPOLE! : http://nespole.itc.it

[21] TC-STAR: http://www.tc-star.org/

[22] via the IST page: under http://www.cordis.lu/ist/projects/projects.htm

[23] TransType2: http://tt2.atosorigin.es/

spheres and supported by the political infrastructure. An infrastructure for language resources, the *European Language Resources Association* (ELRA), has been operational for 10 years.

3.2. United States of America

3.2.1. Strategic Role of HLT

The strategic role of HLT in the USA is substantially different from that in Europe. The United States has a huge domestic market with basically one language. This remains true despite the fact that there is a considerable Hispanic population speaking American Spanish. However, there is no strong economic necessity to serve this economically weak group, and there is no legal requirement to support this language. The United States is essentially a single language market.

The international dominance of English as a foreign language is an advantage for the US in many respects but carries a severe disadvantage in relation to homeland security. This gives translation, especially from other languages into English, a crucial role in gathering information from and about the world, be it of a more general kind or in relation to surveillance operations.

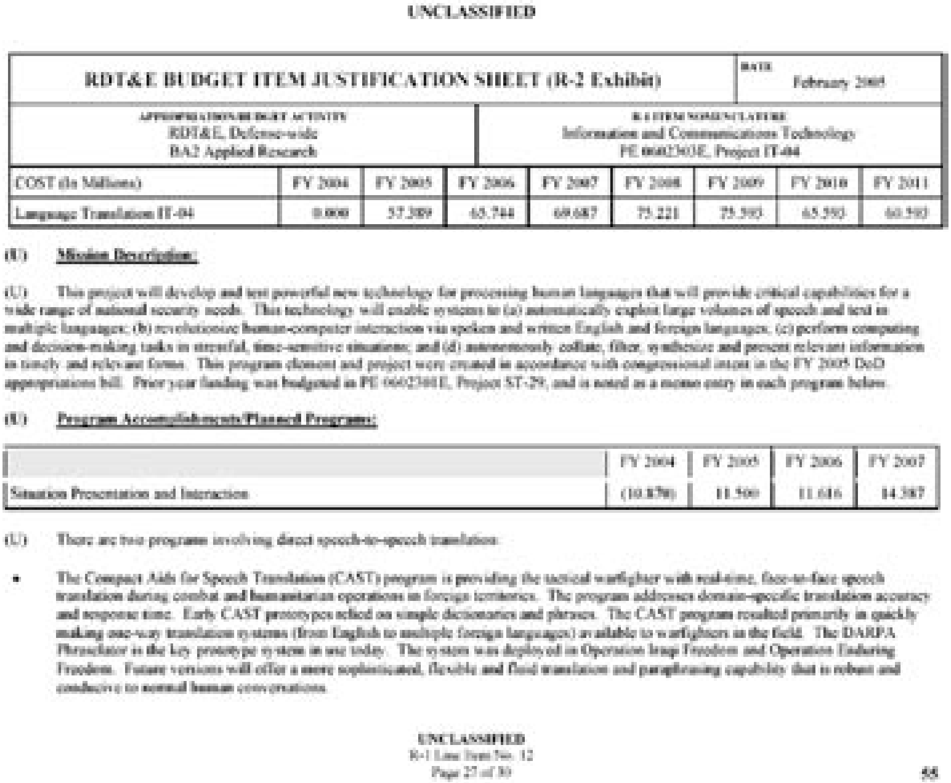
The US interest in HLT is largely driven by homeland security aspects.

3.2.2. Research Programs

After the attacks on the World Trade Center on September 11, 2001, it became apparent that the need for translators and competence in any language could dramatically rise at short notice. This has been a reason for the United States to launch a research program on cross-language information gathering from multiple sources, including machine translation, of an enormous volume. Expenditure is in the range of US$ 50+ million per year.

This money – a multiple of the amount spent by the European Commission – as well as the program itself will have a driving impact on the research community. While some of the scientific and technological advancements will be in line with the specific European needs, others won't – information gathering for military intelligence is in several respects different from the task of translating for a multilingual community. Moreover, the essentially monolingual structure of the US domestic market also raises reasonable doubt as to whether the US language technology market will necessarily deliver what Europe needs.

Without going into the research program details, it can generally be stated that the US tends to fund a small number of large projects while the EU funds large programs consisting of smaller projects. The current flagship project is GALE (Global Autonomous Language Exploitation) which focuses on information extraction from multilingual text and audio documents within an *unrestricted* domain (i.e. it also addresses spoken language translation). There are other ongoing translation projects like TransTac, STR-DUST and ACTD.



UNCLASSIFIED

RDT&E BUDGET ITEM JUSTIFICATION SHEET (R-2 Exhibit)					DATE February 2005			
APPROPRIATION/BUDGET ACTIVITY RDT&E, Defense-wide BA2 Applied Research				R-1 ITEM NOMENCLATURE Information and Communications Technology PE 0602305E, Project IT-04				
COST (In Millions)	FY 2004	FY 2005	FY 2006	FY 2007	FY 2008	FY 2009	FY 2010	FY 2011
Language Translation IT-04	0.000	57.389	65.744	69.687	75.221	75.593	65.593	60.593

(U) Mission Description:

(U) This project will develop and test powerful new technology for processing human languages that will provide critical capabilities for a wide range of national security needs. This technology will enable systems to (a) automatically exploit large volumes of speech and text in multiple languages; (b) revolutionize human-computer interaction via spoken and written English and foreign languages; (c) perform computing and decision-making tasks in stressful, time-sensitive situations; and (d) autonomously collate, filter, synthesize and present relevant information in timely and relevant forms. This program element and project were created in accordance with congressional intent in the FY 2005 DoD appropriations bill. Prior year funding was budgeted in PE 0602301E, Project ST-29, and is noted as a memo entry in each program below.

(U) Program Accomplishments/Planned Programs:

	FY 2004	FY 2005	FY 2006	FY 2007
Situation Presentation and Interaction	(10.870)	11.500	11.616	14.387

(U) There are two programs involving direct speech-to-speech translation:

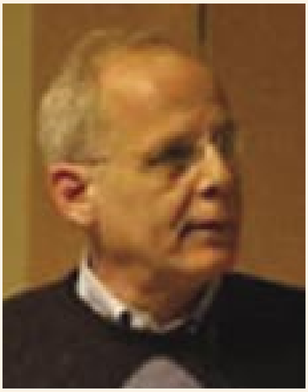
- The Compact Aids for Speech Translation (CAST) program is providing the tactical warfighter with real-time, face-to-face speech translation during combat and humanitarian operations in foreign territories. The program addresses domain-specific translation accuracy and response time. Early CAST prototypes relied on simple dictionaries and phrases. The CAST program resulted primarily in quickly making one-way translation systems (from English to multiple foreign languages) available to warfighters in the field. The DARPA Phraselator is the key prototype system in use today. The system was deployed in Operation Iraqi Freedom and Operation Enduring Freedom. Future versions will offer a more sophisticated, flexible and fluid translation and paraphrasing capability that is robust and conducive to normal human conversations.

UNCLASSIFIED
R-1 Line Item No. 12
Page 27 of 30

55

Fig. 4: US DoD (Department of Defense) budget for language translation technology. Fiscal year 2005 budget estimates for research, development, test and evaluation (RDT&E), defense-wide, at DARPA (Defense Advanced Research Projects Agency). – Source: DARPA.

Interview with Joseph Olive, Program Manager, DARPA



Joseph Olive
Program Manager
IPTO (Information Processing Technology Office)
DARPA - Arlington, VA, USA

Dr. Joseph Olive is the program manager of DARPA's Information Processing Technology Office. His current portfolio comprises one large program, GALE (Global Autonomous Language Exploitation).

Dr. Olive has had over thirty years of experience in research and development at Bell Laboratory. He has been the world leader in research of text-to-speech synthesis and has managed a world-class team in computer dialogue systems and human-computer communication. In his role as Director of speech research and CTO of Lucent Speech Solutions, he supervised the productization of Bell-Labs core speech technologies: Speech Recognition, Text-to-Speech and Speaker Verification.

Dr. Olive graduated from the University of Chicago with a degree in Physics (computational atomic physics) and an M.A. in music composition. After leaving the university, he combined his interest in computation and music and began research in acoustics and signal processing.

Dr. Olive was a recipient of the National Endowment for the Arts grant in 1974 to write a computer opera. He was also the recipient of the Bell-Labs' Distinguished Member of Technical Staff award in 1984.

Which research activities and goals are pursued in the GALE program?

Global Autonomous Language Exploitation (GALE) is a program which will provide distilled, concise and actionable information to our military. Since the source may be in a foreign language GALE requires a translation and a distillation engine. The input may be text or speech and the output maybe a complete translation or a distilled answer to a query (not necessarily in natural language). GALE's aim is to achieve translation and distillation accuracy to make it usable by the military. GALE follows two DARPA programs in HLT: EARS – a transcription program and TIDES – a program for translation, detection extraction and summarization. The previous programs were not designed to be a single end-to-end language system, but they provided great advances in HLT.

What is your position concerning a trans-Atlantic scientific cooperation?

There are several European groups involved with GALE. They are subcontractors on the large teams that were formed by GALE's Principle Investigators (PI) and are they are collaborating within their teams. In general, I believe that there are good collaborations between US scientists and European/Asian scientists. This is accomplished within government, corporate and university research.

How much public US funding goes into human language technologies like MT?

The total HLT program at DARPA is roughly US$ 50M for 2006. I do not believe that it should be broken down any further because the HLT technologies are (or should be) connected and integrated to achieve success in this field. I do not have information about the spending of other government entities on HLT, but I am sure that DARPA is not the only organization interested in the field.

How important are HLT, specifically MT and related technologies, to the US?

I personally feel that HLT is extremely important to the US government and to the military. It is necessary to communicate with and understand our allies and enemies, and many of them do not speak English. Also because of the information explosion, it is necessary to find the important information in the "haystack".

How is public funding of HLT research organized in the United States?

I have not seen an overall plan in the US for HLT research. There is an investment by various government agencies and industry. There is a great deal of collaboration between government sponsored research, corporate and university research, but it is more informal. As far as DARPA is concerned, we have a very ambitious vision and goals for HLT. We do have periodic evaluations to insure progress. We work both through cooperation and competition. Our teams are in it to win the evaluations. In addition, we have tied our goals and evaluations to the utility of the technology rather than just measuring accuracy. HLT deals with language and the first and foremost goal is to preserve the meaning of the language in the documents (spoken or written).

What are the state of the art and the next scientific challenges to be tackled in human language technology, in particular in translation?

The MT technology has improved a great deal in the past two or three years. A great deal of this improvement is due to the statistical paradigms coupled with the optimization procedures using BLEU. However, I am afraid that this paradigm is quickly reaching its limit, if it has not done so yet. I would assess that machine translation is only about a half way to where it should be. I would like to see new thinking into the problem even if it means that at the beginning the results may not be competitive. It is necessary for both MT and ASR technology to adopt a multi prong approach by incorporating other NLP techniques such as IR, parsing, extraction, etc. Although these are also statistical, they do not work in the same way and thus a combined system using all of the technologies could enhance the results leading to a solution. It is extremely important to couple ASR and MT more closely, not using the 1-best ASR result to drive the MT component.

3.3. Eastern Asia

3.3.1. English as a lingua franca in Eastern Asia

Considering its variety of people, cultures and languages, Asia bears resemblance more to Europe than to the US. But while Europe has started to establish itself as a unit, there is, as yet, no coherent policy in Asia and no identity as an autonomous unit. Among one another the countries in essence use English and the wide-spread use of English as a lingua franca in Asia implies that there is significantly less demand for the direct translation of language pairs like Thai – Japanese. As an advantage, each country primarily has to ensure that translation of its own language both from and into English is guaranteed, the restriction to just one language pair per language reduces complexity.

3.3.2. Asian Language Pairs and the Growing Importance of Chinese

There is also a downside to Asia's reliance on English as a lingua franca. Take, for example, the task of translating between Japanese and Chinese: given the common roots of these languages, the etymology of words is similar. It would be natural to take advantage of this fact, as well as of the cultural similarities. The use of English as a mediator is indirect and leads to unnecessary complications. As a new trend, Japanese companies are now increasingly interested in direct translation between Japanese and Chinese. This relates to the flourishing Chinese economy and the fact that many Japanese companies relocate their manufacturing to China. The interest in translation from and to Chinese is rising both on a global level and in Japan, and the relations between China and Japan are closer than in the past.

English is a lingua franca in Asia. Chinese gains relevance.

Economically strong and with a language that has its roots in old Chinese, Korea is a country with strong relations to both China and Japan. Both Japanese and Korean words are (in general) of Chinese origin, i.e. they share a similar etymology, albeit the fact that Korean doesn't use Chinese characters. Technological terms are often first used in Japanese, utilizing Chinese characters, and then taken up in Chinese. This similarity of the repository of words helps a lot in translation. Nevertheless, the three languages have different pronunciation and are linguistically quite different. Given this common cultural background and the fact that trade and interaction between the three countries is stronger than with other Asian countries, it can be assumed that the three languages, Chinese, Japanese and Korean, are likely to be directly translated in the next stage of development, while other Asian languages will be served mostly through English.

In the next stage Chinese, Kapanese and Korean will increasingly be directly translated into one another.

3.3.3. Research Programs

In Japan, research in human language technology reached a climax in the eighties with a high level of public funding as well as corporate interest in machine translation projects such as the EDR *Electronic Dictionary Project*. When the high expectations for machine translation could not be met at that time, reminiscent of the European *EUROTRA* project, interest and public funding started to decline. Today, there are indicators of a growing interest.

Given the dominant role of English, the focus of translation research is on translation from English to the home language and vice versa, which inherently leads to research programs conducted on a national level. It would also be somewhat difficult to execute transnational Asian projects, as no appropriate Asian organizational body exists. However, some aspects of the research work require international cooperation, which is hopefully provided and supported by Asian associations like the recently founded *Asian Federation of Natural Language Processing*[24].

In order to foster research on translation between Asian languages, it would be quite important to build language resources for these language pairs. Given the effort required as well as the necessity of sharing resources, the natural way to achieve these goals is through international cooperation. The next step of major importance would be the establishment of an evaluation agency.

[24] http://afnlp.org

Interview with Jun-ichi Tsujii, Director of the National Centre for Text Mining in Manchester and Professor in Manchester and Tokyo

Jun-ichi Tsujii
Director, National Centre for Text Mining, Manchester, UK
Professor, School of Informatics, University of Manchester, UK
Professor, Department of Computer Science, University of Tokyo, Japan

Professor Jun-ichi Tsujii was appointed, in July, 2005, director of the National Centre for Text Mining and Professor in Text Mining in the School of Informatics, University of Manchester, UK. He is also Professor in Natural Language Processing in the Department of Computer Science, University of Tokyo, Japan. He has worked in natural language processing since 1976. Starting with machine translation, he has widened his research to grammar formalisms for practical NLP application, HPSG-based parsing, information extraction and intelligent question answering. His successful research team on NLP at the University of Tokyo recently succeeded in applying a deep parser to produce semantic representations of all Medline abstracts (1.4 billion words).

He is recognized as one of the leading figures in bio-text mining, machine translation and multilingual NLP, and he has been keen in promoting Asia-wide cooperation. He has been invited to give tutorials, invited talks and keynote speeches at numerous major conferences, both in bio-informatics and NLP. He is permanent member of the ICCL (International Committee of Computational Linguistics, since 1992), vice-president (2005) and president (2006) of the ACL (Association of Computational Linguistics), president (2003-2005) of the IAMT (International Association of Machine Translation), and vice-President of AFNLP (Asian Federation for Natural Language Processing).

Machine translation was the topic of large research programs in the eighties, but the results did not meet the public expectations. How do you view these efforts today?

It is certainly true that the efforts in the 80's did not meet expectation of a large potential market. As one who was involved in some of these projects, I would say it was disappointing. However, thanks to these efforts, a firm basis of MT research and development systems was established. In Japan, at least six or seven MT vendors are still active in the market. More importantly, I think the visions we had at the time were well ahead of the time. That is, necessary technologies were not there for achieving our goals. I believe that, at present, since we have the technologies that were missing then we will be able to revitalize the field. For example, the same task of parsing a sentence by sophisticated grammar formalisms which took hours then takes less than a second by our program at the University of Tokyo.

Human language technologies including machine translation should be important for Asia. Is this perceived so on the national levels?

No, unfortunately not. Unlike Europe, English has long been considered as a single, international language for communication, and because of this, people do not think that it is important to treat "local languages" like Chinese, Korean, Japanese, etc. Obviously, this is wrong, and people are beginning to realize this. Because of the Internet, we suddenly realize that there is a huge demand for processing "local" languages, or that our local language is not really local.

What are the main differences of Asian and European languages? How close are the primary Asian languages among each other?

It all depends on definitions of European and Asian languages. Europeans can claim that languages in Europe are diverse, and I agree. However, the diversity of Asian languages is enormous and far beyond that of European languages, I think. Many languages in India, Middle East, Malay, etc, are completely unrelated with languages in Far East, Chinese, Korean and Japanese. Furthermore, Chinese and Japanese/Korean, while they share common vocabularies through long history of cultural exchanges, belong to completely different language families.

How would you describe the situation in all Asian countries in general?

I am not an appropriate person for answering this, but the situation in Asian countries is diverse again. Quite a few languages do not have agreed ways of transcribing them, let alone standard character codes. However, since technological progresses in many countries, like India, Thailand, China, etc. have been accelerating, the technological levels of these countries are more or less similar. We see many interesting research papers published by researchers in these countries.

Are there international activities in Asia that coordinate research in this sector or even collaboration on a global level?

We established an academic association, AFNLP (Asian Federation for Natural Language Processing) two years ago to promote cooperation and coordinate activities. There are other initiatives as well. However, compared with coordination in Europe, we are still far from the ideal.

What are the next steps to be done to promote HLT and MT in Asia?

We need more government involvement. The EU has played significant roles in promoting research and development in HLT and MT. Till the end of the 80's, Japan had played a leading role not only in technology but in financing the regional cooperation in the field. The situation has changed dramatically since then. We see several countries which can contribute to the field financially as well as intellectually. However, we do not have any pan-Asian governmental body to coordinate the activities. The academic community is ready to cooperate, but we need, for example, funding bodies which support Asia-wide projects.

3.4. India

Can you think of a country with as much language diversity as the united Europe? Well, India[25] and its thirty-five states have twenty-two languages which are official, i.e. approved by the constitution. As if this were not enough diversity, each language has on average about twenty dialects. The languages belong to three different language families, and there are different writing systems as well. The state language is Hindi, but the Indian constitution states that English can also be used for official purposes. In the cities, people often know three languages: the state language (e.g. Hindi, Bengali, etc.), Hindi (the official language of India), and English. Roughly 30% of the population speaks Hindi, and about 5% of the population is comfortable with English.

All official documents in the state capitals have to be in three languages (in English, Hindi and the state language). The most frequently employed translation directions are from English into Hindi and from English or Hindi into the respective state languages. Like in other countries, however, most translation is performed by humans, and there are not enough human translators to accommodate the demand. Concerning technical tools, translation memory is not widely used, but translation is often supported by electronic dictionaries and tools like morphological analyzers.

There is public and government interest in automatic translation and both government funding and Indian research on machine translation. International companies like IBM, Microsoft, Google and Yahoo are investing in MT, although they do not yet cover the Indian languages. As these companies are typically pursuing the statistical approach, a lot of parallel corpora[26] exist but they are not publicly available. Starting in 2006, this gap will be filled by the new government sponsored LDC-IL[27] (*Linguistic Data Consortium for Indian Languages*).

3.5. Economic Boundary Conditions

The economic boundary conditions concerning human language technologies are different between Europe and the USA. In fact, market conditions for speech recognition and machine translation in the US and the EU are directly opposed.

Let's first have a look into automatic speech recognition. At this point in time, the economically most interesting use of automatic speech recognition and dialogue technology lies in the area

[25] Acknowledgements to Professor Dr. Pushpak Bhattacharyya (Indian Institute of Technology, Mumbai) for a briefing on the situation in India.

[26] A parallel corpus is a collection of texts in two language versions, together with the information which sentences are associated to each other. (E.g., the first two sentences in language A might correspond to the first three sentences in language B, etc.)

[27] http://www.ciilcorpora.net/ldcil.htm

of customer self service or, to put it differently, in the automation of human call center services. Concerning this market and application, Europe differs from the United States in two respects. First, while in the US so-called IVR[28] systems were widely deployed, they had less usage in Europe. The American public had used the often somewhat tedious IVR systems for quite a while and was generally very happy with the migration from DTMF (Dual Tone Multi-Frequency) tones and rigid menus to speech dialogues[29].In Europe, where in some countries call center services have been available for free and have provided quite a high service level, the introduction of automatic systems is often perceived by callers as a regression. In addition, the United States represents a large and fairly homogeneous market with one language, much larger in size that any of the monolingual European markets. As a consequence, there is a better economy of scale in the US, as the development of a speech dialog application requires significant effort. Given these arguments, it is not astonishing to see that automatic speech recognition has been taken up more strongly in the US than in the EU.

The economic conditions for machine translation are exactly converse: difficult for the US but very favorable for Europe for both spoken language translation and machine translation. This is due to the fact that there are many languages spoken in Europe, while any system used for the huge domestic American market would cover only a small fraction of all activities. In Europe, doing business means being multilingual.

Favorable economic conditions for European companies as providers of MT services.

There are also strong economic forces that support a translation industry in Asia, but due to the dominant role of English, the market as well as the prospects for a language industry is smaller than in Europe. Given the current situation and company landscape, we would assume that, under normal circumstances, Europe will take the leading role here.

3.6. An Action Point for Europe

Looking back, the strategic role of HLT as well as the situation in the three large regions can be summarized as follows:

Europe: Language technology is an economic, political and cultural

[28] IVR stands for interactive voice response. Classical IVR systems react on callers' key presses by playing voice messages.

[29] A frequently quoted report – Nuance Communications - Market Research: "Nuance Speech User Scorecard, May 2000 – states that overall satisfaction with voice recognition was high (among 87% of respondents) and significantly higher than satisfaction with DTMF systems. While at least this finding is reasonable and consistent with the author's experience, care should be taken as the original report is no longer available on the Nuance site or in the part of the internet that is accessible to Google.

necessity. Breaking the language barrier would boost communication and the economy. While HLT is already the focus of considerable European research effort, the strategic importance of the technology to Europe warrants a much higher priority on the research agenda.

USA: The use of HLT is dominated by military considerations and the fight against terror. Very significant funding is currently allocated to HLT research and technology.

Asia: Translation to and from English is the first priority. Lack of common political identity and infrastructure make it difficult for Eastern Asia to establish a leading role. There is a high demand for translation in India.

Regarding development and com mercial exploitation, Europe has an urgent need and is at the same time in a privileged position. Given the generic nature of the technology, it offers the option to be commercialized in other world regions as well. On the other hand, it cannot be expected that our needs be satisfied by other suppliers.

Interview with Joseph Mariani, Director, French Ministry of Research

Joseph Mariani
Director, Information and Communication Technologies Department
French Ministry of Research, and senior researcher at LIMSI-CNRS
Paris, France

Joseph Mariani's research activities relate to language technology, multimodal human-machine communication, speech recognition, language resources and evaluation.

He was president of the European Language Resources Association (ELRA), president of the European (now International) Speech Communication Association (ISCA), a member of the board of the European Network on Language & Speech (ELSNET), and the coordinator of the francophone FRANCIL network.

Dr. Mariani was the director of LIMSI and the head of its Human-Machine Communication department (1989- 2001), a member of the CNRS Scientific Council, the chair of the CNRS Information Science and Technology Advisory Committee and a member of the Evaluation Committee of INRIA.

Since 2001 he has been director of the ICT department at the French Ministry of Research, where he is responsible for the research programs in telecommunications, software technologies, multimedia and nanotechnologies, including a specific program on language technologies.

Concerning Europe, the United States and Asia: Where do you see similarities, and where are the differences?

The United States with their largely mono-lingual domestic market see multilingualism primarily from the military and security point of view: everyone understands English, but they have a hard time understanding foreign languages! Asia, like Europe, uses many languages, so there is a commercial need for multilingualism as well, but it is not a common market like the EU. For Europe, multilingualism is of utmost importance, with more than 20 languages spoken in the 25 EU member states. Besides the economical dimension, which makes multilingualism a pure necessity, there are also political, cultural and societal dimensions. This makes our situation very special.

What is your attitude towards cooperation with the US or Asia?

International cooperation together with healthy competition is good for the scientific and technological progress, and I appreciate and support when our countries join forces in science and infrastructure to tackle the hard problems in HLT. On the other hand, international cooperation cannot replace our own agenda serving our own needs. Concerning European multilingualism, the EU must lead, no one else can or will do it for us.

Which developments are necessary in order to enhance Europe's position?

Despite some considerable amount of effort in the past, I believe that the level of research and technology funding is not yet compatible with the size of the challenge. Although the Commissioner for education, culture and multilingualism, Ján Figel, recently stressed the importance of multilingualism for Europe, language technologies appear as a tiny part of the ICT content in the FP7 preliminary program, lost at the end of the "Simulation, visualisation, interaction and mixed reality" technology pillar. And the topics presently selected for possible large Article 169 actions are "Research in the Baltic Sea", "Assistance to the elderly" or "Metrology", which are quite respectful, but, in my opinion, less strategic for Europe than the language issue. Language technology is probably the topic which best fits the idea of a European coordinated action, as the effort to cover the various technologies and the various European languages is too large to be carried out by the EC alone, and could easily be shared with the EU member states, with their own languages and their own programs. While the member states should primarily take into account what is specific to their language, or languages, such as language resources (speech and text corpora, dictionaries), and language specific technology adaptation, the EC could primarily address the aspects independent of a specific language, the general coordination, generic technology development and assessment, and standards.

Europe still also needs to settle an infrastructure to evaluate language technologies comparable to what exists in the US with NIST, and funding agencies should take into account the core technology performance evaluation when selecting projects, in order to avoid supporting the development of applications which obviously necessitate a technology of better quality.

This would complement the European Language Resources Association, ELRA, which has just celebrated its 10th anniversary.

Yes indeed. The situation on the aspects of language resources and language technology evaluation has greatly improved due to the existence of permanent entities such as NIST and LDC in the US, or ELRA in Europe, but the challenge for Europe necessitates even more. I advocate the creation of a Language Technology Agency at the EC level, a permanent structure which would coordinate the efforts of the EC and the EU member states towards the necessary language technologies for a multilingual Europe, and would put Europe at the forefront of HLT worldwide, thus taking advantage of its linguistic challenge.

4. Where We Stand Today

4.1. Translation Work Today

While technology is certainly involved, in the end, translation as well as localization is a service provision.

Translation work is very diverse. It ranges from the translation of the school diploma of a guest student to the handling of a huge Fortune 500 company web site, from language pairs of similar languages like Dutch and German to radically different ones like Italian and Chinese, and it could cover standard language, prose or highly technical content.

In part, translation takes place inside the institutions or companies who need it, and in part it is outsourced to translation service providers. The top 20 translation service providers cover only 16.3% of the market[30], so the market is indeed very fragmented. If we count companies with at least five employees, there are 5,000 worldwide offering translation services. Many translators work as freelancers, and even the large translation service providers and the large translation users use freelancers for their translation work. Take the DGT (Directorate-General for Translation of the European Commission) as an example: There, the proportion of work done by freelancers has risen from 11.8% in 1992 to 23.0% in 2004.

4.2. Technologies Used in Professional Translation

Major technologies used in translation services today are translation memory and terminology databases.

Quite naturally, the technical sophistication of the translators' works depends very much on both the company he or she works for and the nature of the material to be translated. If the translations vary and texts are short, the situation is significantly different from, for example, large software packages which need to be localized in the new version when localization of the previous version already exists. In this case, it is sensible to take into account previous translations. For this reason, translation memory is widely used in parts of the industry, and there are automatic checking methods that make sure that words are translated in a consistent way. The consistent and high -quality use of terminology can be supported by terminology databases.

Machine translation is not yet used on a broad scale but it is being employed in certain applications. Machine translation comes into play when raw translations are needed very fast. In the European Union, for example, raw translations are being used as a means to facilitate internal communication.

[30] Source: Common Sense Advisory [CSA].

When material is being presented externally, however, translation is either performed completely by hand, or the machine translation output is revised by a translator.

Technologies like TM and terminology databases not only reduce cost but also improve the quality and consistency of the translation.

It is important to understand that translation memory and terminology databases do not only increase the effectiveness of a translator but also the quality and consistency of the work. It has been noted that machine translation is being used as a raw translation input to the human translation in order to improve efficiency. The main advantage of machine translation lies, however, in the rapid translation of texts that need to be understood for some purpose of work and where the translation cost or, more often, the translation turnaround time prohibits the use of other methods. Such up-to-the-minute online content that needs to be translated at short notice and in rather small quantities (for example newswire) is growing rapidly.

Interview with Kevin Bolen, Chief Marketing Officer, Lionbridge

Kevin Bolen
Chief Marketing Officer
Lionbridge
Waltham, MA, USA

Kevin Bolen leads Lionbridge's global marketing function responsible for setting market direction, solutions management, competitive positioning, and brand promotion. Mr. Bolen joined Lionbridge in 2005 through the acquisition of Bowne Global Solutions where he had served as the VP of Marketing since joining in early 2002. In addition to leading the global marketing function, Mr. Bolen also ran sales for the Eastern Region of the Americas. Prior to Bowne, he led the marketing function at LexiQuest, Inc. an enterprise software developer specializing in natural language technology. Before joining LexiQuest, Mr. Bolen spent six years with IBM Global Services where he held a variety roles spanning consulting, corporate development, and marketing. His last position was as a Senior Marketing Manager responsible for worldwide solutions development and strategic planning for the Retail, Packaged Goods, Transportation, and Industrial sectors.

Mr. Bolen holds a BBA in International Business from Pace University and an MBA in Marketing and Management from the Stern School of Business at New York University.

In a nutshell, could you describe the localization market?

The localization market is highly fragmented with thousands of small competitors competing on a regional, vertical or functional level. Some offer the full range of localization support while others offer pure translation service, often as a sub-contractor to the larger providers. Several of the larger firms also offer complementary services like authoring, creative design, and interpretations or supporting software to manage the localization process and assets.

What are the main market drivers, also concerning market growth?

Companies continue to expand globally to reach new consumers or to reduce their operating costs by moving work to lower cost regions. Entering these new markets requires localization of the product, the marketing and other packaging content, the support or regulatory documentation, sales and customer training materials, and employee communications.

Consumer-driven markets like China and India are pushing localization for a wide variety of products and services into languages and cultures not traditionally served by the legacy enterprise thus fueling new revenue opportunities for the service providers. On the other side, product cycles are accelerating meaning more features and more information is delivered more frequently. This generates new localization opportunities but requires the service providers to reevaluate their production models as turn-around time and file transfers become highly sensitive in the race to beat the competition.

What are the most relevant technologies used in human language translation?

Translation memory (TM) remains the leading technology in terms of efficient production. However, traditional licensing models and desktop-level applications are becoming obsolete in the era of on-demand, web-enabled software models. TM via the web allows multiple translators to work simultaneously on a single project, leveraging the work done by each instantly to improve the collective performance.

Terminology tools are also a key support element as they enable greater consistency across a wide array of content and production teams. The ability to embed this language control into the production process ensures accuracy while accelerating production times, a key issue in the on-demand world.

What is your experience with machine translation?

Lionbridge, through its acquisition of Bowne Global Solutions, owns one of the industry's most respected rules-based MT engines, "Barcelona". Though statistical and example based systems have shown promise in research, their corpus of available languages and domains remains too limited to be of practical commercial use in today's marketplace. Lionbridge has successfully utilized its Barcelona engine on a variety of customer projects serving to reduce the time and cost associated with the translation element of an overall localization project. These projects must be carefully scoped as the system requires some degree of custom dictionary and rules development before the resulting quality is sufficient to allow the work to be reviewed by a post-editor as opposed to a full translator. A sufficient volume of words is required to ensure a positive ROI on the initial customization expense.

High-quality localization and the provision of low-cost automatic translation aim at different markets. Do you intend to also cover the "disruptive technology" end?

Lionbridge already offers a free website for MT-based translation and has enabled clients to utilize MT to translate content they would not have otherwise addressed due to the cost associated with HT. In the future, as MT becomes a more viable alternative, clients will have a variety of solution levels from which to choose, however, quality is very subjective. It is our belief the clients are seeking a more integrated solution leveraging web-based TMs, MT, and HT/post-editing to deliver HT quality faster and at a lower cost than today's TM/HT model alone. This would allow for expansion into more markets faster and with a better ROI.

4.3. Research on Speech to Speech Translation and its Component Technologies

Human language, both in its spoken and its written form, has been worked on in scientific research in recent decades by thousands of researchers worldwide. Let's look into the different technologies.

Automatic speech recognition, the creation of a written word sequence from a spoken word sequence, has proven to be a very difficult scientific problem, indeed much harder than playing chess at the level of human performance. The research community has been successful in successively treating first the simpler and then the more difficult problems. The technology initially dealt with small vocabularies, speaker dependent systems in quiet environments and isolated words input, but has now advanced to the level of very large vocabulary, speaker-independent connected-speech recognition. While public awareness of the technology remains low, the task of text dictation in a professional setting is now often supported by speech recognition functionality, although recognition errors still occur. The research community has since focused on the next challenge, namely speech from speakers who are not cooperative. Not cooperative in this sense means that the speaker does not speak with the intention of being recognized by an automatic system. Typical research scenarios are the recognition of conversation between people in natural, unconstrained speech. Speech recognition is now often one component of a multimodal perceptual interface[31].

Text-to-speech (TTS), also called speech synthesis, generates speech from written text. While TTS systems have been intelligible for quite a few years, the last major break-through of "concatenative systems using the unit selection method" led to the availability of more naturally sounding systems only a few years ago. TTS systems play an important role in making human-computer communication mobile by the introduction of spoken language systems. While prerecorded speech can be used for essentially static content, concatenating word groups into sentences, this is not possible with varying content, and TTS becomes a necessity. Current effort in text-to-speech development is focused on aspects such as the fast generation of different voices and research into the generation of emotion and of contrastive intonation.

Machine translation translates a written text from a source language into a target language. Much effort has been invested into this very difficult problem in recent decades. As with speech recognition, where a methodological paradigm shift took place around 1990 when rule-based methods were replaced by the statistical approach,

[31] Such as in the European projects CHIL (Computers in the Human Interaction Loop, http://chil.server.de/) and AMI (Augmented Multi-party Interaction, http://www.amiproject.org/).

there are again two fundamentally different approaches to machine translation. The earlier approach relies heavily on linguistic methods and the explicit coding of human knowledge, while the later one is very much data driven and exploits variants of statistical methods which have proven to be successful in the speech recognition field. For any of the approaches, state-of-the-art systems produce translations of unconstrained text which remain of poor quality compared to human performance and can only be used for raw translations which deliver the basic meaning, intermingled with mistakes. While linguistic methods tend to perform well on some sentences but completely fail on others, statistical methods spread errors more evenly, typically translating everything without complete failure but with few instances of exemplary performance. Having put the current performance into the right perspective, it should be noted that progress in this area is rapid and international competition and objective benchmarking show impressive progress, particularly on the part of data-driven systems[32].

Spoken language translation (SLT) translates speech from a source language into a text in a target language. SLT systems basically consist of a speech-

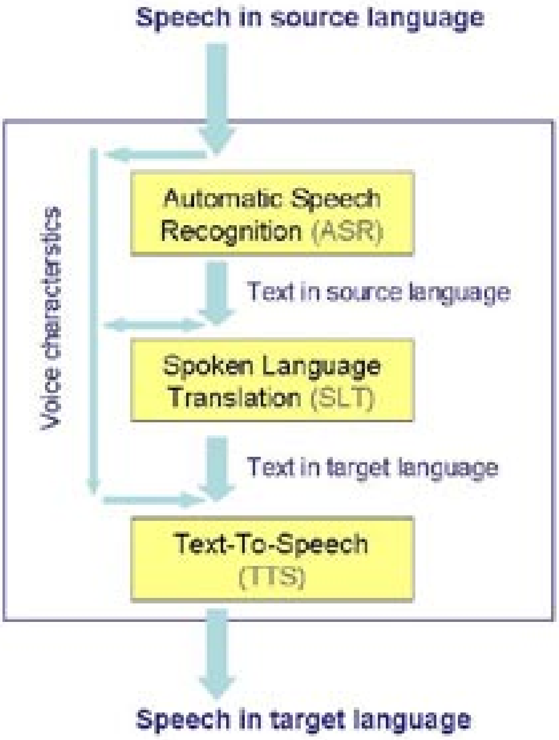

Speech in source language
Voice characteristics
Automatic Speech Recognition (ASR)
Text in source language
Spoken Language Translation (SLT)
Text in target language
Text-To-Speech (TTS)
Speech in target language

Fig. 5: Functional diagram of a *speech-to-speech translation* (SST) system. Incoming speech in the source language is converted into text by the *automatic speech recognition* (ASR) module. This text is fed into a machine translation component called *spoken language translation* (SLT) module which is specific to the fact that the text contains recognition errors and exploits the characteristics of speech in contrast to written language – ungrammatical sentences, hesitations, false starts. Its output, a text in the target language, serves as input to the *text-to-speech* (TTS) module which generates the speech in the target language. In order to preserve the original voice characteristics, additional information is passed to the TTS module.

[32] Cf. [Ney].

recognition component and machine translation component, as you might assume, but there is more to it: the text contains recognition errors which have to be taken care of. Another difficulty is that the scenarios in which researchers work consider speech which is not directly intended to be recognized by a system, i.e. speech which is uttered on television, in the course of a meeting or in a person-to-person conversation.

Speech-to-speech translation (SST) combines all the above mentioned methods. Here, speech in the source language is translated into speech in a target language, such that people can to talk with each other in different languages, using the computer as an interpreter. To achieve this, an SLT system is cascaded with a TTS system, and some additional measures are taken. First, the speaker identity should be preserved in the target language speech along with prosodic features appearing in the source speech. These features are of course not apparent in any intermediary written text form. This very difficult problem is currently being tackled by the European project TC-STAR, which works on the speech-to-speech translation of European parliamentary debates. While the task is challenging, it is interesting to note that the quality of the translations is already good enough to convey the gist of parliamentary speech in an unfamiliar source language.

4.4. The TC-STAR Project

Speech to speech translation is a challenging research task. To simplify the task and make it somewhat more feasible, the first research projects addressing the topic in the 1990's (already with spontaneous speech) worked on restricted domains such as appointment scheduling[33].

TC-STAR is the first joint research project to address speech-to-speech translation in an unrestricted domain.

It is hard to describe what a big step it is to go from a restricted to an unrestricted domain, but if you have ever successfully managed to use a tourist pocket guide to order your hotel breakfast in Portugal and then tried to listen to a parliamentary debate on a random topic, you might have an idea. TC-STAR[34] (*Technology and Corpora for Speech to Speech Translation*) is the first joint research project addressing speech-to-speech translation in an unrestricted domain. Apart from the translation of Chinese broadcasts into English, which was chosen in order to link to available international benchmarks, TC-STAR works on European parliamentary speeches in native or non-native English and Spanish. The topics include anything that is worked on in the European Parliament.

[33] The German Verbmobil project, 1993-2000, which a total funding volume of 53M undertook a large effort in this direction. The international initiative C-STAR, a large global consortium devoted to speech-to-speech translation with twenty partners, started in 1991 and is still ongoing.

[34] TC-STAR (http://www.tc-star.org) is an integrated project (IP), i.e. a large project, within the FP6 programme.

This effort focuses on advanced research in all core technologies for speech-to-speech translation (SST): speech recognition, speech translation and speech synthesis. The objectives of the project are extremely ambitious: making a breakthrough in SST research to significantly reduce the performance gap between human and machine performance. The first half of the three year project has shown very satisfying performance on this difficult task, but there is still a long way to go.

Besides the scientific work, TC-STAR has created the infrastructure needed for accelerating the rate of progress in the field. It has collected the data needed for the data driven methodology pursued in the project, and it has implemented an evaluation infrastructure based on competitive evaluation. This evaluation driven approach ensures that research progress is recognized as such and that the methods developed by the various project partners can be compared and appropriately validated. Together with a mixture of cooperation and healthy competition, this approach should maximize scientific progress. At the same time, the consortium benchmarks within the external scientific community, and project partners have reached top ranks in the recent international IWSLT evaluations[35].

[35] International Workshop on Spoken Language Translation. IWSLT 2005 in Pittsburgh, USA, http://www.is.cs.cmu.edu/iwslt2005/. IWSLT 2004 in Kyoto, Japan, http://www.slt.atr.jp/IWSLT2004/

5. The Power of an Enabling Technology

5.1. Insatiable Human Needs

So there we are: machine translation has just started to be used in practice, speech-to-speech translation is on the forefront of research, and many related technologies are in the research phase. On the other hand, there is a huge latent demand to be fulfilled, and we know that the technology will one day meet this demand. Can we see the future, even in a vague sense?

In order not to get lost in detailed market prognoses that might have to be revised every two years, let's try to rely on an argumentation identifying the relevant forces that determine demand and supply in a future market. There are some trends which are largely accepted by many specialists and non-specialists alike, such as the trends toward mobility, networking and distributed computation. If we understand the key technological and commercial drivers, the value chains (or, more precisely, the value networks) and the dynamics of the market, we can be fairly certain that our view of the future is not too inaccurate. And keep in mind a trick that makes prediction even more reliable: The best way to predict the future is to invent it[36].

When you look at the real success stories of new technologies entering the market, you find two ingredients that regularly coincide: a new technology and an old human need. Mobile telephony is a typical example, serving the insatiable human need for communication with other people and the desire to have an impact even at a distance. Almost-real-time but non-isochronous communication through small messages existed in your childhood when you were writing small paper notes at school – kids now use SMS. Or would it be more accurate to compare the SMS of today with the telegram of the past? There are similarities: in both cases, a short text is delivered via an electronic channel directly to the addressee. However, due to the enormous price difference and the very different usage scenarios, it doesn't make sense to consider SMS as a modern form of the telegram. Any attempt to predict the amount of SMS traffic today based on the number of telegrams sent twenty years ago would have failed completely. Along the same lines, it would have been misleading and highly inaccurate to estimate the current public usage of Google based on the figures of desktop paper research or research in databases twenty years ago. However, even in the past it was clear that any technology that met basic, essentially insatiable, human

[36] Quote attributed to Alan Kay, PC pioneer and user interface specialist.

needs would be embraced: people need to communicate with other people (telephone, mobile phone, e-mail, SMS, chat); people have a desire to master space (car, plane; also communication); people are thirsty for information and entertainment (internet, television).

Insatiable human needs: Communication with people, mastery of space and time, thirst for information and entertainment.

The existing market for translation and localization tells us little about what would happen if translation were available in real-time and at very reasonable cost. In very many situations, paying a Euro for a line of translated text is simply out of any reasonable relation to the added value. Dropping the cost by one or two orders of magnitude would enable a huge increase in usage of translation, even if the quality were not perfect. There is a strong human desire to communicate and to get information. Markets are constantly changing. Value migrates out of some value chains and into others. Old demands meet new possibilities, and new products and markets are born.

What would it mean if real-time translation at reasonable cost were available? What would be the effects of such translation of text and speech, of documents, web-sites, video, streaming content? What if there were no language barrier anymore?

We should not base our estimates on the translation business as it is today. The current amount of translation is just the tip of the iceberg, compared to the latent demand for translation. Let me outline what I would like to do.

I would like to watch Al Jazeera – subtitled in my language – to get an idea how the Arab world thinks. When speaking English with my Italian and Spanish peers, we could sometimes do with a little help to broaden our communication channel's bandwidth. It is hard for me to read a French document, especially when I do not know the terminology – an approximate translation into German or English would substantially speed up reading. During my vacation in Portugal, I would like to understand the menu and be able to exchange a few friendly words with the locals.

Imagine a world with real-time translation available cheaply and on the spot.

Would you like to browse a Chinese website? Cook this wonderful recipe from your vacation in Spain which, unfortunately, is in Spanish? What are your demands?

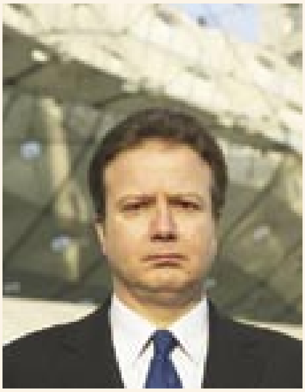
We Europeans will communicate more easily with one another, and we will come closer. Using human language technologies will provide significant economic advantage. End customer contact is being automated where possible, including the handling of multi-lingual FAQ lists – this is where HLT comes in. Much of a corporation's non-tangible assets are in the minds and files of their employees, in many languages. Getting easy access to information across language borders is a pure necessity for any globally operating company. This will be an interesting market, and companies like IBM and SAP, Yahoo and Google are preparing themselves. Real-time translation of any data source – newswire, dynamic web content, video; cross-lingual document search; easier access to foreign language markets in particular for professionals and SMEs: There are endless opportunities for Europe.

The advent of automatic translation will be supported by three strong market drivers:

- low cost
- real-time operation
- automatic processing properties

Any one of these market drivers should be strong enough to drive a business on its own.

Interview with Dimitris Sabatakakis, CEO, SYSTRAN



Dimitris Sabatakakis
CEO of SYSTRAN
Paris, France

SYSTRAN is the market leading provider of language translation software products and solutions for the desktop, enterprise and Internet that facilitate communication in 40 translation directions (20 language pairs) and in 20 domains. With over three decades of expertise and research and development, SYSTRAN's software is the choice of leading global corporations, portals and public agencies. Use of SYSTRAN products and solutions enhance multilingual communication and increase user productivity and time-savings for B2E, B2B and B2C markets as they deliver real-time language solutions for search, content management, online customer support, intra-company communications, and eCommerce.

Dimitris Sabatakakis was born in 1962 in Athens, Greece. A graduate of Strasbourg University in Economic Sciences, he began his career in finance, then in industry. Joined by investors, he took over and managed the recovery of the Gachot company, which was sold to the KEYSTONE/TYCO Group in 1995. Mr. Sabatakakis has managed SYSTRAN since February 1997.

How does the machine translation market relate to the traditional translation market?

The traditional translation market is stable and represents a human process like writing. The MT (machine translation) market is very different. The offering includes real-time web-based translation services and applications. It is practical to employ MT for multilingual publications in situations when high volumes of content must be translated combined with user control of the source text. 99% of current MT market applications are used for gisting purposes as millions of pages on portals like Yahoo! and Google are translated on a daily basis.

MT or language translation software is an automatic process which allows a user to:

- Understand foreign language content in his or her native language in real-time and at no cost
- Publish content in different languages in real-time, by way of "controlling" the source text

Do people understand the capabilities of automatic systems, do they tend to overestimate or to underestimate them?

The value of MT used for gisting is obvious. The proof is millions of translated pages every day.

The value of using MT for multilingual publication is also obvious but the significant investment required to ‘control’ or ‘structure’ the source text is often underestimated and overlooked. In order to accomplish this, the entire publication workflow must consider multilingual issues from the beginning. In practice, corporations produce content (such as product datasheets, marketing material, technical support info, knowledge bases, etc.) in one source language which is usually English and only consider the localization of content as an additional step. Due to budget restrictions, the amount of localized text is much smaller than the original text.

Also, the maintenance of localized versions of content is challenging as it is expensive, slow (a human process) and does not fully resolve inconsistent terminology problems.

Do you see MT more as a rationalization tool for human translation or more as translation of a new type?

MT is both a tool for human translation and translation of a new type. Human translators should adopt MT as it increases user-productivity and time-savings. The software has not yet attained massive reach because the ergonomics and user interface were not created explicitly for human translators. SYSTRAN has made great efforts in providing such tools so that MT can today be used within a traditional translation services environment to boost productivity.

6. Conclusion

Fast, reliable and cheap ways to communicate and to transport and process data are the backbone of a modern information society. What is true for normal information and communication technology holds for technical systems which translate from one language to another, be it for communication or for information access. For the European Union, with its twenty official languages and many more spoken languages, the availability of fast, reliable and cheap translation is a necessity, and translation technology should be considered as strategically important.

Other parts of the world need automated translation services as well, but in different contexts. They will not solve our problems; we are better off solving them ourselves. And we have the means to do it.

Due to the need for machine translation and spoken language translation in Europe, there are favorable market conditions for companies that plan to offer translation technology and services. Several strong European research groups can serve as suppliers for translation technology. Beyond the European market, the demand for machine translation will be high in Japan, China and Korea as well as in India.

The language industry is rapidly growing, but the full potential of machine translation will go beyond an efficiency improvement of current human translation. It is expected to unfold in market segments that barely exist at present and that are so low-margin that they will never be attractive for human translation services. Just as it was impossible to extrapolate SMS usage from the usage of telegrams, we have a hard time predicting the translation volume of the future. It will be very large, and it will help unfold the part of our economic potential that is still being blocked by language barriers.

All these innovations take time, but they are on their way, and they will have an enormous impact. In the meantime, let us set the course for a future that is beneficial to us. It is up to us to shape it.

7. References and Other Supplementary Information

7.1. References

[Chr] C. Christensen: The Innovator's Dilemma. When New Technologies Cause Great Firms to Fail. Harvard Business School Press, Boston, Mass., 1977.

[CRa] C. Christensen, M. Raynor: The Innovator's Solution. Creating and Sustaining Successful Growth. Harvard Business School Press, Boston, Mass., 2003.

[Cry] D. Crystal: English as a Global Language. Cambridge University Press. 2nd edition, 2003 (1st ed. in 1997).

[CSA] R. Beninatto, D. DePalma: Ranking of Top 20 Translation Companies. Common Sense Advisory, Inc., June 2005. Can be downloaded from the Common Sense Advisory website *http://www.commonsenseadvisory.com* under *http://www.commonsenseadvisory.com/members/res_cgi.php/050701_QT_top_20.php*

[DGT] The Directorate-General for Translation of the European Commission (DGT). *http://europa.eu.int/comm/dgs/translation/index_en.htm*

[EB1] Europeans and languages. Eurobarometer 54 Special. INRA Report, 60 pages, Feb. 2001. The report can be downloaded under *http://europa.eu.int/comm/education/policies/lang/languages/barolang_en.pdf* and is referred to in *http://europa.eu.int/comm/education/policies/lang/languages/index_en.html*. Note that the figures relate to the old EU of 15 nations.

[EB5] Europeans and languages. Eurobarometer 63.4, September 2005. A survey in 25 EU Member States, in the accession countries (Bulgaria and Romania), the candidate countries (Croatia and Turkey) and among the Turkish Cypriot Community. The report can be downloaded under *http://europa.eu.int/comm/public_opinion/archives/ebs/ebs_237.en.pdf* and is referred to in *http://europa.eu.int/languages/en/document/80/20*

[ELP] Languages of Europe. On the official website of the European Commission concerning the languages spoken in the EU. *http://europa.eu.int/comm/education/policies/lang/languages/index_en.html*

[Gra] D. Graddol: The Future of English? A guide to forecasting the popularity of the English language in the 21st century. (A report commissioned by The British Council). The English Company (UK), 64 p., 2000 (first published 1997), ISBN 0-86355-356-7.

[Gri] B. Grimes: Ethnologue Language Database. *http://www.sil.org/ethnologue/*The

Ethnologue is a catalogue of more than 6,700 languages spoken in 228 countries.

[Gor] R. G. Gordon (ed.): Ethnologue: Languages of the World. Fifteenth edition. Dallas, Tex., 2005, SIL International. Online version: *http://www.ethnologue.com/*

[Ney] H. Ney: One Decade of Statistical Machine Translation: 1996-2005. In Proceedings of the MT Summit X, pp. i-12 - i-17, Phuket, Thailand, September 2005.
The paper can be downloaded under *http://www-i6.informatik.rwth-aachen.de/web/Publications/index.html*

[TMC] Translating for a multilingual community. Directorate-General for Translation of the European Commission (DGT), April 2005, 19 pages. The pdf brochure can be downloaded under *http://europa.eu.int/comm/dgs/translation/bookshelf/brochure_en.pdf*

[TTW] Translation tools and workflow. Directorate-General for Translation of the European Commission (DGT), April 2005, 25 pages. The pdf brochure can be downloaded under *http://europa.eu.int/comm/dgs/translation/bookshelf/tools_and_workflow_en.pdf*

7.2. Further Reading

Common Sense Advisory *http://www.commonsenseadvisory.com/*

Europa languages portal *http://europa.eu.int/languages/en/home*

LISA – Localization Industry Standards Organization *http://www.lisa.org/*

Multilingual Computing, Inc. *http://www.multilingual.com/*

EAMT – European Association for Machine Translation *http://www.eamt.org/*

ELRA – European Language Ressources Association *http://www.elra.info/*

ELSNET – European Network in Language and Speech *http://www.elsnet.org/*

GALA – The Gloabalization and Localization Association. *http://www.gala-global.org/*

7.3. Tables

Table 3: 'Global influence' of the 12 major languages according to the engco model (see table 2). An index score of 100 represents the position of English in 1995 [Gra].

Table 4: Disciplines in which German academics claim English as their working language [Gra].

Table 5: The 20 official languages of the European Union and their abbreviations [ELP]. Erse (Irish) will become the 21st official language of the EU from 1 January 2007.

Table 6: Foreign language skills in the *EU:* Proportion of European citizens speaking the respective language in the EU (as mother tongue / as secondary or foreign language / either one). The survey was fielded in the European Union of 25 Member States and, in addition, in the accession countries (Bulgaria and Romania), the candidate countries (Croatia and Turkey) and among the Turkish Cypriot community.

Table 7: Ranking of top 20 language service providers - 2004 revenue. (The two important acquisitions of 2005, the acquisition of Bowne Global Solutions by Lionbridge and the acquisition of TRADOS by SDL, have been consolidated in the figures.)

Table 8: Examples for disruptive technological innovations, together with the corresponding established technology. – Source: [Chr], [CRa].

7.4. Figures

Fig. 1: The proportion of the world's books annually published in each language. English is the most widely used foreign language for book publication: over 60 countries publish titles in English [Gra].

Fig. 2: Languages of the world. Each dot represents the primary location of a living language listed in the Ethnologue.

Fig. 3: Sustaining and disruptive technological change. – Source: [CRa].

Fig. 4: US DoD (Department of Defense) budget for language translation technology. Fiscal year 2005 budget estimates for research, development, test and evaluation (RDT&E), defense-wide, at DARPA (Defense Advanced Research Projects Agency). – Source: DARPA.

Fig. 5: Functional diagram of a speech-to-speech translation (SST) system.

7.5. List of Acronyms

AFNLP	Asian Federation of Natural Language Processing
ASR	automatic speech recognition
BLEU	(a statistical quality measure for translations which correlates with human judgment; the higher the figure the better)
CEC	Commission of the European Community
DARPA	defense advanced research projects agency
DGT	directorate general for translation
DTMF	dual tone multiple frequency
ELRA	European Language Resources Association
EU	European Union
GALE	global autonomous language exploitation
HLT	human language technology / human language technologies
HT	human translation
IP	*here:* integrated project (in other contexts: internet protocol; intellectual property)
IR	information retrieval
IT	information technology
IVR	interactive voice response
LISA	localization industry and standards association
MT	machine translation
NLP	natural language processing
ROI	return on investment
SLT	spoken language translation
SME	small or medium-sized enterprise
SMS	short message service
SST	speech-to-speech translation
TC-STAR	technology and corpora for speech-to-speech translation
TM	translation memory
TTS	text-to-speech

Finito di stampare
nel mese di maggio 2006
dalla Tipolitografia TEMI - Trento