META-NET

Languages in the European Information Society

Dr. John Doe (DFKI), Prof. Dr. Felix Sasaki (DFKI), Pepe Pérez (Academia de LaTeX), Tan Ah Kao (USEV)

October 5, 2011

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under contracts T4ME (Grant Agreement 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

Preface

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses educators, jour- nalists, politicians, language communities and others. The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technolo- gies also differ for each language. The required actions depend on many factors, such as the complexity of a given language and the size of its community. META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies. This analysis focused on the 23 official European languages as well as other important national and re- gional languages in Europe. The results of this analysis suggest that there are many significant research gaps for each language. A more detailed expert analysis and assessment of the current situation will help maximise the impact of additional research and minimize any risks. META-NET consists of 47 research centres from 31 countries that are working with stakeholders from commercial businesses, gov- ernment agencies, industry, research organisations, software com- panies, technology providers and European universities. Together, they are creating a common technology vision while developing a strategic research agenda that shows how language technology applications can address any research gaps by 2020.

Contents

1 Executive Summary — Zusammenfassung

Many European languages run the risk of becoming victims of the digital age because they are underrepresented and under-resourced online. Huge regional market opportunities remain untapped today because of language barriers. If we do not take action now, many European citizens will become socially and economically disadvan- taged because they speak their native language. Innovative language technology is an intermediary that will enable European citizens to participate in an egalitarian, inclusive and economically successful knowledge and information society. Multi- lingual language technology will be a gateway for instantaneous, cheap and effortless communication and interaction across lan- guage boundaries. Today, language services are primarily offered by commercial pro- viders from the US. Google Translate, a free service, is just one example. The recent success of Watson, an IBM computer system that won an episode of the Jeopardy game show against human candidates, illustrates the immense potential of language technolo- gy. As Europeans, we have to ask ourselves several urgent ques- tions:

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

This whitepaper for the German language demonstrates that a lan- guage technology research and development community exists in Germany, Austria and Switzerland. However, many large compa- nies have

stopped their activities in language technology research and development, which is nowadays almost exclusively performed by small and medium enterprises that can hardly compete on the global market. While Germany used to be a European hub in this area in the past, we now have to ask ourselves if we want to actively compete in the global market for research and development in this future technology or not. Although a number of technologies and resources for Standard German exist, there are fewer technologies and resources for the German language than for the English language. The existing technologies and resources are also poorer in quality. According to the assessment detailed in this report, immediate action must occur before any breakthroughs for the German lan- guage can be achieved.

2 Risk for Our Languages and a Challenge for Language Technology

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

We are witnesses to a digital revolution that is dramatically impact- ing communication and society. Recent developments in digital information and communication technology are sometimes com- pared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information socie- ty and our languages in particular?
After Gutenberg's invention, real breakthroughs in communication and knowledge exchange were accomplished by efforts such as Luther's translation of the Bible into vernacular language. In sub- sequent centuries, cultural techniques have been developed to bet- ter handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citi- zens to communicate within certain (often political) boundaries;

- the teaching and translation of languages enabled exchanges across languages;
- the creation of editorial and bibliographic guidelines assured the quality and availability of printed material;
- the creation of different media like newspapers, radio, televi- sion, books, and other formats satisfied different communica- tion needs. In the past twenty years, information technology has helped to automate and facilitate many of the processes:
- desktop publishing software has replaced typewriting and type- setting;
- Microsoft PowerPoint has replaced overhead projector trans- parencies;
- e-mail send and receive documents faster than a fax machine;
- Skype offers cheap Internet phone calls and hosts virtual meet- ings;
- audio and video encoding formats make it easy to exchange multimedia content;
- search engines provide keyword-based access to web pages;
- online services like Google Translate produce quick, approxi- mate translations;
- social media platforms such as Facebook, Twitter, and Google+ facilitate communication, collaboration, and information shar- ing.

Although such tools and applications are helpful, they are not yet capable of supporting a sustainable, multilingual European society for all where information and goods can flow freely.

2.1 Language Borders Hinder the European Information Society

We cannot predict exactly what the future information society will look like. But there is a strong likelihood that the revolution in 6 communication technology is bringing people speaking different languages

together in new ways. This is putting pressure on indi- viduals to learn new languages and especially on developers to create new technology applications to ensure mutual understand- ing and access to share- able knowledge.

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

In a global economic and information space, more languages, speakers and content interact more quickly with new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter, YouTube, and, recent- ly, Google+) is only the tip of the ice- berg. Today, we can transmit gigabytes of text around the world in a few seconds before we recognize that it is in a language we do not un- derstand. According to a recent report from the European Commission, 57% of Internet users in Europe purchase goods and services in languages that are not their native lan- guage. (English is the most common foreign lan- guage followed by French, German and Span- ish.) 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the Web.i A few years ago, En- glish might have been the lingua fran- ca of the Web — the vast majority of content on the Web was in Eng- lish — but the situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has ex- ploded. Surprisingly, this ubiquitous digital divide due to language borders has not gained much pub- lic attention; yet, it raises a very pressing question: Which European languages will thrive in the networked information and knowledge society, and which are doomed to dis- appear?

2.2 Our Languages at Risk

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies and Applications)							
Speech Recognition	5	1	4	4	4	3	3
Speech Synthesis	5	3	4	5	4	3	3
Text analysis	4	2,5	4	4	4	2,5	2,5
Text interpretation	2	2	3	2	2	2	1
Language generation	2	1	2	2	2	1	2
Machine translation	5	3	2	3	4	1	2
Language Resources (Resources, Data and Knowledge Bases)							
Text corpora	3	2	3,5	3	4	4	2,5
Speech corpora	3	1	3	2	3	3	2
Parallel corpora	2	1	2	2	2	2	1
Lexical resources	3	2,5	3,5	2,5	4	4	2,5
Grammars	3	2	3	3	3	2	1

While the printing press helped step up the exchange of infor- mation in Europe, it also led to the extinction of many European languages. Regional and minority languages were rarely printed and languages such as Cornish and Dalmatian were limited to oral forms of transmission, which in turn restricted their scope of use. Will the Internet have the same impact on our languages? Europe's approximately 60 languages are one of its richest and most important cultural assets, and a vital part of its unique social model.[1] While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many European lan- guages could become irrelevant in a networked society. This would weaken Europe's global standing, and run counter to the strategic goal of ensuring equal participation for every European citizen regardless of language. According to

The wide variety of languages in Europe is one of its richest and most important cultural assets.

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

a UNESCO report on multilin- gualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society.[2]

2.3 Language Technology is a Key Enabling Technology

In the past, investment efforts in language preservation focused on language education and translation. According to one estimate, the European market for translation, interpretation, software localisation and website globalisation was €8.4 billion in 2008 and is ex- pected to grow by 10% per annum.iv Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use appropriate technology, just as we use technology to solve our transport, energy and disability needs among others. Digital language technology (targeting all forms of written text and spoken discourse) helps people collaborate, conduct business, share knowledge and participate in social and political debate re- gardless of language barriers and computer skills. It often operates invisibly inside complex software systems to help us:

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

- find information with an Internet search engine;
- check spelling and grammar in a word processor;
- view product recommendations in an online shop;
- hear the verbal instructions of a car navigation system;
- translate web pages via an online service.

Language technology consists of a number of core applications that enable processes within a larger application framework. The pur- pose of the META-NET language white papers is to focus on how ready

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

these core technologies are for each European language. To maintain our position in the frontline of global innovation, Eu- rope will need language technology adapted to all European lan- guages that is robust, affordable and tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

Europe needs robust and affordable language technology for all European languages.

2.4 Opportunities for Language Technology

In the world of print, the technology breakthrough was the rapid duplication of an image of a text (a page) using a suitably powered printing press. Human beings had to do the hard work of looking up, reading, translating, and summarizing knowledge. We had to wait until Edison to record spoken language – and again his tech- nology simply made analogue copies. Digital language technology can now automate the very processes of translation, content production, and knowledge management for all European languages. It can also empower intuitive language/speech-based interfaces for household electronics, machin- ery, vehicles, computers and robots. Real-world commercial and industrial applications are still in the early stages of development, yet R&D achievements are creating a genuine window of opportunity. For example, machine translation is already reasonably accurate in specific domains, and experimental applications pro- vide multilingual information and knowledge management as well as content production in many European languages. As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for highly specialised domains, and often exhibit limited perfor- mance. But there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, cultural heritage sites, edutainment packages, libraries, simulation environments and training programmes. Mobile infor- mation services, computer-assisted language learning software, eLearning environments, self-assessment tools and

plagiarism detection software are just some of the application areas where language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a further need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emo- tional responses, identify copyright infringements or track misuse.