META NET

STRATEGIC RESEARCH AGENDA FOR MULTILINGUAL EUROPE 2020

edited by the
META Technology Council

STRATEGIC RESEARCH AGENDA FOR MULTILINGUAL EUROPE 2020

edited by the
META Technology Council

What is Language Technology?

Language technologies are specialised information technologies for processing automatically the most complex information medium in our world: *human languages* – in both modalities (spoken and written language) and also in both directions (analysis and generation of language).
Language technologies are developed by experts involved in computer science, linguistics, computational linguistics and related disciplines. References: [1, 2, 3, 4]

What are common Language Technology applications?

Spell and grammar checking in text processing applications and editing tools; web search; voice dialing; interactive dialogue systems (from banking over the phone to train reservation systems to Apple's Siri); cross-lingual search in digital libraries (such as, e. g., Europeana); synthethic voices for navigation systems; recommender systems for online shops; machine translation systems such as Google Translate, etc.

What are the major topics?

Information	Communication	Translation
Access and management	Human-human; human-machine	Spoken and written
Example: Information retrieval	Example: Spoken dialogue system	Example: Document translation

The development of the META-NET Strategic Research Agenda for Multilingual Europe 2020 has been funded by the Seventh Framework Programme of the European Commission under the contract T4ME (Grant Agreement 249 119).

Version 0.9 (July 11, 2012) for public discussion – Please send feedback to this SRA to georg.rehm@meta-net.eu with the subject line "META-NET SRA: feedback" or participate in our online discussion forum at http://www.meta-net.eu/forum.

WHAT EUROPE SAYS

Czech Republic: "The META-NET project brings a significant contribution to the technological support for languages of Europe and as such will play an indispensable role in the development of multilingual European culture and society." — Prof. Ing. Ivan Wilhelm, CSc., Dr. h. c. mult. (Deputy minister for education, youth and sport)

Denmark: "If we have the ambition to use the Danish language in the technological universe of the future, an effort must be made now to maintain and further develop the knowledge and expertise that we already have. This emerges from the META-NET report with great clarity. Otherwise we run the risk that only people who are fluent in English, will profit from the new generations of web, mobile and robot technology which are up and coming." — Dr. Sabine Kirchmeier-Andersen (Director of the Danish Language Council)

Estonia: "If we do not implement the development plan for language technology or do not cooperate with other countries in the same direction, in future Estonian will [...] be marginalized in information society." — Development Plan of the Estonian Language 2011–2017

Finland: "Without languages we could not communicate. The META-NET network is a valuable support for a multilingual Europe." — Alexander Stubb (Minister for European Affairs and Foreign Trade of Finland)

France: "The META-NET Network of Excellence provides an invaluable contribution to the development of a genuine European strategy in support to multilingualism, based on existing technologies while encouraging the development of new innovative technologies. This approach aims at a better understanding between citizens and community administrations, and will facilitate the recognition of linguistic diversity, at the national and regional levels, including in the overseas French territories." — Xavier North (Délégué Général à la Langue Française et aux Langues de France)

Germany: "Europe's inherent multilingualism and our scientific expertise are the perfect prerequisites for significantly advancing the challenge that language technology poses. META-NET opens up new opportunities for the development of ubiquitous multilingual technologies." — Prof. Dr. Annette Schavan (German Minister of Education and Research)

Greece: "Further support to language technologies safeguards the presence of Greek language and culture in the digital environment, while at the same time promoting development and fostering communication among citizens within the Information Society." — George Babiniotis (Minister of Education, Lifelong Learning and Religious Affairs)

Hungary: "META-NET is making a significant contribution to innovation, research and development in Europe and to an effective implementation of the European idea." — Valéria Csépe (Deputy General Secretary of Hungarian Academy of Sciences)

Iceland: "Language technology is an essential tool in a variety of linguistic research, and supports the official Icelandic policy of promoting the national language in all aspects of communication." — Dr. Guðrún Kvaran (Chair of the Icelandic Language Council)

Ireland: "Language technology is no longer a luxury for most European languages - it is now essential to their survival as viable means of expression across the whole range of areas from business to the arts, and this is as much the case for Irish as any other European language." — Ferdie Mac an Fhailigh (CEO, Foras na Gaeilge)

Latvia: "For such small languages like Latvian keeping up with the ever increasing pace of time and technological development is crucial. The only way to ensure future existence of our language is to provide its users with equal opportunities as the users of larger languages enjoy. Therefore being on the forefront of modern technologies is our opportunity." — Valdis Dombrovskis (Prime Minister of Latvia)

Lithuania: "Conserving [Lithuanian] for future generations is a responsibility of the whole of the European Union. How we proceed with developing information technology will pretty much determine the future of the Lithuanian language." — Andrius Kubilius (Prime Minister of the Republic of Lithuania)

Malta: "It will be really unfortunate if we do not exploit the developments in technology to apply them to the linguistic field. The synergy between these two sciences has to be brought to the service of the people so that it makes our life easier and helps to break the barriers in a globalised world. Thus the technology support for the Maltese language should serve our language to be continuously cultivated, used and placed on the same level as other languages." — Dolores Cristina (Minister for Education and Employment)

Poland: "Language technologies [...] will have a growing influence on capabilities and communication models of the contemporary world as well as on the way human natural languages, such as the Polish language, take part in this process. The text data analysis, speech synthesis and speech recognition, machine translation and text summarisation are more and more present in our everyday life. For their presence to be rational and functional, for it to serve the needs of the economy, as well as the social and cultural life well, further large-scale work in this area is needed." — Prof. Michał Kleiber (President of the Polish Academy of Sciences)

Portugal: "The research carried out in the area of language technology is of utmost importance for the consolidation of Portuguese as a language of global communication in the information society." — Dr. Pedro Passos Coelho (Prime-Minister of Portugal)

Slovenia: "It is imperative that language technologies for Slovene are developed systematically if we want Slovene to flourish also in the future digital world." — Dr. Danilo Türk (President of the Republic of Slovenia)

Sweden: "High-quality language technology may be the most effective means of preserving the linguistic diversity of Europe. Being able to use all languages fully in modern society is a question of democracy. In this connection META-NET fulfils a central, even crucial, function." — Lena Ekberg (head of the Swedish Language Council)

UK: "Language technology has the potential to add enormous value to the UK economy. Without language technology, and in particular text mining, there is a real risk that we will miss discoveries that could have significant social and economic impact." — Prof. Dr. Douglas B. Kell (Research Chair in Bioanalytical Science, University of Manchester)

See http://www.meta-net.eu/whitepapers/all-quotes-and-testimonials for additional quotes and testimonials.

"The Commission will [...] work with stakeholders to develop a new generation of web-based applications and services, including for multilingual content and services, by supporting standards and open platforms through EU-funded programmes." – *A Digital Agenda for Europe* [5], p. 24

"Everybody must have the chance to communicate efficiently in the enlarged EU. This does not only affect those who already are multilingual but also those who are monolingual or linguistically less skilled.
The media, new technologies and human and automatic translation services can bring the increasing variety of languages and cultures in the EU closer to citizens and provide the means to cross language barriers. They can also play an important role to reduce those barriers and allow citizens, companies and national administrations to exploit the opportunities of the single market and the globalising economy.
Faced with the globalising online economy and ever-increasing information in all imaginable languages, it is important that citizens access and use information and services across national and language barriers, through the internet and mobile devices. Information and communication technologies (ICT) need to be language-aware and promote content creation in multiple languages." – *Multilingualism: an Asset for Europe and a Shared Commitment* [6], p. 12 f.

"The Council of the European Union [...] encourage[s] the development of language technologies, in particular in the field of translation and interpretation, firstly by promoting cooperation between the Commission, the Member States, local authorities, research bodies and industry, and secondly by ensuring convergence between research programmes, the identification of areas of application and the deployment of the technologies across all EU languages." – *Council Resolution of 21 November 2008 on a European strategy for multilingualism* [7]

"The language of Europe is translation." – Umberto Eco (1993)

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The unique multilingual setup of our multicultural European society imposes considerable challenges in political, economic and social integration, especially in the creation of the united digital information and market space targeted by Europe's Digital Agenda [5].

Of the world's 6,000 plus languages, more than 2,000 may not survive this century, and more than 4,000 have little chance of ever playing any role in the information society. Although the situation looks far better on our continent, many European languages run the risk of becoming victims of the digital age as they are under-represented in digital content and software products and under-resourced with respect to language tools and technologies.

Today huge regional market opportunities remain untapped because of language barriers [8, 9]. If no action is taken, European citizens will find that speaking their mother tongue is a severe social and economic disadvantage. While the preservation of our linguistic and cultural diversity could become a serious economic burden for an integrated European society, it could also turn into a strong competitive advantage, since the technologies needed to overcome language barriers and support languages in the digital age are key enabling technologies for the next revolution in IT.

One of the last remaining frontiers of information technology is the deep rift that still separates our rapidly evolving technological world of mobile devices, PCs and the internet from the most precious and powerful asset of mankind: the human mind – the only system capable of thought, knowledge and emotion. Although we use computers to write, telephones to chat and the web to search for knowledge, today's information technology has no direct access to the meaning, purpose and sentiment behind our trillions of written and spoken words. Language technology will bridge this rift through sophisticated technologies for understanding. Spectacular recent innovations powered by language technology such as Google's web search, Autonomy's text analytics, Nuance's speech tools, free online translation services, IBM Watson's question answering and Apple's personal assistant Siri have given us but a first glimpse of the massive potential behind this emerging key technology. Today's computers cannot understand texts and questions well enough to provide reliable answers, translations and summaries, but in less than ten years from now, such services will be offered for many languages. Technological mastery of human language will enable a host of innovative IT products and services in business administration, commerce, government, education, health care, entertainment, tourism and many other sectors of our life.

With its special needs, experience, potential and markets, Europe is the most appropriate place for progress in this technology area. Europe has half a billion citizens speaking more than 40 European languages and many non-European ones as their mother tongue, with more than 2,500 small and medium sized enterprises in language, knowledge and interface technologies, and more than 5,000 enterprises providing language services that can be improved and extended by technology. It also has a longstanding research tradition at more than 800 centres of scientific and technological research on all European and economically relevant non-European languages.

The European LT community is dedicated to fulfilling the technology requirements of Europe's multilingual society and turning these needs and the emerging business opportunities into competitive advantages for our economy. Recognizing Europe's exceptional demands and opportunities, 60 leading IT research centres in 34 European countries with a strong track record in language technology joined forces in META-NET, a European Network of Excellence dedicated to the technological foundations of the multilingual European society in the digital age.

META-NET assembled the Multilingual Europe Technology Alliance (META) with more than 600 organisations and experts representing stakeholders such as industries that provide or use language technologies, professional associations, public administrations and language communities. Working together with numerous additional stakeholder organisations and experts from a variety of fields, META (including META-NET) has developed this Strategic Research Agenda. The plan is based on a shared vision and a thorough planning process involving a total of more than 1,200 experts and stakeholders.

In the first four chapters, we analyse the multilingual technology needs that arise from the multicultural setup of our continent and its emerging single digital market. Some of the new findings come from 30 comprehensive Language White Papers compiled by META-NET, each of which summarises the state of one European language with respect to its role and vitality in the digital age. We also survey the major trends in information technology with respect to their relevance for language technology.

Chapter 5 summarises our shared vision of the role of language technology in the year 2020 in non-technical terms. In line with many widely cited and respected forecasts, we predict that a new generation of information technology will be able to deal with human language, knowledge and emotion in competent and meaningful ways. These essential new competencies will enable an endless stream of novel services that will improve understanding. Many services will help people to learn and understand the world including history, technology, economy and nature. Others will help us to better understand each other across language and knowledge boundaries. These new capabilities will also enable many other IT services including programs for commerce, personal assistance, and enable robots and appliances to better understand what their human users want and need without even knowing.

In Chapter 6 five action lines for large-scale research and innovation have been identified:

- Three priority themes connected with powerful application scenarios that can drive research and innovation. These will demonstrate novel technologies in attractive show-case solutions of high economic impact. At the same time they will open up numerous new business opportunities for European language-technology and -service providers.
- A steadily evolving system of shared, collectively maintained interoperable core technologies and resources for the languages of Europe (and selected economically relevant languages of its partners). These will ensure that all of our languages will be sufficiently supported and represented in the next generations of IT solutions.
- The creation of a pan-European language technology service platform for supporting research and innovation by testing and showcasing research results, integrating various services even including professional human services. This showcase platform will allow SME providers to offer component and end-user services, and share and utilise tools, components and data resources.

The three solution scenarios are:

- **Translation Cloud** – generic and specialised federated cloud services for instantaneous reliable spoken and written translation among all European and major non-European languages.
- **Social Intelligence** – understanding and dialogue within and across communities of citizens, customers, clients and consumers to enable e-participation and more effective processes for preparing, selecting and evaluating collective decisions.
- **Socially Aware Interactive Assistants** – socially aware pervasive assistants that learn and adapt and that provide proactive and interactive support tailored to specific situations, locations and goals of the user through verbal and non-verbal multimodal communication.

These priority themes have been designed with the aim of turning our joint vision into reality and letting Europe benefit from a technological revolution that will overcome barriers of understanding between people of different languages, between people and technology, and between people and the accumulated knowledge of mankind. The three research priority themes connect societal needs with LT applications and concrete roadmaps for the organisation of research, development and scientific innovation. The themes are contextualized in the advanced networked society and cover the main functions of language: storing, sharing and using information and knowledge, and improving social interaction among humans and enabling social interaction between humans and technology. As multilingualism is at the core of European culture and becoming a global norm, one theme is devoted to overcoming language barriers.

The SRA proposes ways in which research and innovation need to be organised in order to achieve the targeted breakthroughs and to benefit from the immense economic opportunities they create. Core components of the strategy are novel modes of large-scale collective research and interaction among the major stakeholder constituencies: researchers in several disciplines, technology providers, technology users, policy makers and language communities. They include effective schemes for sharing resources such as data, computational language models, and generic base technologies.

Of central importance is a rapid and effectual flow of intermediate results into profitable solutions of societal impact contributing to the fertile culture of technological, social and cultural innovation targeted by the Digital Agenda [5] as well as the programmes Horizon 2020 [10] and Connecting Europe Facility (CEF) [11].

We believe that this Strategic Research Agenda has the potential to become the starting point and compass for the envisaged massive cooperation of the contributing communities that is needed to turn our shared vision into reality. The Multilingual Europe Technology Council welcomes all comments and suggestions that can help us to reach this ambitious goal.

1 INTRODUCTION

During the last 60 years, Europe has become a distinct political and economic structure. Culturally and linguistically it is rich and diverse. However, from Portuguese to Polish and Italian to Icelandic, everyday communication between Europe's citizens, enterprises and politicians is inevitably confronted with language barriers. The EU's institutions spend about *one billion Euros* a year on maintaining their policy of multilingualism [12], i. e., translating texts and interpreting spoken communication.

The European market for translation, interpretation, software localisation and website globalisation was estimated at 8.4 billion Euros in 2008. Are these expenses necessary? Are they even sufficient? Despite this high level of expenditure, the actual documents translated represent only a fraction of the information available to the whole population in countries with a single predominant language, such as the USA, China or Japan.

Language technology and linguistic research can significantly contribute to removing linguistic barriers. Combined with intelligent devices and applications, language technology can help Europeans talk and do business together even if they do not speak a common language.

The economy benefits from the European single market. For example, in 2010, trade within the European Union accounted for 60.3% of German exports and with other European countries totalled another 10.8%. But language barriers can bring business to a halt, especially for SMEs who do not have the financial means to compete on a European (or a global) level. The only (unacceptable) alternative to a multilingual Europe [13] would be to allow a single language to take a predominant position and replace all other languages in transnational communication. Another way to overcome language barriers is to learn foreign languages, and language technologies can play a key role in this.

But given the multitude of European languages (23 official languages and 60 or more others), language learning on its own just cannot solve the problem of cross-border communication. Without technological support such as machine translation, European linguistic diversity will be an insurmountable obstacle for the entire continent. Only about half of the 500 million people who live in the European Union speak English – it is evident that there is no such thing as a lingua franca shared by the vast majority of the population of our continent.

In addition, less than 10% of the EU's population accept to use online services in English which is why multilingual services based on robust and high-quality language technologies are badly needed to support and to unify the EU online market – from more than 20 country- or language-specific sub-markets to one unified single digital market with more than 500 million users and consumers. The main idea, foreseen in the Digital Agenda EU policy framework [5], is to build a single digital market in which content and services can flow freely. In order to support cross-border exchanges between users, consumers, countries and regions, robust and high-quality cross- and multilingual language technologies need to be developed urgently. In fact, according to the Digital Agenda [5] the current situation with "many fragmented markets" is one of the main obstacles that seriously undermine Europe's efforts to exploit ICT to their fullest!

Language technology is a key enabling technology for sustainable, cost-effective and socially beneficial solutions to language barriers. Language technologies will offer European stakeholders tremendous advantages, not only within the common European market, but also in trade relations with non-European countries, especially emerging economies. Language technology solutions will eventually serve as a unique bridge between Europe's languages. One important prerequisite to develop these solutions, is to carry out a systematic survey of the linguistic particularities of all European languages, and the current state of language technology support for them. With the publication of the META-NET White Paper Series "Europe's Languages in the Digital Age" [14] this important step has now been taken.

As early as the late 1970s, the EU realised the profound relevance of language technology as a driver of European unity, and began funding its first research projects, such as EUROTRA. After a longer period of sparse funding on the European level [15, 16], the European Commission set up a department dedicated to language technology and machine translation a few years ago. In recent years, the EU has been supporting language technology projects such as EuroMatrix and EuroMatrix+ (since 2006) and iTranslate4 (since 2010), which use basic and applied research to generate resources for establishing high-quality solutions for all European languages.

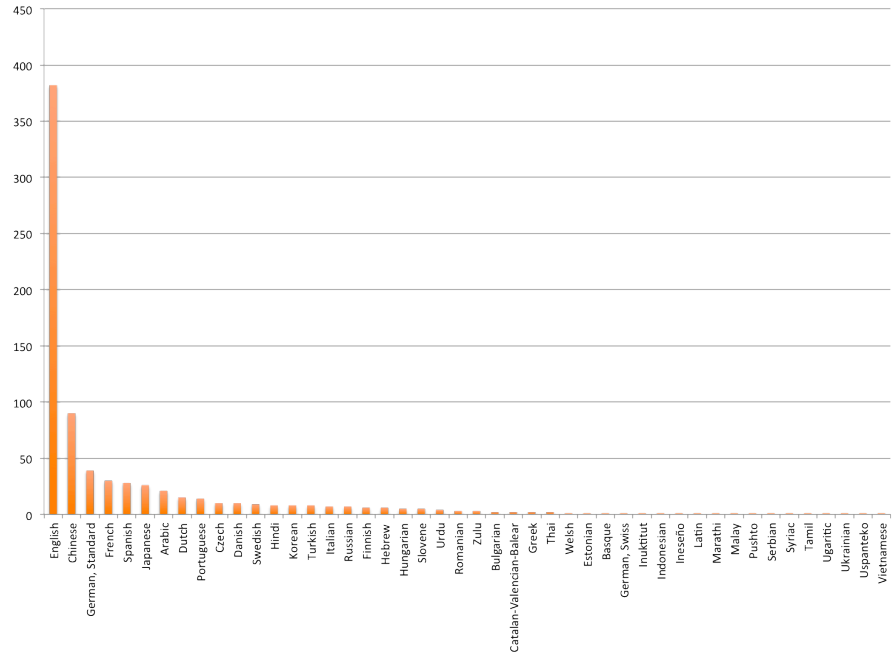
These selective funding efforts have led to a number of valuable results. For example, the translation services of the European Commission now use the Moses open source machine translation software, which has been mainly developed in European research projects. However, these projects never led to a concerted European effort through which the EU and its member states systematically pursue the common goal of providing technology support for all European languages. Figure 1 depicts the languages that have been studied by Language Technology researchers in 2010, taking into account major conferences and journals. It illustrates how technology research has focussed mainly on English followed by Chinese, German, French, and a few other bigger languages. Many European languages are without any reference whatsoever, e. g., Slovak, Maltese, Lithuanian, Irish, Albanian, Croatian, Macedonian, Montenegrin, Romansh, Galician, Occitain, or Frisian.

Research activities have tended to be isolated, delivering valuable results, often failing to make a decisive impact on the market. Even worse, in many cases research funded in Europe eventually bore fruit outside Europe. Google and Apple have been noteworthy beneficiaries. In fact, many of the predominant actors in the field today are privately-owned for-profit enterprises based in North America.

Statistical methods have been succesful, scaling to many languages and application domains but in many cases reached a performance plateau and inclusion of and combination with deep linguistic methods and insights is seen as a promising a way forward.

In the pure statistical approach, sentences are automatically translated by comparing each new sentence against thousands of sentences previously translated by humans; the quality of the output largely depends on the size and quality of the available data. While the automatic translation of simple sentences in languages with sufficient amounts of available textual data can achieve useful results, statistical methods are likely to fail in the case of languages with a much smaller body of sample data or in the case of new sentences with complex structures. Analysing the deeper structural properties of languages is a promising avenue if we want to build applications that perform well across the entire range of European languages.

Europe now has a well-developed research base. Through initiatives like CLARIN and META-NET the research community is well-connected and engaged in a long term agenda that aims gradually to strengthen language technology's role. Yet at the same time, our position is worse compared to other multilingual societies. Despite fewer



1: Languages treated in research published in the 2010 edition of the Journal of Computational Linguistics and the conferences of ACL, EMNLP and COLING

financial resources, countries like India (22 official languages) and South Africa (11 official languages) have set up long-term national programmes for language research and technology development.

What is missing in Europe is a lack of awareness and of political determination and courage that would take us to a leading position in this technology area through a concerted funding effort, a major dedicated push.

Drawing on the insights gained so far, today's hybrid language technology mixing deep processing with statistical methods should be able to bridge the gap between all European languages and beyond. In the end, high-quality language technology will be a must for all of Europe's languages for supporting the political and economic unity through cultural diversity.

Language technology can help tear down existing barriers and build bridges between Europe's languages. In the digital age, communication with people and machines, as well as the unrestricted access to the knowledge of the world should be possible for all languages.

The European LT community is dedicated to fulfilling the technology demands of the multilingual European society and to turn these needs and the emerging business opportunities into competitive advantages for our economy. To this end, we have developed this Strategic Research Agenda based on a shared vision and careful planning involving the major stakeholder communities.

In the first chapters we analyse the multilingual technology needs arising from the multicultural setup of our continent with its emerging single digital market. We also discuss the current state of technologies for European languages and the situation of the provider industries. The two core chapters of this document summarize our shared vision of the role of language technology in the year 2020 in non-technical terms (Chapter 5, p. 28 ff.) and outline three priority themes for large-scale research and innovation (Chapter 6, p. 37 ff.):

1. **Translation Cloud** – Services for instantaneous reliable spoken and written translation among all European and major non-European languages
2. **Social Intelligence and e-Participation** – understanding and dialogue within and across communities of citizens, customers, clients, consumers
3. **Socially Aware Interactive Assistants** – analysis and synthesis of non-verbal, speech and semantic signals

These thematic directions have been designed with the aim of turning the joint vision into reality and to letting Europe benefit from a technological revolution that will overcome barriers of understanding between people of different languages, between people and technology and between people and the accumulated knowledge of mankind.

The three priority research themes build the bridge between societal needs, LT applications, and concrete roadmaps for the organization of research, development and scientific innovation. The priority themes are contextualized in the networked information society and cover the main functions of language: storing, sharing and using of information and knowledge, as well as improving social interaction among humans and enabling social interaction between humans and technology. As multilingualism is at the core of European culture and becoming a global norm, one theme is devoted to overcoming language barriers.

We also present ways in which research and innovation need to be organized, in order to achieve the targeted breakthroughs and to benefit from the immense economic opportunities they create. Core components of the sketched strategy are novel modes of large-scale collective research and interaction among the major stakeholder constituencies: research in several disciplines, technology providers, technology users, policy makers and language communities. Effective schemes for sharing resources such as data, computational language models, and generic base technologies are also an integral part of the designed strategy. Of central importance is a rapid and effectual flow of intermediate results into commercially viable solutions of societal impact contributing to the fertile culture of technological, social and cultural innovation targeted by the Digital Agenda [5] and the programmes Connecting Europe Facility (CEF) [11] and Horizon 2020 [10].

The three priority research themes presented in this Strategic Research Agenda are mainly aimed at the programme Horizon 2020 which is foreseen to run from 2014 until 2020. The more infrastructural aspects, platform design and implementation and concrete language technology services are aimed at the programme Connecting Europe Facility. Our suggestion for integrating multilingual technologies into the wider CEF framework is to develop innovative solutions that enable providers of online services to offer their content and services in as many EU languages as possible, in a most cost effective way. These services are to include public services (e. g., eGovernment, eHealth, eCulture and open data portals), commercial services and user-generated content. An integral component of our strategic plans are the member states and associated countries: it is of utmost importance to set up, under the overall umbrella of our SRA and priority research themes, a coordinated initiative both on the national (member states, regions, associated countries) and international (EC/EU) level, including research centres as well as small, medium and large enterprises who work on or with language technologies. Only through close cooperation and tightly coordinated collaboration can we realise the ambituous plan of researching, designing, developing and putting into practice a European platform that supports all citizens of Europe and beyond by providing, among others, sophisticated services for communication across language barriers.

2

MULTILINGUAL EUROPE: FACTS, CHALLENGES, OPPORTUNITIES

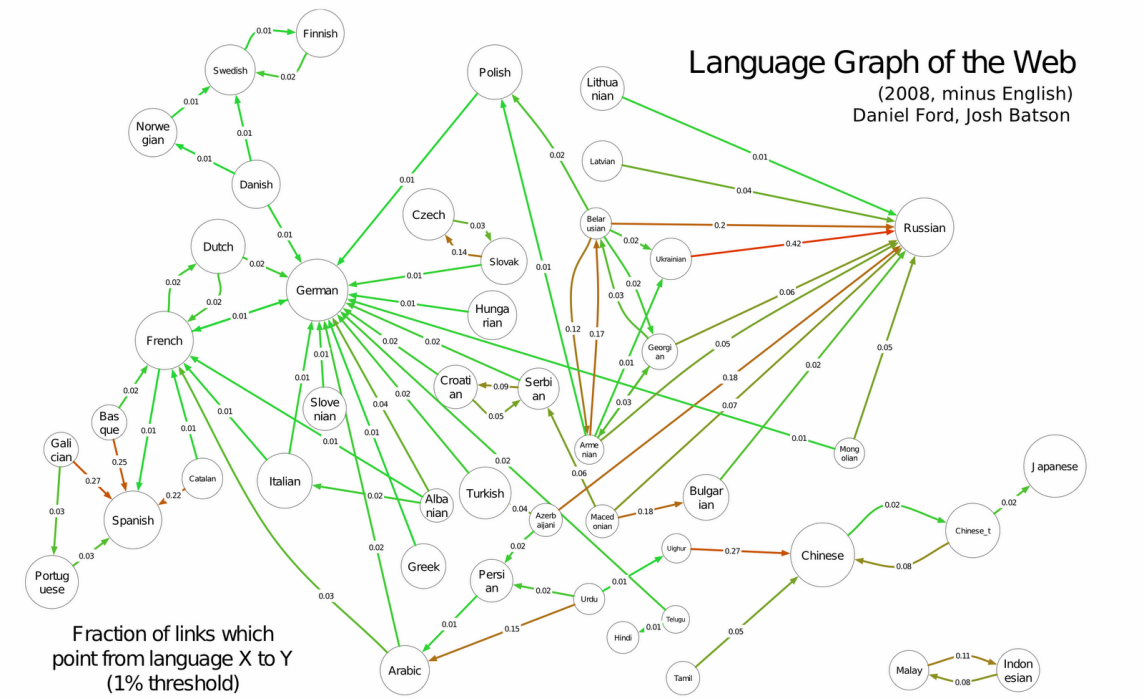
2.1 EUROPE'S LANGUAGES IN THE NETWORKED SOCIETY

Europe's more than 80 languages are one of its richest and most important cultural assets, and a vital part of its unique social model [6]. While languages such as English and Spanish are likely to thrive in the emerging digital marketplace, many European languages could become marginal in a networked society. This would weaken Europe's global standing, and run counter to the goal of ensuring equal participation for every European citizen regardless of language. A recent UNESCO report on multilingualism states that languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society [17]. From the very beginning, Europe had decided to keep its cultural and linguistic richness and diversity alive during the process of becoming an economic and political union. For maintaining the policy of multilingualism, the EU's institutions spend about one billion Euros a year on translating texts and interpreting spoken communication. For all European economies the translation costs for compliance with the laws and regulations are much higher.

A single European market that secures wealth and social well-being is possible, but linguistic barriers still severely limit the free flow of goods, information and services. With the increased number of EU members and the general trend towards timely trans-border interaction, everyday communication between Europe's citizens, within business and among politicians is more and more becoming confronted with language barriers. Many Europeans find it difficult to interact with online services and participate in the digital economy. According to a recent study, 57% of internet users in Europe purchase goods and services in languages that are not their native language (English is the most common foreign language followed by French, German and Spanish). 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the web [18]. A few years ago, English might have been the lingua franca of the web – the vast majority of content on the web was in English – but the situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has exploded. Already today, more than 55% of web-based content is not in English.

The fragmentation of languages on the web is highlighted by a study carried out by Google [19]. Figure 2 shows cross-lingual links excluding the English language, demonstrating that many European languages are practically isolated on the web. Figure 3 shows the European language communities of Twitter: the map was created by identifying automatically the languages millions of tweets are written in and placing them onto a map using their GPS-coordinates [20]. To a large degree the resulting map replicates Europe's language borders – and barriers.

Surprisingly, this ubiquitous digital divide due to language borders and language barriers has not gained much



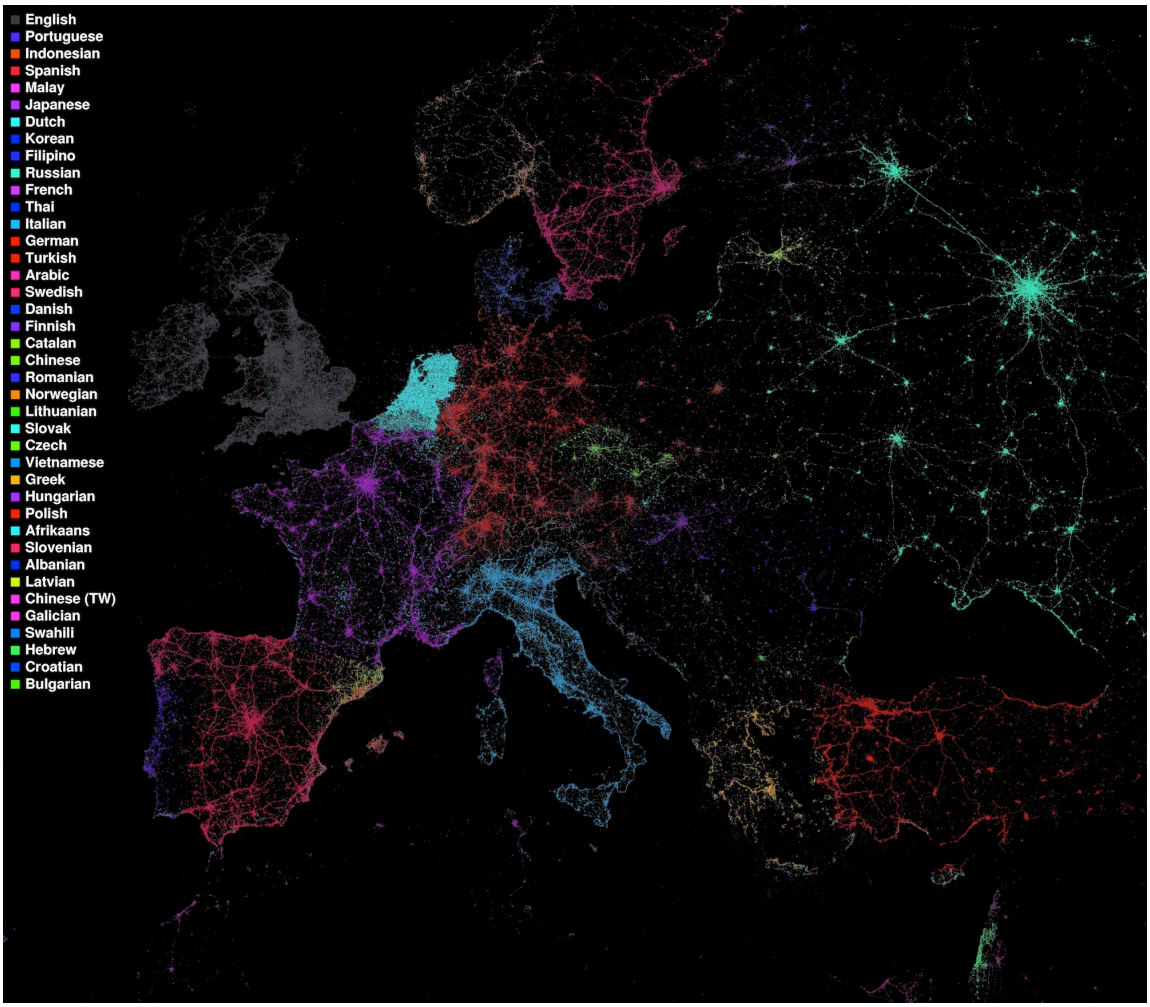
2: Language graph of the web

public attention. Yet, it raises a very pressing question: Which European languages will thrive in the networked information society, and which are doomed to disappear?

The European market for translation, interpretation and localisation was estimated to be 5.7 billion Euros in 2008. The subtitling and dubbing sector was at 633 million Euros, language teaching at 1.6 billion Euros. The overall value of the European language industry was estimated at 8.4 billion Euros and expected to grow by 10% per year, i. e., resulting in an approx. 16.5 billion Euros in 2015 [8]. Yet, this existing capacity is not enough to satisfy current and future needs, e. g., with regard to translation [21]. Already today, Google Translate translates about the same volume per day that all human translators on the planet translate in one year [22].

Despite recent improvements, the quality, usability and integration of machine translation into other online services is far from what is needed. If we rely on existing technologies, automated translation and the ability to process a variety of content in a variety of languages – a key requirement for the future internet – will be impossible. The same applies to information services, document services, media industries, digital archives and language teaching. There is an urgent need for innovative technologies that help save costs while offering faster and better language services to the European citizen.

The most compelling solution for ensuring the breadth and depth of language usage in the Europe of tomorrow is to use appropriate technology. Still, despite recent improvements, the quality and usability of current technologies is far from what is needed. META-NET has conducted an analysis on the current state of the official EU languages as well as other important European languages with special emphasis on their language technol-



3: Language communities of Twitter (European detail)

ogy support. The result of this analysis is published in a series of white papers [14] showing that, already today, especially the smaller European languages suffer severely from under-representation in the digital realm. Moreover, there are tremendous deficits in technology support and significant research gaps for all languages. For example, machine translation support for 23 out of the studied 30 languages was evaluated as having limited quality and performance, which is an alarming result!

2.2 HOW CAN LANGUAGE TECHNOLOGY HELP?

One way to overcome language barriers is to learn foreign languages. Yet without technological support, mastering the 23 official languages of the EU and some 60 other European languages is an insurmountable obstacle for Europe's citizens, economy, political debate, and scientific progress. The solution is to build key enabling technologies: language technologies (LT) will offer all European stakeholders tremendous advantages, not only within the single market, but also in trade relations with

non-European countries, especially emerging economies. Language technologies will eventually serve as the bridge between Europe's languages.

Language technology is a key enabling technology for the knowledge society. LT supports humans in everyday tasks, such as writing e-mails, searching for information online or booking a flight. It is often used behind the scenes of other software applications. We benefit from language technology when we

- use spelling checkers in a word processor;
- check product recommendations in an online shop;
- hear the spoken instructions of a navigation system;
- translate web pages with an online service.

Several popular language technology services are provided by American companies, some of them free of charge. The recent success of Watson, an IBM computer system that won against human candidates in the game show Jeopardy, illustrates the immense potential of language technology. As Europeans, we urgently have to ask ourselves a few crucial questions:

- Can we afford our information, communication and knowledge infrastructure to be highly dependent upon monopolistic services provided by US companies?
- What is Europe's fallback plan in case the language-related services provided by US companies that we rely upon are suddenly switched off?
- Are we actively making an effort to compete in the global landscape for research and development in language technology?
- Can we expect third parties from other continents to solve our translation and knowledge management problems in a way that suits our specific communicative, societal and cultural needs?
- Can the European cultural background help shape the knowledge society by offering better, more secure, more precise, more innovative and more robust high-quality language technology?

We believe that *Language Technology made in Europe* will significantly contribute to future European cross-border and cross-language communication, economic growth and social stability while establishing for Europe a world-wide, leading position in technology innovation, securing Europe's future as a world-wide trader and exporter of goods, services and information.

2.3 LANGUAGE TECHNOLOGY AND SOCIETAL CHALLENGES

With regard to the future information society there is a strong likelihood that the revolution in communication technology will bring people speaking different languages together in new ways. This is putting pressure on individuals to learn new languages and especially on developers to create new technology applications. In a global economy 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, Google+, Pinterest, Instagram etc.) is only the tip of the iceberg.

Many societal changes and economic trends confirm the urgent need to include sophisticated language technology in our European information and communication technology (ICT) infrastructure. Research, development and innovation efforts in LT must increase to go beyond what is possible today.

Linguistic, Commercial and Knowledge Barriers. A recent study on cross-border online commerce in the EU clearly indicates that language barriers are economic barriers [23]. Only 59% of retailers can handle transactions in more than one language. Translation and localisation costs must be drastically lowered before broad participation in Europe's single digital market is a reality. Multilingual language technology is the key, especially for SMEs.

At the same time, user expectations are increasing: 81% of all internet users think that websites run in their country should also be available in other languages. 44% of European users think they miss out on interesting information because websites are not available in a language they understand [18]. These facts can no longer be ignored. The availability of reliable LT can help establish a potentially vast market for information as well as consumer and entertainment goods in any language.

Ageing Population. Demographic changes suggest the need for more assistive technologies, especially those that drastically improve spoken language access. An aging population requires technology that can help master everyday situations and provide proactive guidance. Such technologies could answer the question, "Where did I leave my glasses?" The economic cost of demographic changes will also mean that more health care services and support systems will be required in our homes. Ambient assisted living (AAL) technologies can greatly benefit from a personalised, spoken method of interaction that is possible due to recent developments in the field of dialogue systems and interactive assistants.

People with Disabilities. The way we deal with disabilities has changed dramatically in the last 20 years. It shifted from an approach based on assistance, recovery or maintenance of functional capabilities to the goal of full integration. New technologies can help us reach the ambitious goal of achieving equal opportunities and promoting independent living. Language technologies already help people with disabilities to participate in society. Noteworthy examples include screen readers, dictation systems and voice-activated services. In addition to the social aspect there is a huge commercial market for future technologies such as, for example, full dialogue systems and interactive assistants, sign language recognition and synthesis, automatic translation, summarisation and content simplification. Approximately 10% of Europeans (50 million citizens) have permanent disabilities.

Immigration and Integration. According to the United Nations' International Migration Report 2002, 56 million migrants lived in Europe in 2000 [24]. The number of migrants has grown roughly to 60 million people today. Facilitating communication, providing access to information in foreign languages and helping people learn European languages can help better integrate migrants into European society. In fact, speech and language technologies can dramatically improve the integration process by providing intelligent language learning environments, automatic subtitling services in real time and automatic translation services.

Personal Information Services and Customer Care. Broadband access to information and services is common, mobile communication is daily routine for millions of Europeans. In this 24/7 economy we expect quick and reliable answers as well as engaging and timely online news broadcasts. However, information overload is also common, and it limits exchange in the digital information society. Citizens, governments and industries would greatly benefit from new technologies that help get the situation under control again. Embedded mobile applications enhanced with language technology will become personal assistants to everyone, offering automatic and intelligent question answering and dialogue capabilities, as well as automatic, personalised and trusted text and speech processing of messages, news items and other content.

Global Cooperation and Human Communication. Companies need to address new markets where multiple languages are spoken and support multinational teams at multiple locations. Many jobs cannot be filled today because linguistic barriers exclude otherwise qualified personnel. A flexible and mobile population requires multilingual language skills. Improvements in language technology can enable richer interactions and provide more advanced video conferencing services. Future technologies like a three dimensional internet can enable new modes of situation-based collaboration in the workplace

as well as support more realistic training and education scenarios. We will soon be able to participate in virtual events as new forms of entertainment, cultural exchange and tourism. Combining 3D virtual worlds and simulations with multilingual language technology including simultaneous translation, automatic minute taking, video indexing and video searching will let us experience being European in a brand new way.

Preservation of Cultural Heritage and Linguistic Diversity. According to the principles of the UN-endorsed World Summit on the Information Society [25], the "Information Society should be founded on and stimulate respect for cultural identity, cultural and linguistic diversity." Much effort has been put into the creation of digital archives and virtual museums that should help promote our cultural heritage. However, digitisation and digital asset management are only the first step. The amount of available information and language barriers still hinder the enjoyment and usage of our cultural treasures. Language technology can make this content accessible, e. g., through cross-lingual and multimedia search and machine translation. Likewise, communication skills need to be trained, especially in the light of today's find-remix-share paradigm of social media. This is underlined by the UNESCO Information for All Programme [26], which seeks to "support the production of local content and foster the availability of indigenous knowledge through basic literacy and ICT literacy training." Computer assisted language learning and language technology should be embedded into didactic software and games to help rescue our linguistic knowledge and diversity.

Social Media and e-Participation. Participation in online social media is a key characteristic of the early twenty-first century. Social media have a tremendous impact on practically all areas of society and life. Social media can help us solve technical problems, research products, learn about interesting places or discover new recipes. At the same time, recent developments in North Africa demonstrate the ability of social media to bring citizens together to express political power. Social media will play a role in the discussion of important, future topics for Europe like a common energy strategy and a common foreign policy. A severe problem is that certain groups are becoming detached from these developments. One can even speak of a broken link regarding communication cultures. This is an issue since both types of bottom-up movements sketched above are highly relevant for politicians, marketing experts, and journalists who would like to know what their customers or citizens think about their initiatives, products, or publications and to be able to react accordingly. However, it is not possible to process manually the sheer amount of information generated in multiple languages on social networks. We need to develop sophisticated language technology that is able to analyse these developments in real time.

Market Awareness and Customer Acceptance. Language technology is a key part of business and consumer software. The exact size of this market is difficult to assess because LT is often hidden inside other, more visible products. Customer acceptance of LT has recently been shown to be high. For example, market research by the Ford Motor Company indicates that their voice control system, Ford SYNC, is widely accepted [27]. 60% of Ford vehicle owners use voice commands in their cars. Non-Ford owners report a three-fold increase in their willingness to consider Ford models while 32% of existing customers admit that the technology played an important or critical role in their purchase decision. Language technology has a tremendous market potential.

One Market, Many Languages. Support for the 23 official languages of the EU has major economic and social implications, but the political dimension is equally important. Europe currently lags behind countries such as India (22 "official" languages) and South Africa (11 national languages). Government programmes in these two countries actively foster the development of lan-

guage technology for a significant number of official languages (India: http://tdil.mit.gov.in; South Africa: http://www.meraka.org.za/humanLanguage.htm). Mobile devices will become an even more important connection point between humans and information technology. Google already provides free translation services in 3,306 different language pairs as well as voice input for 16 languages and speech output for 24 languages. Apple's App and iTunes Store has demonstrated how premium content and products can be marketed for free and for a fee. Europe must address this global competition.

Secure Europe. The evolving information and knowledge society has improved human communication and information access, but the same communication networks also help some to commit crimes such as identity theft and internet fraud. The effective persecution of illegal activities requires automatic tools that can help detect crimes and monitor offenders. Language technology can help to build systems that can monitor, analyse and summarise large amounts of text, audio and video data in different languages (European and non-European) and from different sources (websites and social media).

The solutions presented above are strongly influenced by larger trends (see the following chapter), such as cloud computing, social media, mobile apps and web services. Many of these products and services are only available online. For example, severely restricting access to Facebook and Twitter strongly influenced recent political developments in North Africa. In Europe, the idea of social innovation has recently gained interest as it "offers an effective approach to respond to social challenges by mobilizing people's creativity to develop solutions and make a better use of scarce resources" [28]. Social innovation, which is also part of Europe's 2020 strategy, critically relies on active involvement of citizens and interaction among them, which calls for supportive multilingual language technologies.

Multilingualism has become a global norm rather than an exception. Future 3D applications that embed information and communication technology require sophisticated language technologies. Fully speech-enabled autonomous robots could help in disaster areas by rescuing travellers trapped in vehicles or by giving first aid. Language technology can significantly contribute towards improving social inclusion. Language technology can also help us provide answers to urgent social challenges while creating genuine business opportunities.

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, machinery, 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 provide multilingual information and knowledge management as well as content production in many European languages.

2.4 MARKET OPPORTUNITIES

There is an immensely large set of interesting and promising market opportunities around Language Technologies in Europe. We agreed with the EC-funded initiative "LT Innovate" ("LT-Innovate is the Forum for Europe's Language Technology Industry", see http://lt-innovate.eu) that we will include a concise description of the market opportunities into this document once the "LT Innovation Agenda" has been prepared.

To provide a rough indication about the estimated size of the different markets: The European market for translation, interpretation and localisation is expected to have a size of ca. 16.5 billion Euros in 2015 [8]. The global speech technology market is to reach the number of ca. 20.9 billion US-Dollars by 2015 [29].

3

MAJOR TRENDS IN INFORMATION AND COMMUNICATION TECHNOLOGIES

3.1 THE CURRENT STATE

Networked computers are ubiquitous. They come in many different shapes and forms (desktop, laptop, mobiles, tablets, etc.) or are embedded in devices, objects, and systems (e. g., smartphones, cameras, washing machines, cars, heating systems, robots, factories, traffic control systems). Software is usually available in multiple human languages. Standardisation efforts such as the introduction of Unicode solved the problem of representing and displaying different scripts, alphabets and special characters. The main use cases for today's computers are text processing, spreadsheets, presentations, communication (e-mail, Facebook, Twitter, Skype etc.), information search and entertainment (photos, music, films, games).

Mobile devices and social media are ever more reshaping how and when we communicate with one another using the tools and devices we use both in business and private life. The way we interact with computers is no longer restricted to graphical user interfaces with limited functionality but it is being extended through touch screens, voice interfaces and dialogue systems, tactile interfaces and mobile devices with built-in accelerometers that tell the device how it is held by the user.

Language technology is currently not well integrated into applications and interfaces – to the end user, spelling and grammar checking as well as, to a certain extent, search seem to be the only notable exceptions. A trend towards more intelligent language-based interaction is illustrated by Apple's introduction of the mobile assistant Siri in the latest iPhone and, recently, Google Now.

The web represents much of our knowledge. It emerged as a collection of static documents. Nowadays it is first and foremost a collection of systems and databases that can be queried through APIs, and applications such as Google Mail, Google Calendar, Facebook, eBay and Amazon. Many people only need one interface application on their computers: a web browser. Others use netbooks whose operating system more or less *is* the browser (Chromium OS). Behind the scenes, there is already a considerable amount of language technology incorporated in web applications such as search engines, dialogue systems, or machine translation services.

3.2 HARDWARE

Networked computers are no longer as big as a refrigerator, the age of the clumsy tower or desktop computer is over. Nowadays, networked computers come in many shapes and forms: small mobile devices (smartphones using, for example, Android or iOS), tablets, netbooks, ultra-portable laptops, small desktop computers, ebook readers, radios, television sets, gaming consoles and other entertainment devices with built-in wireless and access to, for example, RSS feeds, internet radio stations or youtube, cameras or house-hold appliances such as vacuum cleaners, coffee machines or scales that push the weight of the user to the cloud from where it can be monitored using an app on the smartphone. The next revolution in the hardware market will be wearable comput-

ers. Google has already demonstrated a prototype of their Google Glasses product in which the computer visuals are projected into a head-up display that looks like a regular pair of glasses. This approach can be used to provide the user with a true augmented reality perspective and a hands-free computing environment which immediately brings up the question if it will be possible to interact with the Glasses, or a similar product, using only your voice.

The shape and size of computers is no longer determined by the shape and size of their internal hardware components. Due to further breakthroughs in miniaturisation, the form of computers now truly follows their function. While computers and devices with embedded systems get smaller and smaller, the distributed data centres around the world get bigger and bigger – both in terms of number and size. The concept of cloud computing and storing data in dedicated data centres from where the data can be accessed by multiple devices (see, for example, Apple's iCloud), is already mainstream and used by millions of consumers world-wide. An important reason for the success of using the cloud to store data is the fact that, by now, people tend to have more than one computer, a not too unusual setup may include a laptop, a smartphone, a tablet and another computer as a dedicated media centre. Cloud services are ideal for synchronising data between all devices without buying, configuring and administering your own server machine.

3.3 SOFTWARE

The trends in the software area are much more multi-dimensional – in this section we can only scratch the surface and highlight several recent developments and current trends.

Communication: Probably the most important cornerstone of today's computer use is communication (both human to human and human to machine), be it more direct communication via traditional e-mail, instant messaging, text-based chat systems, video chat between two people or larger groups (Skype, Facetime, Google Hangout) or indirect communication and staying in touch with friends, acquaintances and colleagues via social networks such as Twitter, Facebook, XING and LinkedIn or social media such as blogs, YouTube, Pinterest or Instagram. An important factor is that millions of people world-wide are, by now, always online using several different networked devices including their phones.

Search and Information Services: Another important use case of any type of computing device is to search for information and to make use of specialised information services. Important applications are web search engines such as Google Search or Microsoft's Bing, Wikipedia, Google News, Google Books, digital libraries such as Europeana, meta-search engines and RSS feed aggregators etc.

Location-based Services: Search queries are nowadays often coupled to the user's current location. Location-based services enable the user to search for certain information in his or her geographic area, to make use of online maps, navigation systems, recommender systems such as Yelp or Qype or to find tweets or photos on Instagram in his or her geographic area.

E-Commerce and Shopping: World-wide billions of Euros are spent each year using general online shops such as Amazon or eBay or shops run by specific brands or services, reservation and booking, online banking and brokering services etc.

Media and Entertainment: Different types of media (photos, videos, music, sounds, text and multimedia documents, audio and video podcasts, ebooks, films, tv programmes etc.) play an important role. Not only personal media and other user-generated content are often connected to social networks (posting photos or videos on Instagram, Facebook, Google+, Flickr or YouTube), songs, photos or videos created and posted by third parties are also often shared using social networks. Almost all of the media mentioned above can be purchased using general or specific online stores, for consumption on

any device. Another important category of software is games, from online Flash games to games that are embedded into social networks, location-based games, multiplayer games with millions of users to very simple but also very successful casual games such as Angry Birds.

App and Media Stores: The success of ecommerce platforms, online shopping and the increased use of digital media led to the development of dedicated app and media stores. By now it is possible to buy or to rent almost every movie ever made (Amazon, iTunes), to buy music (iTunes music store), to stream music from the cloud onto your device (Spotify) and to buy software and mobile apps through dedicated stores (e. g., Apple's app stores for MacOS and iOS) without any need to ship physical media. An important development is in-app purchasing, especially on mobile devices: with a single tap of a finger it is possible to buy, within a specific app (which is usually available for free), additional modules, components or data sets for a small price.

Personal Information Management: With the ever increasing number of personal and professional contacts (including social networks), meetings and personal errands to run, there is a big trend towards personal information management. This includes address and contacts databases that are often integrated into larger applications such as Google Contacts (embedded in, among others, Google Mail) or Apple's AddressBook (used in Apple Mail). Cloud-integration is an important feature, so that contact information (including names, email address, phone numbers, photos etc.), calendar entries, "to do" items and the data from other productivity tools are always available on all devices.

Office Applications: The classic office applications – word processors, spreadsheets, presentations – are still important in the professional context and also in home use. Nowadays, there are several applications to choose from including open source software, cloud-based services and applications for Apple's iOS (MS Office, Apple iWork, Open Office, Google Docs). Except for Open Office all office suites use the cloud to enable the user to, for example, finish work on a presentation at the desktop computer where the document is automatically pushed to the cloud and to continue working on the presentation on a mobile device on the way home.

One of the most basic common denominators of all pieces of software is language – language plays a central and integral part in practically every single app, tool or application. However, language technology as such (including text analysis, information retrieval and extraction, spelling and grammar checking, speech recognition and synthesis, dialogue systems etc.) is usually completely hidden from the user, integrated into bigger applications, working behind the scenes. There is, however, a clear trend to embed language technologies not only at the level of the single application but on the level of the operating system. Another important factor of current computing is communicating and interacting with other people or groups of people, both on the personal level and also for business purposes. A third crucial ingredient of computing today is information, especially structured information which is annotated based on specific standards (see, for example, the family of standards around XML, Semantic Web, Linked Open Data, Web Services etc.).

3.4 CURRENT TRENDS AND MEGA-TRENDS

In the following we briefly sketch some of the current trends and mega-trends, loosely grouped into three sections.

Internet: The internet will continue to be *the* main driving force behind future developments in information and communication technologies. There are several mega-trends tightly coupled to the internet and network technologies: among these are cloud computing and cloud services, including cloud storage, as well as linked open

data and the semantic web. Social media and social networks will continue to change everything and to penetrate the market further, including niche markets, driven by location-based services. With the predominance of social networks we expect a certain convergence of digital identities that will enable users to have and to maintain one central digital identity that feeds into their multiple social network profiles including also a merger of the business-self and the private-self. Along those lines, exchanging and distributing personal data and information (photos, videos, music etc.) in a secure way will become easier. We further expect more broad deployment and general acceptance of services in the areas of e-democracy and e-government (including open data portals) and a continued increase of e-commerce platforms and services. A perceived general information overload will continue to be a problem, although modern search engines, aggregation services and user interfaces help a lot; web search is generally considered a solved problem. New business models and ways to distribute content or services to the end-user will continue to emerge (see the different app stores and approaches such as in-app purchases).

People: Information and communication technologies are used by people – the predominance of social networks and being always-on using smartphones, tables and laptops, is responsible for the fact that the way people interact, communicate and do business with one another will continue to be redefined and reshaped completely, including novel approaches for participation and public deliberation processes. Communication tools such as email, Twitter, Facebook etc. are mainstream by now and used across all age groups. This trend will continue. A popular phrase that characterises the main essence of the success factor of social networks is "faces and places" as this is what people are mainly interested in: other people first and foremost as well as certain buildings, restaurants, cinemas, landmarks and many others. The trend to use location-based services to find current friends, items of interest or even new friends with similar interests on social networks will continue (along with a more in-depth discussion of privacy issues). We also expect a tighter connection between the data stored in social networks and the linked open data cloud as well as a tighter connection between tools for personal information management and linked open data.

Hardware and Software: By now many internet companies operate under the slogan "mobile first". Accessing the internet or using web services on mobile devices will overtake the use of desktops and laptops very soon. There is also a clear tendency for completely novel mobile devices with Apple's iPad and Google's Glasses being two prime examples; in addition, there is a tendency for more household-appliances connected to the internet (tv, radio, gaming consoles, refrigerator, scales, coffee machine, lamps etc.; see the Internet of Things). Many of these devices will not have any displays but voice-driven interfaces. We expect a seamless integration of mobile devices into the hardware landscape at home including very simple file, data and application transfer and exchange among arbitrary mobile or stationary devices, playing music or movies on arbitrary displays or video projectors etc. Very soon there will not be a need anymore for the average user to own a laptop or desktop computer because mobile devices (phones and tablets) will cover all basic needs. As regards networks, their capacity and bandwidth will continue to grow, mobile telecommunication networks will gradually become more important than, for example, ADSL lines. The quality of voice or video calls (Skype, Facetime, Google Hangout) will continue to improve, phones and all other devices will continue to become faster, have more storage as well as 3D-capable displays that offer more intricate modes of interacting with the device. Mobile phones will have built-in facilities to replace credit cards for payment purposes (for example, using Near Field Communication), effectively replacing the wallet. Finally, the market for apps, especially mobile

apps, will continue to grow. Nowadays many companies, services and events have their own app that users can interact with and that usually offer added value when compared to the respective website. In order to be successful on the app market, usability will continue to be a decisive factor: only those apps will be successful that users can interact with intuitively right away.

To sum up, information and communication technologies will continue to be ubiquitous, available wherever and whenever needed. These technologies will be services that combine widely distributed applications, resources and data. They will be able to adapt to the location, situation and needs of the user including current emotions, habits and goals. As can be seen by the success of Wikipedia and other collaboratively edited knowledge bases, it is only a matter of time until a gigantic digital model of our world will exist that consists of interlinked and overlapping components. Naturally, languages and especially the automatic processing of languages using sophisticated language technologies will play a key role in this development. Now is the time to realise the needed breakthroughs. High performance, robust machine translation and related language technology services are urgently needed. There is a huge window of opportunity for consumer-oriented language technology: mobile devices are fast enough and have enough computing power, memory and a direct internet connection; they have a camera and are always online; it is easy to buy apps or add-ons.

While the LT-related aspects will be further discussed in the following chapters, we provide a more in-depth discussion of two selected trends in Sections 3.5 and 3.6.

3.5 SELECTED TREND: LINKED OPEN DATA AND THE DATA CHALLENGE

Data is considered one of the main topics of the future. At the European Data Forum 2012 and several other occasions the "data challenge" (big data, open data, linked data and the data value chain) is seen as one of the main themes and driving forces for future developments in information technology. Language technology and the priority research themes described later have strong relations to the data challenge, both as contributors and as beneficiaries. On the one hand, LT can help to exploit the immense volumes of information, knowledge and data encoded through natural language in text documents. LT can extract information from texts and make them accessible as structured data for automatic processing. On the other hand, LT will be able to analyse and interpret language data much better if it can use the growing volume of available structured data as background knowledge. Most of humankind's knowledge, reflection, communication and planning is encoded in and through human language. Conceptualising language as an integral part of the growing data universe is the ultimate challenge for the "big data movement". Interpreting and interlinking textual knowledge with the linked data world will help in the process of extracting new knowledge from the masses of newly produced structured data.

The Translation Cloud will benefit from data available across languages. The translation technologies being developed will also help to address data challenges, like building and cleaning data sets that span across languages or building links between existing data sets within one or between several languages. Multilingual access is an important requirement for a European vision of e-Government and e-Participation services. On the one hand, language technology can make use of open, governmental data that is being developed in portals like

data.gov.uk or within the upcoming European data portal. On the other, improving language technologies is inevitable for realizing multilingual access to public sector data for all European citizens, as recommended by the European Interoperability Framework for European public services [30]: the sheer amount of data and language barriers between data sets are obstacles that can only be removed with technologies in the realm of, e. g., machine translation, cross-lingual information access and information extraction. Finally, one application scenario of Socially-Aware Interactive Assistants are multilingual virtual meetings that make use of shared data sets that provide information about individuals, organizations and interactions settings. The creation of these data sets is a challenge in terms of privacy and re-use of data. This leads to various open issues that need to be resolved both for the data challenge in general and for language technologies:

Public and private data infrastructures need to be made available with different implications in terms of, e. g., licensing schemes or provenance of data. Provenance in general is an important aspect to achieve trustworthiness of data and to assure data quality. The language technology community has created many language resources of high quality, and with adequate provenance information, these resources will play an important role for creating truly multilingual, linked open data. As a prerequisite, the data itself developed within language technology and localisation (terminological or lexical data, translation memories or language resources in general) needs to be made available as linked open data, using standardised, e. g., Semantic Web technologies. In addition, for creating applications based on these resources, language technologies need to be made ready-to-use beyond general textual input or output, for areas relying on quite specific formats like e-Government or e-Business.

There is another aspect of data in general that needs to be taken into account: From the perspective of human and machine translation workflows, and actually every LT application, there are two types of relevant data. The first are resources that are part of, e. g., an MT process: a statistical language model, rules for grammar-based machine translation, lexicons etc. The other are metadata, necessary to organize and improve translation or other LT-related processes. A big challenge for LT is the proliferation of formats and metadata types. The combination of input and output formats, of languages and domains to be taken into account, of customer relations in real-life scenarios and many others, lead to a multitude of problem situations that need to be solved individually.

The only way to tackle this problem is to develop standardised metadata which would help in various areas. First, workflows can be organised more easily, from source content through LT processes and back again, including CMS, TMS and CAT systems. Second, re-use of resources becomes easier by providing standardised metadata for identifying resources or pieces of content. Finally, metadata will foster interoperability of components in agile workflows, e. g., to ease the integration of the output of text analytics (e. g., standard tags for named entities) with terminology management and MT systems.

Metadata also need to be accompanied by reference implementations that help to achieve wide adoption. In addition, all metadata standardisation efforts need to involve not only consumers of metadata. It is important that producers of content are brought to the table; only high quality content with the appropriate metadata can lead to high quality results in LT applications.

With support from the 7th Framework Programme, the data and LT communities already have started building bridges in projects and infrastructures such as DBpedia, Monnet, Wikidata and META-SHARE. For the topic of metadata standardisation including LT, various organisations have proven to be helpful for wide range community building, including ISO TC 37, GALA and the World Wide Web Consortium (W3C). We are now in a good position to strengthen these relations and to assure the

long-term availability of data and metadata for the European multilingual information society.

3.6 SELECTED TREND: FROM CLOUD COMPUTING TO SKY COMPUTING

A major megatrend is known as cloud computing. An increasing proportion of IT solutions is offered through the internet, forecasts predict that this proportion will rapidly increase. Computing may be offered on different levels of abstraction ranging from "Infrastructures as a Service" (IaaS) via "Platforms as a Service" (PaaS) to the powerful concept of providing any suitable software product as an internet service (Software as a Service, SaaS). Especially the latter concept has far-reaching, mainly beneficial, implications for distribution, support, customization, maintenance and pricing. It also opens new opportunities for software evolution by emerging dynamic schemes of integration, evaluation, adaptation and scaling. A well-known example is the Google Docs online suite of office applications. In language technology an increasing number of solutions are already offered as free or commercial web services, among them machine translation, language checking and text-to-speech conversion.

A special challenge for cloud computing is the need for trust. Since the services are rendered outside their sphere of control, customers demand sufficient safeguards securing performance, data protection, and persistence. Large European users of translation technology do not send their corporate language data to the existing large online translation services because the service providers do not offer such mechanisms. The situation is even more severe for business intelligence applications where the confidentiality of the collected information can be mission critical for the relevant planning and decision processes.

The most far-reaching and promising development within the cloud computing trend is the inter-cloud or sky computing paradigm. Although the cloud metaphor originated from the widely used graphical icon for the internet symbolising the entire global network outside the user's computer, soon the term became applied to any individual computing service provided on the internet.

Sky computing extends the notion of cloud computing beyond its original meaning. The term was coined for a setup in which clouds are combined into complex services, environments with workflows realising functionalities that exceed the capabilities of the individual services. A new line of research and development is dedicated to the creation of sky computing platforms that permit such integration.

Language technologies are prime candidates for sky computing setups since they are often a component of complex applications such as services supporting knowledge discovery, business intelligence or text production. Taking into account the large number of languages, language variants and subject domains, a sky computing setup can provide a much larger number of language and task-specific workflows through service composition than a traditional software product. Moreover, small and medium technology enterprises will be able much more easily to enter the market, stay on the market and improve their services without having to cast all demanded service combinations into their product family or into a range of bilateral OEM partnerships.

4

LANGUAGE TECHNOLOGY 2012: CURRENT STATE AND OPPORTUNITIES

4.1 CURRENT STATE OF EUROPEAN LANGUAGE TECHNOLOGY

Answering the question on the current state of a whole R&D field is both difficult and complex. For language technology, even though partial answers exist in terms of business figures, scientific challenges and results from educational studies, nobody has collected these indicators and provided comparable reports for a substantial number of European languages yet. In order to arrive at a comprehensive answer, META-NET prepared a White Paper Series that describes the current state of language technology support for 30 European languages [14]. The White Paper Series has been in preparation since mid 2010, has been finalised in the Spring of 2012 and is currently in print. More than 160 co-authors participated to the 30 volumes, more than 50 additional experts contributed supporting information, data and figures. Language White Papers were written for the following 30 European languages (including all 23 official EU languages):

Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Spanish, Swedish

The current state of support through language technology varies considerably from one language community to another. In order to compare the situation between languages, the META-NET Language White Papers introduce an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis) as well as basic language resources needed for building LT applications (for example, very large collections of texts for machine learning purposes). For each language, support through language technology was categorised using a five-point scale (1. excellent support; 2. good support; 3. moderate support; 4. fragmentary support; 5. weak or no support) and measured according to the following key criteria:

Machine Translation: quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Speech Processing: quality of existing speech recognition and synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Text Analysis: quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars.

Resources: quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage

of existing lexical resources and grammars.

The more than 160 co-authors of the Language White Papers prepared an initial language-specific assessment of language technology support using an approach in which ca. 25 different applications, tools and resources were assessed along seven different axes and criteria. Later on, the 30 individual and language-specific matrices were condensed in multiple iterations in order to arrive at a single score per language and area.

Figures 4 to 7 (p. 26 and 27) show the results. The figures demonstrate that there are dramatic and alarming differences in LT support between the various European languages and technology areas. In all four areas, English is ahead of the other languages but even support for English is far from being perfect. While there are good quality software and resources available for a few larger languages and application areas, others, usually smaller or very small languages, have substantial gaps. Many languages lack even basic technologies for text analysis and essential language resources. Others have basic tools and resources but the implementation of, for example, semantic methods is still far away. Therefore, a large-scale effort is needed to attain the ambitious goal of providing high-quality language technologies for all European languages.

The 30 volumes of the Language White Paper Series contain detailed assessments of LT support for each of the 30 languages. Due to space limitations we are unable to reproduce the results in this document. Two key results of this study are that currently no language, not even English, has the technological support it deserves. Also, the number of badly supported and under-resourced languages is unacceptable if we do not want to give up the principles of solidarity and subsidiarity in Europe.

4.2 CHALLENGES AND CHANCES

As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for highly specialised domains and purposes, and often exhibited rather limited performance. By now, however, there are huge market opportunities in the communication, collaboration, education and entertainment industries for integrating language technologies into general information and communication technologies, games, cultural heritage sites, edutainment packages, libraries, simulation environments and training programmes. Mobile information services, computer-assisted language learning software, e-learning environments, self-assessment tools and plagiarism detection software are just a few application areas in which language technology can and will play an important role in the years to come. The success of social media networks such as Twitter and Facebook demonstrates a further need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology represents a tremendous opportunity for the European Union. It can help address the complex issue of multilingualism in Europe. Citizens need to communicate across language borders, criss-crossing the European common market – language technology can help overcome this final barrier while supporting the free and open use of individual languages. Looking even a bit further into the future, innovative European multilingual language technology will provide a benchmark for other multilingual communities in the world.

The automated translation and speech processing tools currently available fall short of the envisaged goals. The dominant actors in the field are primarily companies based in the US. As early as the late 1970s, the European Union realised the profound relevance of language tech-

nology as a driver of European unity, and began funding its first research projects, such as EUROTRA. At the same time, national projects were set up that generated valuable results, but never led to a concerted European effort. In contrast to these highly selective funding efforts, other multilingual societies such as India (22 official languages) and South Africa (11 official languages) have recently set up long-term national programmes for language research and technology development.

Today the predominant actors in language technology rely on statistical approaches that have reached a performance plateau and that do not make use of deeper linguistic methods and knowledge. For example, sentences are translated automatically by comparing each new sentence with thousands of sentences previously translated by humans. The quality of the output completely depends on the size and quality of the available data. While the automatic translation of simple sentences in languages with sufficient amounts of available textual training data can achieve surprisingly good results, shallow statistical methods are likely to fail in the case of languages with a much smaller body of sample data or in the case of new sentences with complex structures. Analysing the deeper structural properties of languages in terms of syntax and semantics is a promising way forward if we want to build applications that perform well across the entire range of European languages.

The European Union is funding projects such as EuroMatrix and EuroMatrix+ (since 2006) and iTranslate4 (since 2010), that carry out basic and applied research and also generate resources for establishing high quality language technology solutions for several European languages. European research in the area of language technology has already achieved a number of outstanding successes. For example, the translation services of the European Union now use the Moses open source machine translation software, which has been mainly developed in European research projects [31]. In addition, national funding used to have huge impact. For example, the Verbmobil project, funded by the German Ministry of Education and Research (BMBF) between 1993 and 2000, pushed Germany to the top position in the world in terms of speech translation research for a time. Rather than building on the important results and success stories generated by these research projects, Europe has tended to pursue isolated research activities with a less pervasive impact on the market. The economic value of even the earliest efforts can be seen in the number of spin-offs. A company such as Trados, founded back in 1984, was sold to the UK-based SDL in 2005.

Drawing on the insights gained so far, today's hybrid language technology mixing deep processing with statistical methods will be able to bridge the gap between all European languages and beyond. But as we have described above, there is a dramatic difference between Europe's languages in terms of both the maturity of the research and the state of readiness with respect to language technology solutions.

Three key ingredients are needed to realise the technology visions described in the next chapter: the right actors, a shared vision and strategic programme and a certain level of support and commitment. Until recently the European community of language technologists and language professionals had to be considered highly fragmented at best. In early 2010 META-NET (see appendix C, p. 70) has started to bring the fragmented community together and to assemble researchers from the different subfields involved in language technology and also related scientific fields, universities, research centres, the language communities, national language institutions, smaller and medium companies as well as large enterprises, officials, administrators, politicians under one roof: META (Multilingual Europe Technology Alliance). By now META has more than 630 members in more than 50 countries (roughly one third of META's membership base are companies). Now that

the European language technology community has been brought together we can present our technology vision and strategic research agenda as illustrated in this very document. The whole META community has shaped this SRA through participating in many discussions around the ideas, approaches, technology visions and strategic goals described in this paper (see, among others, the list of key contributors on p. 68 f.). META-NET hopes to raise enough awareness, enthusiasm and, eventually, support to develop and, finally, to bring about a truly multilingual Europe based on sophisticated language technologies. To this end, we suggest to set up a shared and coordinated programme with the goal of concentrating our research efforts on the three priority research themes described in the next chapter. This shared and coordinated programme is foreseen to span all member states and associated countries and also the level of the European Commission.

4.3 MARKET OPPORTUNITIES

There is an immensely large set of interesting and promising market opportunities around Language Technologies in Europe. We agreed with the EC-funded initiative "LT Innovate" ("LT-Innovate is the Forum for Europe's Language Technology Industry", see http://lt-innovate.eu) that we will include a concise description of the market opportunities into this document once the "LT Innovation Agenda" has been prepared.

To provide a rough indication about the estimated size of the different markets: The European market for translation, interpretation and localisation is expected to have a size of ca. 16.5 billion Euros in 2015 [8]. The global speech technology market is to reach the number of ca. 20.9 billion US-Dollars by 2015 [29].

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	Catalan Dutch German Hungarian Italian Polish Romanian	Basque Bulgarian Croatian Czech Danish Estonian Finnish Galician Greek Icelandic Irish Latvian Lithuanian Maltese Norwegian Portuguese Serbian Slovak Slovene Swedish

4: Machine translation: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French German Italian Portuguese Spanish	Basque Bulgarian Catalan Danish Estonian Galician Greek Hungarian Irish Norwegian Polish Serbian Slovak Slovene Swedish	Croatian Icelandic Latvian Lithuanian Maltese Romanian

5: Speech processing: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch French German Italian Spanish	Basque Bulgarian Catalan Czech Danish Finnish Galician Greek Hungarian Norwegian Polish Portuguese Romanian Slovak Slovene Swedish	Croatian Estonian Icelandic Irish Latvian Lithuanian Maltese Serbian

6: Text analysis: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German Hungarian Italian Polish Spanish Swedish	Basque Bulgarian Catalan Croatian Danish Estonian Finnish Galician Greek Norwegian Portuguese Romanian Serbian Slovak Slovene	Icelandic Irish Latvian Lithuanian Maltese

7: Speech and text resources: State of support for 30 European languages

5

LANGUAGE TECHNOLOGY 2020: THE META-NET TECHNOLOGY VISION

5.1 THE NEXT IT REVOLUTION

People store and exchange information using the language they have known since early childhood. Computers have been ignorant about the languages of their masters for a long time. It took a while until they could handle scripts of languages different from English. It took even longer until computers could check the spelling of texts and read them aloud for the visually impaired.

On the web we can now get rough translations and we can search for texts containing a word, even if the word occurs in the text in a different form such as in plural number or in the genitive case. But when it comes to interpreting, actually making sense of certain input, and correctly responding, computers only "understand" simple artificial languages such as Java, PHP, Python, C++ and HTML.

After the next IT revolution computers will have mastered the languages of their users. Just as measures and formats for dates and times, the operating systems of tomorrow will know human languages. They may not reach the linguistic performance of educated people and they will not yet know enough about the world to understand everything, but they will be much more useful than today and further enhance our work and life.

5.2 COMMUNICATION AMONG PEOPLE

Since language is our most natural medium for interpersonal communication, computers cannot help much in regular conversations. However, when the communication partner speaks a different language this situation changes. With thousands of languages spoken on our planet, chances are high that we do not understand our partner. Rudimentary speech translation has been successfully demonstrated for limited numbers of languages and themes. By the year 2020, reliable and robust dialogue translation for face-to-face conversation and telecommunication can be achieved at least for hundreds of languages if research concentrates sufficient efforts on solving the problems of high-quality automatic translation and robust accurate speech recognition.

We use the computer as a tool for producing texts as well as for reading – from emails or instant messages to novels or technical documents. It checks spelling and grammar and its thesaurus suggests alternatives for words. LT products are already successfully employed in enterprises for checking conformance to corporate terminologies and style guidelines. In 2020 authoring software will also check for appropriate style according to genre and purpose including comprehensibility. It will flag potential errors and suggest appropriate corrections. It will employ authoring memories to proactively suggest completions of started sentences or whole paragraphs.

Today, Google Translate and other translation services provide access to information and knowledge for hundreds of millions of users who do neither speak English nor any other of the languages that make up most of the global web content. The technology is important for personal use and for numerous professional applications,

e. g., intelligence jobs at which analysts have to search through masses of texts for relevant information nuggets. However, often automatic translation outputs are still far away from the quality standards needed for a true impact on the translation and globalisation markets. Although the European Commission uses similar technology provided by European research projects, the translations in their current quality can only be used internally, which is great progress but does not yet help with the skyrocketing costs for outbound translation. Many translation services have started using machine translation but further economic breakthroughs through increased translation quality are still ahead of us. It will come in stages over the next ten years when the existing barriers for quality are overcome by new technologies that get closer to the structure and meaning behind human language.

In 2020 affordable high-quality translation for numerous domains and genres will be available among hundreds of languages if the proposed big push in research and innovation is implemented. We will be able to access such services online for written as well as for spoken language.

By 2020 many professional meetings will be tele-meetings utilizing large displays and comfortable presentation technology. LT will be able to record and transcribe face-to-face and virtual meetings. It will produce drafts and also summaries of minutes. For both types of meetings, it will simultaneously translate (interpret) the contributions of participants into as many languages as needed. The incrementally drafted records and summaries will be used for displaying the state of the discussion including intermediate results and open issues. The software will be guided by partial understanding of the contents, i. e., by their semantic association with concepts in semantic models of domains and processes. Brainstorming will be facilitated by semantic lookup and structured display of relevant data, proposals, charts, pictures and maps.

Language technology will also be used massively for helping with the ever-growing volume of correspondence. It will actively help to draft messages through automatic authoring techniques. Today many businesses and other organisations already employ e-mail response management software to filter, sort and route incoming email and to suggest replies to recognised types of requests. By 2020, business email will be embedded in semantically structured process models automating standardised communication. Already before 2020, email communication will be semantically analysed, checked for sentiment indicators and summarized in reports.

LT will also help to integrate the contents of all communication channels: telecommunication, meetings, email and chat, among others. Semantic integration into the work processes, threading and response management will be applied across channels. Machine translation and analytics will be available for all communication channels.

The extremely popular and powerful Web 2.0 mechanisms – social networks and user-generated content – have confronted LT with a new set of challenges. Every user can become a content producer and large numbers of people can participate in interpersonal communications. Some of these many-way mass communications have turned into effective instruments for support solicitation, idea creation, opinion formation and solution search. Communities can emerge in a matter of hours or days around admired works of art, shared preferences or social issues. Citizen action movements, international NGOs, patient self-help groups, expert circles and communities of concerned consumers can organize mutual support schemes, arrive at optimal and broadly supported social solutions and exert pressure on decision makers.

However, the social web cannot yet unfold its true potential because the large volumes of user-generated content become intransparent and unmanageable in no time. Both for participants and outside stakeholders or concerned decision makers it requires considerable efforts to stay on top of new developments. Much of the often-cited wisdom of the crowds and quite a bit of the aggregated

motivation is wasted because of information overflow. LT can and will eventually harness the inevitable information deluge resulting from Web 2.0 communities and discussions. If dedicated research efforts are focussed, technologies will be in place by 2020 that monitor, analyse, summarise, structure, document and visualise social media dynamics. Democracy and markets will be enriched by powerful new mechanisms for improved collective solution development and decision making.

Another important role of LT in interpersonal communication is the automatic conversion of language between different modes. Early examples are dictation systems and text-to-speech tools that convert between spoken and written language. These technologies are already successful but within the next few years they will reach full maturity opening up much larger markets. They will be complemented by reliable conversion from spoken or written language into sign language and vice versa. LT will also be utilised for improved methods of supported communication and for conversion of everyday language into strongly simplified language for special types of disabilities (augmentative alternative communication).

5.3 COMMUNICATION WITH TECHNOLOGY AND THE REST OF THE WORLD

Through language technology, human language will become the paramount medium for communication between people and the rest of the world. Today's voice-control interfaces to smartphones and the query fields of search engines are just the modest beginning of overcoming the communication barrier between humankind and the non-human part of the world.

This world consists of plants, animals and natural as well as man-made objects. The realm of artefacts ranges from small simple objects via all kinds of technical devices such as machines, appliances and vehicles all the way to complex units such as robots, airplanes, buildings, traffic systems and cities. But the artificially created world also consists of information and knowledge contained in books, films, recordings and digital storage. Virtually all information and knowledge will soon be available in digital form. The volume of data and thus potential information created daily about all parts of our world keeps increasing at a fast rate. The result is a gigantic distributed digital model of our world, let's call it second world, which continuously grows in complexity and fidelity. Through massive networking of this information by meta-information services and the linking of open data, the second world is getting more useful as a resource for information, planning and knowledge creation.

Today we still have a rather clear distinction between intelligent beings, i. e., humans, artificial agents with some autonomous behaviour and all other kinds of objects. We can easily communicate with people and we would like to communicate with computers and robots. We do not feel a pressing need to speak with a cup or with a power drill. However, the situation keeps changing fast since more and more products come equipped with sensors, processors and some information services such as descriptions, specifications or manuals. Many of these objects are connected to the internet (Internet of Things) or at least represented on the web (Web of Things). Thus, eventually, we can and will communicate with such objects.

Depending on the function, complexity, relevance and autonomy of the artefacts, the nature of communication can strongly vary. Some objects will come with interesting information, often represented in the second world, that we would like to query and explore (such as, for example, user and maintenance manuals, historical digests and consumer information). Other objects will provide information on their state and will also have their own individual memory that can be queried. Objects than can perform actions such as vehicles and appliances will accept and carry out voice commands.

Recently the old concept of a personal digital assistant has gained popularity due to the successful launch of Siri on the iPhone. In the near future, much more sophisticated virtual characters will follow equipped with expressive voices, faces and gestures. They will become an interface to any information that is provided on the web in the appropriate form. Thus this assistant could speak for or about machines, locations, the weather, the Empire State Building or the London Stock Exchange. The metaphor of a personal assistant is powerful and extremely useful, since such an assistant can be made sensitive to the user's preferences, habits, moods and goals. It can even be made aware of socio-emotional signals and learn appropriate reactions from experience.

Realizing this ambitious vision will require a dedicated thoughtfully planned massive effort in research and innovation. By the year 2020 we could have a highly personalised, socially aware and socially interactive virtual assistant. Having been trained on the user's behaviour and educated from his digital information and communication space it will be proactive by offering valuable unrequested advice. Voice, gender, language and mentality of the virtual character could be adjusted to the user's preferences. The agent will be able to speak in the language and dialect of the user but also digest information in many other natural and artificial languages and formats. Because of these skills, the assistant can translate or interpret without the user even realising it.

In the future, many providers of information on products, services and touristic sites will try to present their information with a specific look and feel. The personality and functionality of the interface may also depend on the user-type; there may be special interfaces for children, foreigners and persons with disabilities, thus, there will be space for interfaces tailored to the providers' corporate identities or to the nature of the objects and services.

By the year 2020 there will be a competitive landscape of intelligent interfaces to all kinds of objects and services employing language and other media such as manual and facial gestures for effective communication. Depending on the complexity of functionalities and provided information, the language coverage will range from simple commands to sophisticated dialogues. Many interface services will be offered as customizable cloud-based middleware, others may be completely custom-tailored. The technologies needed for such interfaces to machines, objects and locations are all part of the socially aware virtual assistant so that our proposed priority theme also creates enabling technologies for other interface products.

Two large application domains are special in their demands and need for additional technologies: robotics and knowledge services.

Although stationary industry robots have already taken over large parts of industrial production, the real era of robots is still ahead of us. But within this decade, specialized mobile robots will be deployed for personal services, rescue missions, household chores and tasks of guarding and surveillance. Natural language is by far the best communication medium for natural human-robot interaction. Since human language is very elaborate when it comes to speaking about perception, motion and action in space and time, the interaction in the physical world poses enormous challenges to LT. Some of these challenges can be addressed within the priority theme of the digital assistant, but without additional LT research in the robotics area, the communication skills of robots will lag behind their physical capabilities for a long time. By 2020 we will have robots around us that can communicate with us in human language, but their user-friendliness and acceptance will largely depend on progress in the coming years of LT research.

The communication with knowledge services raises a different set of problems. Here it is the inherent complexity of the represented knowledge that requires considerable advances in technology. The complexity arises from both the intricate structures of the subject domains and

the richness of linguistic expressivity, in particular the great variety of options to implicitly or explicitly express the same fact or question. Moreover, most information that we can learn from a text is not encoded explicitly but stands "between the lines." For the human reader it follows from the text but for language technology it needs to be derived by applying reasoning mechanisms and inference rules along with large amounts of explicitly encoded knowledge about the world.

From watching the crew of spaceship Enterprise in the famous TV series Star Trek, we expect that, eventually, we will be able to just say "Computer" followed by any question. As long as an answer can be found or derived from the accumulated knowledge of mankind, it will come back in a matter of milliseconds. In the Jeopardy game show, the computer giant IBM Watson was recently able to find correct answers that none of its human competitors could provide. Erroneously one may think now that the problem of automatic question answering is solved. Undoubtedly Watson is a great achievement demonstrating the power of LT. But some of the questions that were too hard for the human quiz champions were actually rather easy for a machine that has stored handbooks, decades of news, lexicons, dictionaries, bibles, databases and the entire Wikipedia. With clever lookup and selection mechanisms for the extraction of answers, Watson could actually find the right responses without a full analysis of the questions or clues.

Most questions that people might ask cannot be answered by today's technology, even if it has access to the entire web, because they require a certain degree of understanding of both the question and the passages containing potential answers. However, research on automatic question answering and textual inferencing progresses fast and by 2020 we will be able to use internet services that can answer huge numbers of non-trivial questions.

One prerequisite of this envisaged knowledge access through natural communication are novel technologies for offline processing of large knowledge repositories and massive volumes of other meaningful data which will be discussed in the following subsection.

5.4 PROCESSING KNOWLEDGE AND INFORMATION

Most knowledge on the web, by far, is formulated in human language. However, machines cannot yet automatically interpret the texts containing this knowledge. Machines can interpret knowledge represented in databases but databases are too simple in structure to express complex concepts and their relations. The logical formalisms of semanticists, on the other hand, that were designed to cope with the complexity of human thought, were too unwieldy for practical computation. Therefore computational logicians developed simpler logical representation languages as a compromise between desired expressivity and required computability. In these languages, knowledge engineers can formulate formal models of knowledge domains, ontologies, describing the concepts of the domains by their properties and their relations to other concepts. Ontologies enable knowledge engineers to specify which things, people, places in the world belong to which concepts. Such a domain model can be queried like a database. Its contents can be automatically analysed and modified. The intellectual creation of domain models, however, turned out to be an extremely demanding and time-consuming task, requiring well-trained specialists. Their encoding of knowledge seemed to be a promising alternative to the current web, so that the vision of the Semantic Web was born.

The main bottleneck of the Semantic Web is the problem of knowledge acquisition. It is unrealistic to believe that the authors of web content will be able to encode knowledge in the semantic web languages based on description logics. Nor will there be any affordable services for the manual conversion of large volumes of content.

Since LT did not have any means for automatically interpreting texts, language technologists had developed methods for at least extracting relevant pieces of information from such texts. This technique became a useful extension to information retrieval, which enables users to find entire documents such as in Google search. A relatively simple information extraction task is the reliable recognition of all person and company names, time and date expressions, locations and monetary expressions (named-entity extraction). Much harder is the recognition of certain relations such as the one between company and customer, company and employee or inventor and invention. Even more difficult are many-place relations such as the four-place relation of a wedding between groom and bride at a certain data and time. Events are typical cases of relations. However, events can have many more components such as the participants, costs, causes, victims and circumstances of accidents. Although research in this area is advancing fast, a reliable recognition of relations is not yet possible.

Information extraction can also be used for populating ontologies. Texts and pieces of texts can be annotated by extracted data. These metadata can serve as a bridge between the "semantic" portions of the web and the traditional web of unstructured data. LT is indispensable for the realization of the vision of a semantic web.

In addition, LT can perform many other tasks in the processing of knowledge and information. It can sort, catalogue and filter content and it can deliver the data for data mining in texts, which has been termed text data mining. LT can automatically connect web documents with meaningful hyperlinks and it can produce summaries of larger collections of texts. The LT techniques of opinion mining and sentiment analysis can find out what people think about products, personalities or problems and analyse their feelings about such topics.

Another class of techniques is needed for connecting between different media in the multimedia content of the web. Some of the tasks are annotating pictures, videos and sound recordings with metadata, interlinking them with texts, semantic linking and searching in films and video content and cross-media analytics including cross-media summarization.

In the next few years we will see considerable advances for all these techniques. For large parts of research and application development, language processing and knowledge processing will merge. The most dramatic innovations will draw from progress in multiple subfields. The predicted and planned use of language and knowledge technologies for social intelligence applications, one of our three priority areas, will involve text analytics, translation, summarisation, opinion mining, sentiment analysis and several other technologies. If the planned massive endeavour in this direction can be realised, it will not only result in a new quality of collective decision-making in business and politics. In 2020, LT will enable forms of knowledge evolution, knowledge transmission and knowledge exploitation that speed up scientific, social and cultural development. The effects for other knowledge-intensive application areas such as business intelligence, scientific knowledge discovery and multimedia production will be immense.

5.5 LEARNING LANGUAGE

Soon every citizen on Earth will learn a second language, many will learn a third, a few will go beyond this by acquiring additional languages. Learning a language after the period of early childhood is hard. It is very different from acquiring scientific knowledge because it requires repetitious practicing by actual language use. The more natural the use, the more effective the practice is.

IT products that help to ease and speed up language learning have a huge market. Already today, the software market for computer-assisted language learning (CALL) grows at a fast rate. Current products are helpful complements to traditional language instruction, however, they

are still limited in functionality because the software cannot reliably analyse and critique the language produced by the learner. This is true for written language and even more so for spoken utterances. Software producers are trying to circumvent the problem by strongly restricting the expected responses of the user. This helps for many exercises but it still rules out the ideal interactive CALL application, which is an automatic dialogue partner ready around the clock for error-free conversation on many topics, a software that analyses and critiques the learner's errors and adapts its dialogue to the learner's problems and progress. LT cannot yet provide such functionality.

This is the reason why research on CALL applications has not yet come into full bloom. As research on language analysis and understanding and on dialogue systems progresses, we predict a boom in research and development in this promising and commercially attractive application area. Research toward the missing technologies is covered by our priority themes. We expect a strong increase in CALL research at some time between 2015 and 2020.

5.6 LEARNING THROUGH LANGUAGE

Since most K-12, academic and vocational instruction happens through language, spoken in classroom and read in textbooks, LT can and will play a central role in learning. Currently LT is already applied at a few places in the preparation of multiple-choice tests and in the assessment of learners' essays.

As soon as dialogue systems can robustly conduct nearly error-free dialogues based on provided knowledge, research can design ideal tutoring systems. But long before LT research will reach this point, we will be able to create systems that test for knowledge by asking questions and that provide knowledge to the learner by answering questions. Thus even adaptive loops of analytic knowledge diagnosis and customized knowledge transmission as they form the core of an effective learning system will become possible through LT. Knowledge structuring and question answering is covered by our priority themes. The transfer to research and development toward educational applications should happen through close cooperation with the active research scene in e-learning.

Although we do not expect to substitute human teachers by 2020, we predict that e-learning technology will have become much more effective and learner-friendly by that time through the integration of advanced LT.

5.7 CREATIVE CONTENTS AND CREATIVE WORK

One of the major cost factors in European TV and film production is the required subtitling and dubbing. Whereas some countries with multiple official languages or with strict legislation on subtitling or sign-language display have a long tradition in providing these services, producers in many other countries still leave all subtitling and dubbing to importing distributors or media partners. With a single digital market, the increase of productions for multiple language communities and with the strengthening of inclusion policies, the demand for fast and cost-effective subtitling and dubbing will grow significantly. In some countries the method of voice-over is widely used as it is cheaper than dubbing. A professional voice talent reads all translations, sometimes shortened, over the original sound track.

The automatic translation of subtitles is easier than the translation of newspaper articles because of shorter and simpler sentences in spoken language. Some commercial services have already started using machine translation for subtitles and audio-description. If monolingual subtitling becomes the norm demanded by law, automated subtitle translation could be deployed at large scale.

Open challenges are the automatic production of sign-language translations and dubbing. Especially automatic

dubbing will be a hard task for speech technology since it requires the interpretation of the intonation in the source language, the generation of the adequate intonation in the target language, and finally lip synchronisation. An easier method would be automatic voice-over for appropriate material and markets. In 2020 we will see wide use of automatic subtitling and first successful examples of automatic voice over for a few languages.

Language can also be a medium for creative work, not only in literature. In traditional fine arts, creation mainly happens by a direct production of visual objects or images in two or three-dimensional space through drawing, sculpting, constructing, painting or photographing. In creative writing, the creation happens in language. But in many other areas of creative work, the creation happens through languages, ranging from musical notation to programming languages. Here the created work is specified in some suitable notation. Often natural language is used, for instance in the formulation of storyboards and scripts for movies or in the design of processes or services.

In computer science, the idea of writing programmes in natural language is almost as old as programming itself. This approach would require the translation of natural language into a programming language. However, the inherent ambiguity and vagueness of natural language has remained a major problem. Another obstacle is the richness of language, i. e., there are often too many ways to express the same statement. Even if we could implement a system that would correctly translate a subset of a natural language into computer programs, how would one specify and memorize this subset?

Computer scientists have created a number of easily learnable interpreted languages, often scripting languages, whose syntax resembles simple sentence structures of English. The idea of natural language programming has recently received renewed attention because of the concept of ontology-assisted programming. Natural language statements are interpreted with respect to an ontology. We expect that the concept of programming in natural language will bear fruit through progress in the semantic interpretation of natural language with respect to formal ontologies. Natural language may never become the programming language of choice for professional programmers, but we will certainly witness means for specifying scripts and simple programs in natural language for the everyday computer user.

The ontology-based interpretation of natural language statements will also permit the specification of processes, services, objects which will then be automatically translated into formal descriptions and finally into actions, models, workflows or physical objects. By 2020 we can expect successful examples of natural language scripting and specification in a few suitable application areas.

5.8 DIAGNOSIS AND THERAPY

Because of the central role of language in human life, psychological and medical conditions affecting language use belong to the most severe impairments people can suffer from. Deficiencies in language can also be strong indicators for other conditions that are harder to detect directly such as damage to brain, nerves or articulatory system. LT has been utilized for diagnosing the type and degree of brain damage after strokes. Since the administration of diagnosis and therapy are time-critical for a successful recovery of brain functions, permanently available software can support the immediate detection and treatment of stroke effects.

Language technology can also be applied to the diagnosis and therapy of aphasia resulting from causes other than strokes, e. g., from infections or physical injuries. Another application area is the diagnosis and therapy of innate or acquired speech impairments, especially in children.

Dyslexia is a widespread condition affecting skills in reading and orthography. Some effects of dyslexia can be greatly reduced by appropriate training methods. Recent advances in the development of software for the therapy

of dyslexia give rise to the hope that specialized CALL systems for different age groups and types of dyslexia will help to treat this condition early and effectively.

Technologies for augmentative alternative communication referred to in Section 5.2 can also perform an important function in therapy since any improvement of communication for language-impaired patients opens new ways for the treatment of causal or collateral conditions. Expected progress in LT together with advances in miniaturisation and endo- and exo-prosthetics will open new ways for helping people who cannot naturally enjoy the benefits of communication.

5.9 LT AS A KEY-ENABLING TECHNOLOGY

The wide range of novel or improved applications mentioned in our shared vision only represent a fragment of the countless opportunities for LT to change our work and everyday life. Language-proficient technology will enable or enhance applications wherever language is present. It will change the production, management and use of patents, legal contracts, medical reports, recipes, technical descriptions, scientific texts and it will permit many new voice applications such as automatic services for the submission of complaints and suggestions, for accepting orders and for counselling in customer-care, e-government, education, community services, etc.

With so many applications and application areas, each of them confronted with different functionalities and types of language, we may be tempted to doubt that there is a common technology core. And indeed there has been a trend of excessive diversification in LT software development. Many tools can only be used for one purpose. This is different from the way humans learn their language. Once we have learned our mother tongue we can easily obtain new skills, always employing the core knowledge acquired during childhood. We learn to read, write, skim texts, summarize, outline, proof-read, edit and translate.

Currently we are witnessing a promising trend in LT giving rise to hope for faster progress. Instead of relying on highly specialised components, powerful core technologies are reused for many applications. We can now compose lists of components and tools that we need for every language since these will be adapted for and integrated into many applications. In addition, we have also identified lists of core data, such as text and speech corpora and language descriptions, such as lexicons, thesauri and grammars, needed for a wide spectrum of purposes.

In information technology, we can differentiate between specialized application technologies, such as credit-card readers, and enabling technologies, such as microprocessors, that are needed for rather diverse types of applications. In hardware technology, certain key-enabling technologies have been identified, technology areas indispensable for projected essential progress (e. g., nanotechnology, microelectronics including semiconductors, biotechnology and photonics). Similar key-enabling technologies also exist on the software side, such as database technology or network technology. Considering the broad range of LT-enabled applications and their potential impact on business and society, LT is certainly becoming a key-enabling technology for future generations of IT. In contrast to some of the key-enabling technologies listed above, Europe has not lost yet a leadership role in this field. There is no reason to be discouraged or even paralysed by the strong evidence of interest and expertise on the side of major commercial players in the US. In software markets the situation can change fast.

If Europe does not take a decisive stand for a substantial commitment to LT research and innovation in the years to come, we may as well give up any ambition in future IT altogether because there is no other software sector in which European research can benefit from a similar combination of existing competitive competence, recognized economic potential, acknowledged societal needs and determined political obligation toward our unique wealth of languages.

6

LANGUAGE TECHNOLOGY 2020: THE META-NET PRIORITY RESEARCH THEMES

6.1 INTRODUCTION

For decades it has been obvious that one of the last remaining frontiers of information technology is still separating our rapidly evolving technological world of mobile devices, personal computers and the internet from the most precious and powerful asset of mankind, the human mind, the only system capable of thought, knowledge and emotion. Although we use computers to write, telephones to chat and the web to search for knowledge, information technology has no direct access to the meaning, purpose and sentiment behind our trillions of written and spoken words. This is why it is unable to summarize a text, answer a question, respond to a letter and to translate reliably. In many cases it cannot even correctly pronounce a simple English sentence.

Visionaries such as Ray Kurzweil, Marvin Minsky and Bill Gates have long predicted that this border would eventually be overcome by artificial intelligence including language understanding whereas science fiction such as the Star Trek TV series suggested attractive ways in which technology would change our lives, by establishing the fantastic concept of an invisible computer that you have a conversation with and that is able to react to the most difficult commands and also of technology that can reliably translate any human and non-human language.

Many enterprises had started much too early to invest in language technology research and development and then lost faith after a long period without any tangible progress. During the years of apparent technological standstill, however, research continued to conquer new ground. The results were a deeper theoretical understanding of language, better machine-readable dictionaries, thesauri and grammars, specialized efficient language processing algorithms, hardware with increased computing power and storage capacities, large volumes of digitized text and speech data and, most importantly, powerful new methods of statistical language processing that could exploit language data for learning hidden regularities governing our language use.

We do not yet possess the complete know-how for unleashing the full potential of language technology for business and society as essential research results are still missing. Nevertheless, the speed of research keeps increasing and even small improvements can already be exploited for innovative products and services that are commercially viable. As a consequence, we are witnessing a chain of new products for a wide variety of applications entering the market in rapid succession.

These applications tend to be built on dedicated computational models of language processing that are specialized for a certain task. People, on the other hand, apply the basic knowledge of the language they have acquired during the first few years of their socialisation, throughout their lives to many different tasks of varying complexity such as reading, writing, skimming, summarizing, studying, editing, arguing, teaching. They even use this knowledge for the learning of additional languages, which explains why second languages are easier to acquire if they are closely related to the learners' native language.

After people have obtained proficiency in a second language, they can already translate simple sentences more fluently than many machine translation systems, whereas truly adequate and stylistically acceptable translation, especially of more demanding texts is a highly skillful art gained by special training.

Today, no text technology software can translate and check for grammatical correctness and no speech technology software could recognize all the sentences it can read aloud if they were spoken by people in their normal voices. But increasingly we observe a reuse of core components and language models for a wide variety of purposes. It started with dictionaries, spell checkers and text-to-speech tools. Google Translate, Apple's Siri and IBM Watson still do not use the same technologies for analysing and producing language, because the generic processing components are simply not powerful enough to meet their respective needs. But many advanced research systems already utilize the same tools for syntactic analysis. This process is going to continue.

In ten years or less, basic language proficiency is going to be an integral component of any advanced IT. It will be available to any user interface, service and application development. Additional language skills for semantic search, knowledge discovery, human-technology communication, text analytics, language checking, e-learning, translation and other applications will employ and extend the basic proficiency. The shared basic language competence will ensure consistency and interoperability among services. Many adaptations and extensions will be derived and improved through sample data and interaction with people by powerful machine learning techniques.

In the envisaged big push toward realising this vision by massive research and innovation, the technology community is faced with three enormous challenges:

1. *Richness and diversity.* A serious challenge is the sheer number of languages, some closely related, others distantly apart. Within a language, technology has to deal with numerous dialects, sociolects, registers, professional jargons, genres and slangs. Each language variant finally is abundant with alternatives to achieve the same goal or to express the same fact.
2. *Depth and meaning.* Understanding language can be a complex and creative process. Human language is not only the key to knowledge and thought, it also cannot be interpreted without certain shared knowledge and active inference. Computational language proficiency needs semantic technologies.
3. *Multimodality and grounding.* Human language is embedded in our daily activities. It is combined with other modes and media of communication. It is affected by beliefs, desires, intentions and emotions and it affects all of these. Successful interactive language technology requires models of embodied and adaptive human interaction with people, technology and other parts of the world.

If we could take these challenges with us into our research labs and reappear after some years with solutions that approximate the grand vision of language-competent IT, this would put expectations and commercial planning on hold and trigger in society the typical mixed attitude of waiting and forgetting, which can be observed with respect to nuclear fusion and manned Mars exploration. It is fortunate for both research and economy that the only way to effectively tackle the three major challenges mentioned above involves submitting the evolving technology continuously to the growing demands and practical stress tests of real world applications.

Google's Translate, Apple's Siri, Autonomy's text analytics and scores of other products demonstrate that there are plenty of commercially viable applications for imperfect technologies that are still far from the envisaged scope and capabilities. Only a continuous stream of technological innovation can provide the economic pull forces and the evolutionary environments for the realization of the grand vision.

In the remainder of the Chapter, we propose five major action lines of research and innovation:

- Three priority themes connected with powerful application scenarios that can drive research and innovation. These will demonstrate novel technologies in attractive show-case solutions of high economic impact. At the same time they will open up numerous new business opportunities for European language-technology and -service providers.
- A steadily evolving system of shared, collectively maintained interoperable core technologies and resources for the languages of Europe (and selected economically relevant languages of its partners). These will ensure that all of our languages will be sufficiently supported and represented in the next generations of IT solutions.
- The creation of a pan-European language technology service platform for supporting research and innovation by testing and showcasing research results, integrating various services even including professional human services. This showcase platform will allow SME providers to offer component and end-user services, and share and utilise tools, components and data resources.

The three priority research themes based on solution scenarios are:

- **Translation Cloud** – generic and specialised federated cloud services for instantaneous reliable spoken and written translation among all European and major non-European languages.
- **Social Intelligence** – understanding and dialogue within and across communities of citizens, customers, clients and consumers to enable e-participation and more effective processes for preparing, selecting and evaluating collective decisions.
- **Socially Aware Interactive Assistants** – socially aware pervasive assistants that learn and adapt and that provide proactive and interactive support tailored to specific situations, locations and goals of the user through verbal and non-verbal multimodal communication.

These priority themes have been designed with the aim of turning the joint vision into reality and to letting Europe benefit from a technological revolution that will overcome barriers of understanding between people of different languages, between people and technology and between people and the accumulated knowledge of mankind. The three research priority themes connect societal needs with LT applications and concrete roadmaps for the organization of research, development and scientific innovation. The priority themes are contextualised in the advanced networked society and cover the main functions of language: storing, sharing and using of information and knowledge, as well as improving social interaction among humans and enabling social interaction between humans and technology. As multilingualism is at the core of European culture and becoming a global norm, one theme is devoted to overcoming language barriers.

The three themes have been thoughtfully selected in a complex process (see Appendix E on p. 74 ff.) to ensure the needed market pull, the appropriate performance demands, the realistic testing environments and a sufficient level of public interest. Each of our identified challenges is covered by the three themes but strongly represented by one of them. The priority themes also represent a good mix of applications with respect to the various user communities. Small businesses, large enterprises, public administration and the general public as personal end users are all well represented among the beneficiaries of the targeted solutions.

6.2 PRIORITY THEME 1: TRANSLATION CLOUD

6.2.1 Solutions for the EU Society and the Citizen

The goal is a multilingual European society, in which all citizens can use any service, access all knowledge, enjoy all media and control any technology in their mother tongues. This will be a world in which written and spoken communication is not hindered anymore by language barriers and in which even specialised high-quality translation will be affordable.

The citizen, the professional, the organisation, or the software application in need of cross-lingual communication will use a single, simple access point for channelling text or speech through a gateway that will instantly return the translations into the requested languages in the required quality and desired format.

Behind this access point will be a network of generic and special-purpose services combining automatic translation or interpretation, language checking, post-editing, as well as human creativity and quality assurance, where needed, for achieving the demanded quality. For high-volume base-line quality the service will be free for use but it will offer extensive business opportunities for a wide range of service and technology providers.

Special components and extensions of the permanent and ubiquitous service are:

- use and provision platform for providers of computer-supported creative top-quality human translation, multilingual text authoring and quality assurance by experts
- trusted service centres: certified service providers fulfilling highest standards for privacy, confidentiality and security of source data and translations
- quality upscale models: services permitting instant quality upgrades if the results of the requested service levels do not yet fulfil the quality requirements
- domain and task specialisation models
- translingual spaces: dedicated locations for ambient interpretation. Meeting rooms equipped with acoustic technology for accurate directed sound sensoring and emission

6.2.2 Novel Research Approaches and Targeted Breakthroughs

The core reason why high-quality machine translation (HQMT) has not been systematically addressed yet seems to be the Zipfian distribution of issues in MT: some improvements, the "low-hanging fruit", can be harvested with moderate effort in a limited amount of time. Yet, many more resources and a more fundamental, novel scientific approach – that eventually runs across several projects and also calls – are needed for significant and substantial improvements that cover the phenomena and problems that make up the Zipfian long tail. This is an obstacle in particular for individual research centres and SMEs given their limited resources and planning horizon. Although recent progress in MT has already led to many new applications of this technology, radically different approaches are needed to accomplish the ambitious goal of this research including a true quality breakthrough. Among these new research approaches are:

- Systematic concentration on quality barriers, i. e., on obstacles for high quality
- A unified dynamic-depth weighted-multidimensional quality assessment model with task profiling
- Strongly improved automatic quality estimation
- Inclusion of translation professionals and enterprises in the entire research and innovation process
- Ergonomic work environments for computer-supported creative top-quality human translation and multilingual text authoring
- Semantic translation paradigm by extending statistical translation with semantic data such as linked open

data, ontologies including semantic models of processes and textual inference models
- Exploitation of strong monolingual analysis and generation methods and resources
- Modular combinations of specialized analysis, generation and transfer models, permitting accommodation of registers and styles (including user-generated content) and also enabling translation within a language (e. g., between specialists and laypersons).

The expected breakthroughs will include:

- High-quality text translation and reliable speech translation (including, among others, a modular analysis-transfer-generation translation technology that facilitates reuse and constant improvement of modules)
- Seemingly creative translation skills by analogy-driven transfer models
- Automatic subtitling and voice over of films
- Ambient translation

6.2.3 Solution and Technological Realisation

The envisaged technical solutions will benefit from new trends in IT such as software as a service, cloud computing, linked open data and semantic web, social networks, crowd-sourcing etc. For MT, a combination of translation brokering on a large scale and translation on demand is promising. The idea is to streamline the translation process such that it becomes simpler to use and more transparent for the end user, and at the same time respects important factors such as subject domain, language, style, genre, corporate requirements and user preferences. Technically, what is required is maximum interoperability of all components (corpora, processing tools, terminology, knowledge, maybe even pre-trained translation models) and a cloud or server/service farm of specialised language technology services for different needs (text and media types, domains, etc.) offered by SMEs, large companies or research centres.

A platform has to be designed and implemented for the resource and evaluation demands of large-scale collaborative MT research. An initial inventory of language tools and resources as well as extensive experience in shared tasks and evaluation has been obtained in several EU-funded projects. Together with LSPs, a common service layer supporting research workflows on HQMT must be established. As third-party (customer) data is needed for realistic development and evaluation, intellectual property rights and legal issues must be taken into account from the onset. The infrastructures to be built include:

- Service clouds with trusted service centres
- Interfaces for services (APIs)
- Workbenches for supporting creative translations
- Novel translation workflows (and improved links to content production and authoring)
- Showcases for services such as ambient and embedded translation

6.2.4 Impact

HQMT in the cloud will ensure and extend the value of the digital information space in which everyone can contribute in her own language and be understood by members of other language communities. It will assure that diversity will no longer be a challenge, but a welcome enrichment for Europe both socially and economically. Based on the new technology, language-transparent web and language-transparent media will help realise a truly multilingual mode of online and media interaction for every citizen regardless of age, education, profession, cultural background, language proficiency or technical skills. Showcase applications areas are:

- Multilingual content production (media, web, technical, legal documents)
- Cross-lingual communication, document translation

Research Priority	Phase 1: 2013-2014	Phase 2: 2015-2017	Phase 3: 2018-2020
Immediate affordable translation in any needed quality level (from sufficient to high)	Development of necessary monolingual language tools (analysis, generation) driven by MT needs; exploitation of novel ML techniques for MT purposes, using large LR and semantic resources, including Linked Open Data and other naturally occuring semantic and knowledge resources (re-purposing for MT and NLP use); experiment with novel metrics, automated, human-centered, or hybrid; use EU languages, identify remaining gaps (LR resources, tools)	Concentrate on High-Quality MT systems using results of Phase 1, deepen development of MT-related monolingual tools; employ novel techniques aimed at HQMT, combination of systems, domain adaptation, cross-language adaptation; develop showcase systems for novel translation workflow; use novel metrics identified as correlated with the aims of HQMT application; continue development on EU languages, identify needs for non-EU languages (MT-related) and their gaps	Deployment of MT systems in particular applications requiring HQMT, such as technology export, government and public information systems, private services, medical applications, etc, using novel translation workflows where appropriate; application- and user-based evaluation driven engagement of core and supplemental technologies; coverage of EU languages and other languages important for EU business and policy
Delivering multimedia content in any language (captioning, subtitling, dubbing)	Multi-media system prototypes, combining language, speech, image and video analysis; employing novel techniques (machine learning, cross-fertilization of features across media types); targeted evaluation metrics for system quality assessment related to MT; aimed at EU languages with sufficient resources; data collection effort to support multi-media analysis	Prototype applications in selected domains, such as public service (parliament recordings, sports events, legal proceedings) and other applications (archive TV series or movie delivery, online services at content providers); continued effort at multimedia analysis, adding languages as resources become available	Deployment of large-scale applications for multi-media content delivery, public and/or private, in selected domains; development of online services for captioning, subtitling, dubbing, including on-demand translation); new languages for outside-of-the-EU delivery, continued improvement of EU languages
Content analytics	...	...	...
Cross-lingual knowledge management and linked open data	...	...	...
Synchronous and asynchronous interpretation	...	...	...
Translingual collaborative spaces	...	...	...

8: Priority Theme 1 – Translation Cloud: Preliminary Roadmap

- Real-time subtitling and translating speech from live events
- Mobile interactive interpretation for business, social services, and security
- Translation workspaces for online services

6.2.5 Organisation of Research

Several very large cooperating and competing lead projects will share an infrastructure for evaluation, resources (data and base technologies), and communication. Mechanisms for reducing or terminating partner involvements and for adding new partners or subcontracted contributors should provide the needed flexibility. A number of smaller projects, including national and regional projects, will provide building blocks for particular languages, tasks, component technologies or resources. A special scheme will be designed for involving EC-funding, member states, industrial associations, and language communities.

Two major phases from 2015 to mid 2017 and from mid 2017 to 2020 are foreseen. Certain services such as multilingual access to web-information across European languages should be transferred to implementation and testing at end of phase 2017. Internet-based real-time speech translation for a smaller set of languages will also get into service at this time as well as HQMT for selected domains and tasks. A major mid-term revision with a thorough analytical evaluation will provide a possible breakpoint for replanning or termination.

A close cooperation of language technology and professional language services is planned. In order to overcome the quality boundaries we need to identify and understand the quality barriers. Experienced professional translators and post-editors are required whose judgements and corrections will provide insights for the analytical approach and data for the bootstrapping methodology. The cooperation scheme of research, commercial services and commercial translation technology is planned as a symbiosis since language service professionals working with and for the developing technology will at the same time be the first test users analytically monitored by the evaluation schemes. This symbiosis will lead to a better interplay of research and innovation.

Although the research strand will focus on advances in translation technology for innovation in the language and translation service sector, a number of other science, technology and service areas need to be integrated into the research from day one. Some technology areas such as speech technologies, language checking, authoring systems, analytics, generation and content management systems need to be represented by providers of state-of-the-art commercial products.

Supporting research and innovation in LT should be accompanied by policy making in the area of multilingualism, but also in digital accessibility. Overcoming language barriers can greatly influence the future of the EU. Solutions for better communication and for access to content in the users' native languages would reaffirm the role of the EC to serve the needs of the EU citizens. A connection to the infrastructure programme CEF could help to speed up the transfer of research results to badly needed services for the European economy and public.

At the same time, use cases should cover areas in which the European social and societal needs massively overlap with business opportunities to achieve funding investment that pays back, ideally public-private partnerships.

Concerted activities sharing resources such as error corpora or test suites and challenges/shared tasks in carefully selected areas should be offered to accelerate innovation breakthrough and market-readiness for urgently needed technologies.

6.3 PRIORITY THEME 2: SOCIAL INTELLIGENCE AND E-PARTICIPATION

6.3.1 Solutions for the EU Society and for the Citizen

The central goal behind this theme is to use networked information technology and the digital content of the web for improving effectiveness and efficiency of decision-making in business and society.

The quality, speed and acceptance of individual and collective decisions is the single main factor for the success of social systems such as enterprises, public services, communities, states and supranational organisations. The growing quantity and complexity of accessible relevant information poses a serious challenge to the efficiency and quality of decision processes. IT provides a wide range of instruments for intelligence applications. Business intelligence, military intelligence or security intelligence applications collect and pre-process decision-relevant information. Analytics programmes search the data for such information and decision support systems evaluate and sort the information and apply problem-specific decision rules. Although much of the most relevant information is contained in texts, text analytics programmes today only account for less than 1% of the more than 10 billion US$ business intelligence and analytics market. Because of their limited capabilities in interpreting texts, mainly business news, reports and press releases, their findings are still neither comprehensive nor reliable enough.

Social intelligence builds on improved text analytics methodologies but goes far beyond the analysis. One central goal is the analysis of large volumes of social media, comments, communications, blogs, forum postings etc. of citizens, customers, patients, employees, consumers and other stakeholder communities. Part of the analysis is directed to the status, opinions and acceptance associated with the individual information units. As the formation of collective opinions and attitudes is highly dynamic, new developments need to be detected and trends analysed. Emotions play an important part in individual actions such as voting, buying, supporting, donating and in collective opinion formation, the analysis of sentiment is a crucial component of social intelligence.

Social intelligence can also support collective deliberation processes. Today any collective discussion processes involving large numbers of participants are bound to become intransparent and incomprehensible rather fast. By recording, grouping, aggregating and counting opinion statements, pros and cons, supporting evidence, sentiments and new questions and issues, the discussion can be summarised and focussed. Decision processes can be structured, monitored, documented and visualised, so that joining, following and benefitting from them becomes much easier. The efficiency and impact of such processes can thus be greatly enhanced.

Since many collective discussions will involve participants in several countries, e. g., EU member states or enterprise locations, cross-lingual participation needs to be supported. Special support will also be provided for participants not mastering certain group-specific or expert jargons and for participants with disabilities affecting their comprehension.

6.3.2 Novel Research Approaches and Targeted Breakthroughs

A key enabler will be language technologies that can map large, heterogeneous, and, to a large extent, unstructured volumes of on-line content to actionable representations that support decision making and analytics tasks. Such mappings can range from the relatively shallow to the relatively deep, encompassing for example coarse-grained topic classification at the document or paragraph level or the identification of named entities, as well as in-depth syntactic, semantic and rhetorical analysis at the level of individual sentences and beyond (paragraph, chapter,

text) or the resolution of co-reference or modality cues within and across sentences.

Language technologies such as, for example, information extraction, data mining, automatic linking and summarisation have to be made interoperable with modern knowledge representation approaches and semantic web methods such as ontological engineering. Drawing expertise from related areas such as knowledge management, information sciences, or social sciences is an important prerequisite to meet the challenge of modelling social intelligence, see [32]. A new research approach should target the bottleneck of knowledge engineering by:

- "Semantification" of the web: bridging between the semantic parts and islands of the web and the traditional web containing unstructured data;
- Merging and integrating textual data with social network and social media data, especially along the dimension of time;
- Aligning and making comparable different genres of content like mainstream-news, social media (blogs, twitter, facebook etc.), academic texts, archives etc.;
- Extracting semantic representations from social media content, i. e., creating representations for reasoning and inferencing;
- Taking metadata and multimedia data into account.

The following list contains specific targeted breakthroughs to be sought in this scenario:

- Social intelligence by detecting and monitoring opinions, demands and needs;
- Detecting diversity of views, biases along different dimensions (e. g., demographic) etc. including temporal dimension (i. e., modelling evolution of opinions);
- Support for both decision makers and participants;
- Support of collective deliberation and collective knowledge accumulation;
- Vastly improved approaches to sentiment detection and sentiment scoring (going beyond the approach that relies on a list of positive and negative keywords).
- Introducing the approach of genre-driven text and language-processing (different genres need to be processed differently).
- Personalised recommendations of e-Participation topics to citizens;
- Proactive involvement in e-Participation activities;
- Understanding influence diffusion across social media (identifying drivers of opinion spreading);
- More sophisticated methods for topic and event detection that are tightly integrated with the Semantic Web, Linked Open Data and machine-readable knowledge bases such as DBpedia.
- Modelling content and opinions flows across social networks;
- Evaluation of created methods by analytic/quantitative and sociological/qualitative means.

6.3.3 Solution and Technological Realisation

Individual solutions should be assembled from a repository of generic monolingual and cross-lingual language technologies, packaging state-of-the-art techniques in robust, scalable, interoperable, and adaptable components that are deployed across sub-tasks and sub-projects, as well as across languages where applicable (e. g., when the implementation of a standard data-driven technique can be trained for individual languages). These methods need to be combined with powerful analytical approaches that can aggregate all relevant data to support analytic decision making and develop new access metaphors and task-specific visualisations.

By robust we mean technologically mature, engineered and scalable solutions that can perform high-throughput analysis of web data at different levels of depth and granularity in line with the requirements of the respective

applications. Technology should also be able to work with heterogeneous sources, ranging from completely unstructured (arbitrary text documents of any genre) to completely structured (ontologies, linked open data, databases).

To accomplish interoperability we suggest a strong semantic bias in the choice and design of interface representations: to the highest degree possible, the output (and at deeper levels of analysis also input) specifications of component technologies should be interpretable semantically, both in relation to natural language semantics (be it lexical, propositional, or referential) and extra-linguistic semantics (e. g., taxonomic world or domain knowledge). For example, grammatical analysis (which one may or may not decompose further into tagging, syntactic parsing, and semantic role labelling) should make available a sufficiently abstract, normalized, and detailed output, so that downstream processing can be accomplished without further recourse to knowledge about syntax. Likewise, event extraction or fine-grained, utterance-level opinion mining should operate in terms of formally interpretable representations that support notions of entailment and, ultimately, inference.

Finally, our adaptability requirement on component technologies addresses the inherent heterogeneity of information sources and communication channels to be processed in this scenario. Even in terms of monolingual analysis only, linguistic variation across genres (ranging from carefully edited, formal publications to spontaneous and informal social media channels) and domains (as in subject matters) often calls for technology adaptation, where even relatively mature basic technologies (e. g., part-of-speech taggers) may need to be customized or re-trained to deliver satisfactory performance. Further taking into account variation across downstream tasks, web-scale language processing typically calls for different parameterizations and trade-offs (e. g., in terms of computational cost vs. breadth and depth of analysis) than an interactive self-help dialogue scenario. For these reasons, relevant trade-offs need to be documented empirically, and component technologies accompanied with methods and tools for adaptation and cost-efficient re-training, preferably in semi- and un-supervised settings. The technical solutions needed include:

- Technologies and platforms for decision support, collective deliberation and e-participation.
- A large public discussion platform for Europe-wide deliberation on pressing issues such as energy policies, financial system, migration, natural disasters, etc.
- Visualization of social intelligence-related data and processes for decision support (for politicians, health providers, manufacturers, or citizens).
- High-throughput, web-scale content analysis techniques that can process multiple different sources, ranging from unstructured to completely structured, at different levels of granularity and depth by allowing to trade-off depth for efficiency as required.
- Mining e-participation content for recommendations, summarisation and proactive engagement of less active parts of population.
- Detection and prediction of events and trends from content and social media networks.
- Extraction of knowledge and semantic integration of social content with sensory data and mobile devices (in near-real-time).
- Cross-lingual technology to increase the social reach and approach cross-culture understanding.

We suggest to structure the research along at least the following five lines (see also Figure 9):

1. *Social influence and incentives:* modeling social diversity of views across languages and cultures; modeling social influence and incentives; multipolar opinion mining (beyond usual sentiment analysis)

2. *Information tracking:* tracking dynamics of information diffusion across languages, cultures and media; prediction of future events and identification of causal relationships from textual and social streams
3. *Multimodal data processing:* joining textual data and social networks, including spatial and temporal dimensions; joining textual and social data with unstructured sources like sensor data (smart cities), video, images, audio
4. *Visualisation and user interaction:* visualization of textual and social dynamics; adaptive user interfaces
5. *Algorithmic fundamentals:* algorithms and toolkits for scalable processing of multi-modal big data; real-time modeling and reasoning on massive textual and social streams

6.3.4 Impact

The 21st century presents us with multiple challenges including efficient energy consumption, global warming and financial crises. It is obvious that no single individual can provide answers to challenging problems such as these, nor will top-down imposed measures find social acceptance as solutions. Language technology will enable a paradigm shift in transnational public deliberation.

The applications and technologies discussed in this section will change how business adapts and communicates with their customers. It will increase transparency in decision-making processes, e. g., in politics and at the same time give more power to the citizen. As a by-product, the citizens are encouraged to become better informed in order to make use of their right to participate in a reasonable way. Powerful analytical methods will help European companies to optimise marketing strategies or foresee certain developments by extrapolating on the basis of current trends. Leveraging social intelligence for informed decision making is recognised as crucial in a wide range of contexts and scenarios:

- Organisations will better understand the needs, opinions, experiences, communication patterns, etc. of their actual and potential customers so that they can react quickly to new trends and optimize their marketing and customer communication strategies.
- Companies will get the desparately needed instruments to exploit the knowledge and expertise of their huge and diverse workforces, the wisdom of their own crowds, which are the most highly motivated and most closely affected crowds.
- Political decision makers will be able to analyse public deliberation and opinion formation processes in order to react swiftly to ongoing debates or important, sometimes unforeseen events.
- Citizens and customers get the opportunity (and necessary information) to participate and influence political, economic and strategic decisions of governments and companies, ultimately leading to more transparency of decisions processes.

Thus, leveraging collective and social intelligence in developing new solutions to these 21st century challenges seems a promising approach in such domains where the complexity of the issues under discussion is beyond the purview of single individuals or groups.

The research and innovation will provide technological support for emerging new forms of issue-based, knowledge-enhanced and solution-centred participatory democracy involving large numbers of expert- and non-expert stakeholders distributed over large areas, using multiple languages.

6.3.5 Organisation of Research

Research in this area touches upon political as well as business interests and at the same time is scalable in reach from the regional to the European scale. Therefore, it is necessary to identify business opportunities and potential impact for society at different levels and to align EU level research with efforts on the national level.

Research Priority	Phase 1: 2013-2014	Phase 2: 2015-2017	Phase 3: 2018-2020
Social influence and incentives	Modelling social diversity of views across languages and cultures	Modelling social influence and incentives through game theoretic approaches using data from textual and social networking streams	Holistic modelling of society (or its segments) through observing variety of data sources
Information tracking	Tracking dynamics of information diffusion across languages, cultures and media	Transforming observed textual and social data streams into actionable deep knowledge representations	Prediction of future events and identification of causal relationships from textual and social streams
Multimodal data processing	Joining textual data and social networks, including spatial and temporal dimensions	Joining textual and social data with unstructured sources like sensor data (smart cities), video, images, audio	Detecting inconsistencies, gaps and completeness of collected knowledge from textual and social sources
Visualisation and user interaction	Visualisation of textual and social dynamics	Adaptive human-computer interfaces boosting specific aims in interaction	Adaptive interaction systems for communication with the whole or parts of society
Algorithmic fundamentals	Scalable processing of multimodal data (Big-Data)	Real-time modelling and reasoning on massive textual and social streams	Algorithms and toolkits being able to deal with planetary scale analytics and reasoning with multimodal data

9: Priority Theme 2 – Social Intelligence and e-Participation: Preliminary Roadmap

Furthermore, this priority theme calls for large-scale, incremental, and sustained development and innovation across multiple disciplines (notable language technology and semantic technologies) and, within each community, a certain degree of stacking and fusion of approaches. Therefore, research organisation needs to create strong incentives for early and frequent exchange of technologies among all players involved. A marketplace for generic component technologies and a service-oriented infrastructure for adaptation and composition must be created, to balance performance-based stearing and self-organisation among clusters of contributing players. In this ecosystem of technology providers and integrators, component uptake by others and measurable contributions against the targeted breakthrough of the priority theme at large should serve as central measures of success.

6.4 PRIORITY THEME 3: SOCIALLY AWARE INTERACTIVE ASSISTANTS

6.4.1 Solutions for the EU Society and for the Citizen

Socially aware interactive assistants are conversational agents. Their socially-aware behaviour is a result of combining analysis methods for speech, non-verbal and semantic signals.

Now is the time to develop and make operational socially aware, multilingual assistants that support people interacting with their environment, including human-computer, human-artificial agent (or robot), and computer-mediated human-human interaction. The assistants must be able to act in various environments, both indoor (such as meeting rooms, offices, appartments),

outdoor (streets, cities, transportation, roads) and virtual environments (such as the web, virtual worlds, games), and also be able to communicate, exchange information and understand other agents' intentions. They must be able to adapt to the user's needs and environment and have the capacity to learn incrementally from all interactions and other sources of information.

The ideal socially aware multilingual assistant can interact naturally with humans, in any language and modality. It can adapt and be personalised to individual communication abilities, including special needs (for the visual, hearing, or motor impaired), affections, or language proficiencies. It can recognise and generate speech incrementally and fluently. It is able to assess its performance and recover from errors. It can learn, personalise itself and forget through natural interaction. It can assist in language training and in education in general, and provide synthetic multimedia information analytics. It recognises people's identity, and their gender, language or accent. If the agent is embodied in a robot, it can move, manipulate objects, and interact with people.

This priority theme includes several components:

- Interacting naturally with humans (in games, entertainment, education, communication, etc.) in an implicit (proactive) or explicit (spoken dialogue and/or gesticulation) manner based on robust analysis of human user identity, age, gender, verbal and nonverbal behaviour, and social context;
- Exhibiting robust performance everywhere (indoor and outdoor environments, mobile applications, augmented reality);
- Overcoming handicap obstacles by means of suitable technologies (sign language understanding, assistive applications, adapted communication to suit cognitively impaired, etc.);
- Interacting naturally with and in groups (in social networks, with humans or artificial agents/robots);
- Exhibiting multilingual proficiency (speech-to-speech translation, interpretation in meetings and videoconferencing, cross-lingual information access);
- Referring to written support (transcription, close-captioning, reading machines, ebooks);
- Providing personalised training (computer-assisted language learning, e-learning in general).

6.4.2 Novel Research Approaches and Targeted Breakthroughs

In addition to significantly improving core speech and language technologies, the development of socially aware interactive assistants requires several research breakthroughs. With regard to speech recognition, accuracy (open vocabulary, any speaker) and robustness (noise, cross-talking, distant microphones) have to be improved. Methods for self-assessment, self-adaptation, personalisation, error-recovery, learning and forgetting information, and also for moving from recognition to understanding have to be developed. Concerning speech synthesis, voices have to be made more natural and expressive, control parameters have to be included for linguistic meaning, speaking style, emotion etc. They also have to be equipped with methods for incremental conversational speech, including filled pauses and hesitations. Likewise, speech recognition, synthesis and understanding have to be integrated, including different levels of evaluation and different levels of automated annotation.

Human communication is multimodal (including speech, facial expressions, body gestures, postures, etc.), crossmodal and fleximodal: it is based on pragmatically best suited modalities. Semantic and pragmatic models of human communication have to be devloped. These have to be context-aware and model situational interdepedencies between context and modalities for arriving at robust communication analysis (multimodal content analytics, infering knowledge from multiple sensory modalities). They have to be able to detect and

recover interactively from mistakes, learning continuously and incrementally. Parsing has to model temporal inter-dependencies within and between modalities in order to maximise the assistant's human-communication-prediction ability. In order to be able to design technologies, adequate semantically and pragmatically annotated language and multimodal resources have to be produced.

A common push has to be made towards more natural dialogue. This includes, among others, the recognition and production of paralinguistics (prosody, visual cues, emotion) and a better understanding of socio-emotional functions of communicative behaviour, including group dynamics, reputation and relationship management. More natural dialogue needs more advanced dialogue models that are proactive (not only reactive), that are able to detect that recognised speech is intended as a machine command, they have to be able to interpret silence as well as direct and indirect speech acts (including lies and humour). Another prerequisite for more natural dialogue is the ability of the assistant to personalise itself to the user's preferences. The digital assistant has to operate in a transparent way and be able to participate in multi-party conversations and make use of other sensory data (GPS, RFID, cameras etc.).

There is also a strong connection to the first priority theme: the multilingual assistant should be able to do speech-to-speech translation in human-human-interaction (e. g., in meetings) and to deal with different languages, accents and dialects effectively. Systems developed should also cover at least all official languages of the EU and several regional languages.

6.4.3 Solution and Technological Realisation

The technological and scientific state-of-the-art is at a stage that allows tackling the development of socially aware multilingual assistants. Progress in machine learning, including adaptation, unsupervised learning from streams of data, continuous learning, and transfer learning makes it possible automatically to learn certain capabilities from data. In addition, existing language and multimodal resources enable the bootstraping of systems. Furthermore, there is interdisciplinary progress made in, e. g., social signal processing.

Technological advances are continuously being achieved in the vision-based human behaviour analysis and synthesis fields. Ubiquitous technologies are now widely available (at lower costs and in reduced size). User-centric approaches have been largely studied and crowd-sourcing is being more and more widely used. Quantitative and objective language technology and human-behaviour understanding technology evaluations, allowing for assessing a technological readiness level (TRL), are carried out more widely, as best practice, and language resources and publicly-available annotated recordings of human spontaneous behaviour are now available.

However there are still some prohibitive factors. Language technology evaluation is still limited and is not conducted for all languages. There is limited availability of language resources, and the necessary resources do not exist yet for all languages. Similarly, publicly-available recordings of spontaneous (rather than staged) human behaviour are sparse, especially when it comes to continuous synchronised observations of multi-party interactions. Limited progress of the technology for automatic understanding of social behaviour like rapport, empathy, envy, conflict, etc., is mainly attributed to this lack of suitable resources. In addition, we still have very limited knowledge of human language and human behaviour perception processes and automated systems often face theoretical and technological complexity of modelling and handling these processes correctly.

6.4.4 Impact

The impact of this priority theme will be wide-ranging. It will impact the work environment and processes, cre-

Research Priority	Phase 1: 2013-2014	Phase 2: 2015-2017	Phase 3: 2018-2020
Interacting naturally with agents	**Provide usable human interface**, reliable speech recognition, natural and intelligible speech synthesis, limited understanding and dialogue capabilities	**Provide usable dialogue interface**, context and dialogue aware speech recognition and synthesis. Recognize and produce emotions, understanding capabilities, context aware dialogue, using other sensors (GPS, RFID, cameras, etc.)	**Provide multiparty (human-agents) interface**, multiple voices, mimicking, advanced understanding and advanced personalised dialogue (indirect speech acts, incl. prosodics, lies, humor)
Using language and other modalities (in parallel or together)	**Multimodal interaction** (speech, facial expression, gesture, body postures)	**Multimodal dialogue**, fusion and fission	**Fleximodal dialogue**, identification of best suited modalities
Conscious of its performing capacities	Confidence in hearing/understanding, interactively recovering from mistakes	Ability to learn continuously and incrementally from mistakes by interaction	Unsupervised learning/forgetting
Exhibiting multilingual proficiency	Ensure availability or portability to major EU languages; recognize which language is spoken; multilingual access to multilingual information	More languages (migrants, foreign languages), accents and dialects; recognize dialects, accents; exploit limited resources; crosslingual access to information	Speech translation in human-human interactions (multiple speakers speaking multiple languages); cross-cultural support; learn new language with small effort
Resources	Install infrastructure, collection of multi-task benchmark data, collaborative production of semantically annotated data (multimodal), incremental production of dialogue data	Use infrastructure, more data, more languages	Use infrastructure, more data, more languages
Evaluation	Multi-task benchmark evaluation; measures and protocols for automated speech synthesis; dialogue systems and speech translation evaluation	Measure of progress; more languages	Measure of progress; more languages

10: Priority Theme 3 – Socially-Aware Interactive Assistants: Preliminary Roadmap

ativity and innovation, leisure and entertainment, and the private life. Several societal and economical facts call for, but also allow for, improved and more natural interaction between humans and the real world through machines. The ageing society requests ambient intelligence. Globalisation involves the capacity to interact in many languages, and offers a huge market for new products fully addressing that multilingual necessity.

The automation of society implies more efficiency and a 24/7 availability of services and information, while green technologies, such as advanced videoconferencing, need to be prioritised. The continuously reduced costs and speed improvement of hardware allow for affordable and better technologies, that can now easily be made available online through app stores.

At the same time we still face prohibitive factors. The cultural, political and economical dimensions of language are well perceived, but not its technical dimension. There is still a psychological barrier for communicating with machines, although this gets more and more common through the use of smartphones and applications such as Skype or Facetime. There is an extra cost for developing personalised systems and the business models are difficult to define as humans are used to communicating at no cost.

6.4.5 Organisation of Research

In order to improve research efficiency within a public-private partnership, the preferred infrastructure would be to handle the various applications in connection with the cooperative development of technologies, including the evaluation of progress, and the production of the language and human naturalistic behaviour resources which are necessary to develop and test the technologies.

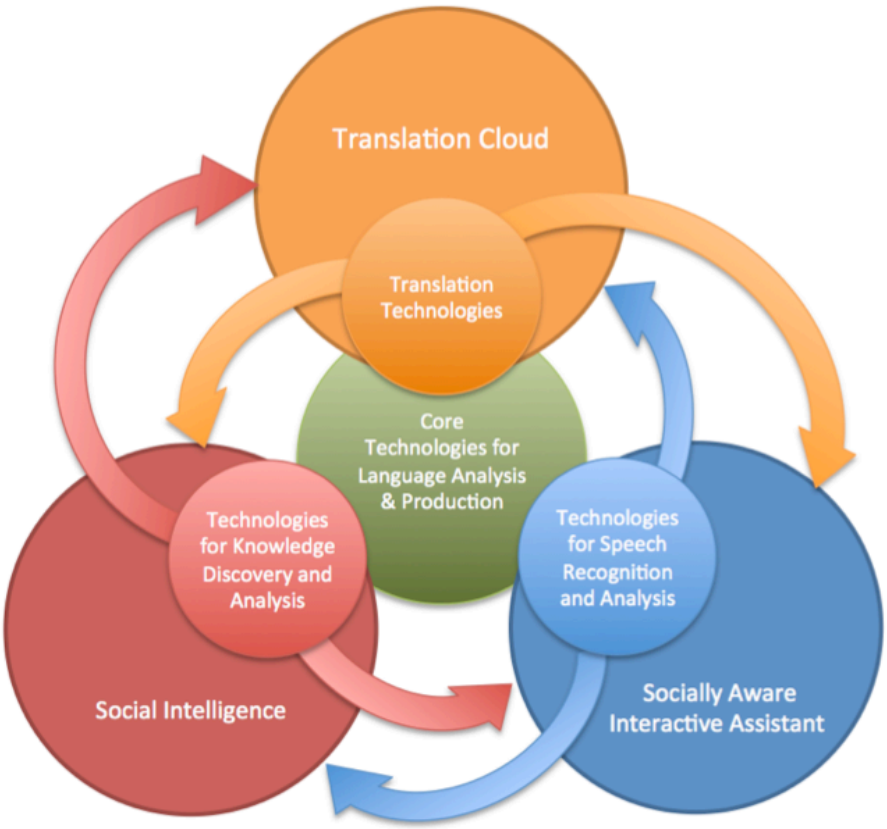
To maximise impact, it is necessary to make a substantial effort in the development of integrated systems based on open architectures, and a multilingual middleware to enable the developed functionalities to be incorporated in a wide range of software. This might best be achieved through a small number of coordinating projects, attached to a federation of strategic projects with complementary goals. These projects should be objective-driven, with clear research, technology and exploitation milestones, coordinated by an on-going road-mapping effort. This includes the production of adequate language and human naturalistic behaviour corpora, semantically annotated including prosodic and non-verbal behavioural cues. This also includes the production (acquisition and annotation) of dialogue corpora from the real world, which implies an incremental system design, and either the use of synchronised continuous observations of all involved parties, or the use of similar data available online (conversations, talks shows).

Dialogue systems evaluation still needs research investigations on the choice of adequate metrics and protocols. The multilingual dimension that is targeted implies the availability of language resources and language technology evaluation for all languages. Handling them all together reduces however the overall effort, given the possibility to use the same best practices, tools and protocols.

6.5 STRUCTURE AND PRINCIPLES OF RESEARCH ORGANISATION

From the description of the three priority themes one can easily see that the proposed research strands overlap in technologies and challenges – this is intended. The overlap reflects the coherence and maturation of the field. At the same time, the resulting division of labour and sharing of resources and results is a precondition for the realisation of the highly ambitious program.

All three themes need to benefit from progress in core technologies of language analysis and production such as morphological, syntactic and semantic parsing and generation. But each of the three areas will concentrate on one central area of language technology: the Translation Cloud will focus on cross-lingual technologies such



11: Scientific cooperation among the three priority research themes

as translation and interpretation; the Social Intelligence strand will take care of knowledge discovery, text analytics and related technologies; the research dedicated to the Interactive Assistant will take on technologies such as speech and multimodal interfaces (see Figure 11).

Except for a few large national projects and programmes such as Techno-langue and Quaero in France, Verbmobil and Theseus in Germany and DARPA Communicator and GALE in the US the field of language technology does not have experience with research efforts of the magnitude and scope required for the targeted advances and plans in this SRA. Nevertheless, our technology area has to follow developments in other key engineering disciplines and speed up technology evolution by massive collaboration based on competitive division of labour and sharing of resources and results. In our reflection on optimal schemes for organizing we tried to draw lessons from our own field's recent history and to capitalise on experience from other fields by adopting approaches that proved successful and evading encountered pitfalls.

The final model for the organisation of collaboration will have to be guided by a thoughtful combination of the following basic approaches.

Flexible collaborative approach: For each priority theme, one or several very large cooperating and competing lead projects will share an infrastructure for evaluation, resources (data and base technologies), and communication. Mechanisms for reducing or terminating partner involvements and for adding new partners or subcontracted contributors should provide flexibility. A number

of smaller projects including some national and regional projects will provide building blocks for particular languages, tasks, component technologies or resources. A cooperation scheme will be designed for effectively involving EC-funding, contributions from member states, industrial associations, and language communities. The choice of funding instruments will be determined in due time after.

Staged approach: Two major phases are foreseen (2015-2017, 2018-2020). For better concertation the major phases should be synchronised among the themes and also projects.

Evolutionary approach: Instead of banking on one selected paradigm, competing approaches will be followed in parallel with shared schemes for evaluation, merging, adopting and discontinuing research threads so that the two elements of successful evolutionary research approaches, selection and cross-fertilisation, are exploited to the maximum extent possible.

Analytical approach: Instead of the currently predominant search for an ideal one-fits-all approach, the research will focus on observed quality barriers and not shun computationally expensive dedicated solutions for overcoming particular obstacles.

Bootstrapping approach: Better systems can be derived from more and better data and through new insights. In turn, improved systems can be used to gain better data and new insights. Thus the combination of the analytical evolutionary approach with powerful machine learning techniques will be the basis for a technology bootstrapping, which has been the by far most fruitful scheme for the development of highly complex technologies.

Close cooperation with relevant areas of service and technology industries: In order to increase chances of successful commercialisation and to obtain convincing and sufficiently tested demonstrations of novel applications, the relevant industrial sectors of industry must be strongly integrated into the entire research cycle.

Tighter research-innovation cycle: Through the collaboration between research, commercial services and commercial technology industries, especially through the shared evaluation metrics and continuous testing, the usual push-model of technology transfer will hopefully substituted by a pull-model, in which commercial technology users can ask for specific solutions. In the envisaged research scheme incentives will be created for competing teams each composed of researchers, commercial users and commercial developers by the participating enterprises for initiating successful innovations

Interdisciplinary approach: A number of science, technology and service areas need to be integrated into the research from day one. Some technology areas such as speech technologies, language checking and authoring systems need to be represented by providers of state-of-the-art commercial products.

Supporting research and innovation in language technology should be accompanied by policy making in the area of multilingualism, but also in digital accessibility. Overcoming language barriers can greatly influence the future of the EU. Solutions for better communication and for access to content in the native languages of the users would reaffirm the role of the EC to serve the needs of the EU citizens. A substantial connection to the infrastructural program CEF could help to speed up the transfer of research results to badly needed services for the European economy and public.

At the same time, use cases should cover areas where the European societal needs massively overlap with business opportunities to achieve funding investment that pays back, ideally public-private partnerships.

The coordination among the three research strands poses administrative challenges. Because of the described interdependencies and also because of the need to maintain and improve the obtained level of cohesion and community spirit in the European Language Technology community, a coordinating body is needed. Whether such an

entity is jointly carried by the three areas or by a separate support project, needs to be determined in the upcoming discussion on the appropriate support instruments for the identified research priorities.

6.6 CORE LANGUAGE RESOURCES AND TECHNOLOGIES

The three priority research themes share a large and heterogeneous group of core technologies for language analysis and production that provide development support through basic modules and datasets (see Figure 11, p. 53). To this group belong tools and technologies such as, among others, tokenisers, part-of-speech taggers, syntactic parsers, tools for building language models, information retrieval tools, machine learning toolkits, speech recognition and speech synthesis engines, and integrated architectures such as GATE and UIMA. Many of these tools depend on specific datasets (i. e., language resources), for example, very large collections of linguistically annotated documents (monolingual or multilingual, aligned corpora), treebanks, grammars, lexicons, thesauri, ontologies and language models. Both the basic tools and especially language resources can be rather general or highly task- or domain-specific, available for free or for a fee, tools can be language-independent, datasets are, by definition, language-specific. As complements to the core technologies and resources there are several types of resources, such as error-annotated corpora for machine translation or spoken dialogue corpora, that are specific to one or more of the three priority themes.

A key component of the suggested research agenda is to collect, develop and make available core technologies and resources through a shared infrastructure so that the research and technology development carried out in all themes can make use of them. Over time, this approach will improve the core technologies, as the specific research will have certain requirements on the software, extending its feature set, performance, accuracy etc. through dynamic push-pull effetcs. Conceptualising these technologies as a set of shared core technologies will also have positive effects on their sustainability and interoperability.

The European academic and industrial technology community is fully aware of the need for sharing resources such as language data (e. g., corpora), language descriptions (e. g., lexicons, thesauri, grammars), tools (e. g., taggers, stemmers, tokenisers) and core technology components (e. g., morphological, syntactic, semantic processing) as a basis for the successful development and implementation of the priority themes. Initiatives such as FLaReNet [33] and CLARIN have prepared the ground for a culture of sharing, META-NET's open resource exchange infrastructure, META-SHARE, is providing the technological platform as well as legal and organisational schemes. It is also important to note that many European languages other than English are heavily under-resourced, i. e., there are no or almost no resources or basic technologies available [14].

All language resources and basic technologies that are created under the core technologies umbrella will be shared and made available through the respective infrastructure. The effort should revolve around the following axes: Infrastructure; Coverage, Quality, Adequacy; Language Resources Acquisition; Openness; Interoperability.

6.6.1 Infrastructure

It is imperative to maintain and further to develop META-SHARE. Broad participation by the whole language technology community is essential in maintaining and extending the infrastructure so that acceptance is ensured. META-SHARE will be the key instrument to make language resources available, visible and accessible, to facilitate sharing and exchange of resources.

Among the important aspects that need to be taken care of when taking the next evolutionary steps of the META-

SHARE infrastructure are the following: definition of the basic data and software resources that should populate META-SHARE, multilingual coverage, the capacity to attract providers of useful resources, improvements in sharing mechanisms, and collaborative working practices between R&D and commercial users. There must also be a business-friendly framework to stimulate the commercial use of resources, based on a sound licensing facility. Close cooperation with the three priority themes is of vital importance, especially for defining the set of needed core technologies and resources.

The content of META-SHARE is not limited to data. Instead, it has to be seen as an international hub of resources and technologies for speech and language services from industries and communities. The development and proposal of ideally free tools and, more generally, web services, including evaluation protocols and collaborative workbenches is deemed essential. The accumulation and sharing of resources and tools in a single place would lower the R&D costs for new applications in new language resource domains.

Sustainability covers preservation, accessibility, and operability (among other things). Collecting and preserving knowledge in the form of existing resources should be a key priority. A sustainability analysis must be part of a resource specification phase, and it is important that funding agencies impose a sustainability plan mandatory for those projects that are concerned with the production of language resources.

Accurate and reliable documentation of resources is an undisputable need. An effort must be made to collect all existing documentations and to make them available as a repository of specifications, guidelines, and documentation of resources. Documentation is also the gateway to resource discovery. Ensuring that resources are discoverable is the first step towards promoting the data economy.

6.6.2 Coverage, Quality, Adequacy

With regard to the data-driven paradigm, innovation in LT nowadays crucially depends on language resources. Despite the vast amount of academic and industrial investment, there are not enough available resources to satisfy the needs of all languages, quantitatively and qualitatively. Language resources should be produced and made available for every language, every register, every domain to guarantee full coverage, high quality and adequacy for various applications. New methods of resource development can be exploited to achieve better coverage, for instance shared or distributed ones. It is important to assess the availability of existing resources with respect to their adequacy to applications and technology requirements. This involves assessing the maturity of the technologies for which new resources should be developed. Specifically for the advancement of LTs, basic language resource kits should be supported and developed for all languages and, at least, key applications.

To reduce the amount of human intervention and revision, automatic techniques should be promoted to guarantee quality through error detection and confidence assessment. The promotion of validation and evaluation can play a valuable role in fostering quality improvement. Evaluation should encompass technologies, resources, guidelines and documentation. But like the technologies it addresses, evaluation is constantly evolving, and new, more specific measures using innovative methodologies are needed to evaluate the reliability of language resources, while maximal use of existing tools should be ensured for the validation of resources.

A "Language Resources Impact Factor (LRIF)" should be defined in order to enforce the practice of citation of resources on the model of scientific paper authoring and to calculate the actual research impact of resources. A reference model for creating resources will help address the current shortage of resources in terms of breadth (languages and applications) and depth (quality and volume).

6.6.3 Language Resources Acquisition

Re-use and re-purposing should be encouraged to ensure the reuse of development methods and existing tools. With production costs constantly increasing, there is a need to invest in innovative production methods that involve automatic procedures, so as to reduce human intervention to a minimum. The coverage problem is so enormous that strategies that approach or ensure full automation for high-quality resource production should be promoted. It is worth considering the power of social media to build resources, especially for those languages where there are no language resources built by experts yet.

There are several promising experiments in crowdsourcing data collection tasks. Crowd-sourcing makes it possible to mobilise large groups of human talent around the world with just the right language skills so that we can collect what we need when we need it. For instance, it has been estimated that Mechanical Turk translation is 10 to 60 times less expensive than professional translation.

6.6.4 Openness

There is a strong trend towards open data, i. e., data that are easily obtainable and that can be used with few, if any, restrictions. Sharing resources (both data and tools) has become a viable solution towards encouraging open data, and the community is strongly investing in facilities such as META-SHARE for the discovery and use of resources. These facilities could represent an optimal intermediate solution to respond to the needs for data variety, ease of retrieval, better data description and community-wide access, while at the same time assisting in clearing the intricate issues associated with intellectual property rights.

The challenge for the community and policy makers is to push for the development of a common legal framework that would facilitate resource sharing efforts abiding by the law, benefiting from the adoption of "fair use" principles and appropriate copyright exceptions. It is of utmost importance that legislation regarding resource use be harmonised, and even standardised, for all types of resources, and that free use be allowed, at least for research or non-profit purposes.

6.6.5 Interoperability

Interoperability of resources seeks to maximise the extent to which they are compatible and therefore integratable at various levels, so as to allow, for instance, the merging of data or tools coming from different sources. The community and the funding agencies need to join forces to drive forward the use of existing and emerging standards, at least in the areas where there is some degree of consensus. The only way to ensure useful feedback to improve and advance is to use standards on a regular basis. It will be thus even more important to enforce and promote the use of standards at all stages.

6.6.6 Organisation of Research

In order to optimise the efficiency of shared core technologies for language analysis and production as well as the further development of the infrastructure, maximise the infrastructure's impact, and ensure that requirements for research and development are met at the necessary depth for all languages in all priority themes, the organisation of this shared component of the research agenda should adopt the following principles: It is necessary to invest in the further development of an integrated infrastructure (i. e., META-SHARE) based on an open architecture, enabling the sharing and further development of resources. The infrastructure should support technology-specific challenges and shared tasks in order to accelerate innovation breakthrough and market-readiness for desperately needed technologies. Concerted activities and policies facilitating the sharing of resources overcoming all stumbling blocks on the way to technical, organisational and legal interoperability should be supported. EU level research must be aligned and tightly coordinated with efforts on the national levels, so that language cov-

Research Priority	Phase 1: 2013-2014	Phases 2 and 3: 2015-2020
Infrastructure	Maintain and extend facility(-ies) for sharing resource data and tools; promote accurate and reliable documentation of resources through metadata; cooperation between infrastructure initiatives to avoid the duplication of effort	Create mechanisms for accumulating descriptions of as well as actual resources; multilingual coverage, ease of conversion into uniform formats; solutions for integrating language processing services to help growth of infrastructures (SaaS)
Coverage, quality, adequacy	Increase quantity of resources available to address language technology and application needs; address formal and content quality of resources by promoting evaluation and validation; promote evaluation and validation activities of resources and the dissemination of their outcomes	Further increase quantity of resources available to address language and application needs; provide high quality resources for all European languages
Acquisition	Ensure public and community support to definition and dissemination of resource production best practices; enforce reusing and repurposing; research work towards the full automation of LR data production; invest in methods for collaborative creation and extension of high-quality resources, also as a means to achieve better coverage; implement workflows of language processing services for acquisition of resources required for the implementation of the priority themes; bridge acquisition methods with linked open data and big data; share the effort for production of LRs between international bodies and individual countries	
Openness	Educate key players with basic legal know-how; elaborate specific, simple and harmonised licensing solutions for data resources; promote copyright exception for research purposes; develop legal and technical solutions for privacy protection; opt for openness of resources, especially publicly funded ones; ensure that publicly funded resources are publicly available free of charge; clear IPR at the early stages of production; try to ensure that re-use is permitted	
Interoperability	Invest in standardisation activities, make standards operational and put them in use; create permanent Standards Observatory or Standards Watch; promote and disseminate standards to students and young researchers; encourage/enforce use of best practices or standards in production projects; identify new mature areas for standardisation and promote joint efforts between R&D and industry	

12: Core language resources and technologies: Preliminary Roadmap

erage and language-specific developments are efficiently achieved. An important aspect of this coordination effort is concerned with the results of the META-NET White Paper Series: in the 30 different white papers we have concrete and specific assessments of the language- and country-specific situation with regard to demands and technology gaps. The next step is to address and to fill these gaps with high-quality and robust core technologies and language resources. In addition we will continue to collect relevant technologies and resources for inclusion in and distribution through META-SHARE.

6.7 CHALLENGES FOR INNOVATION

Language Technology is innovation-friendly in the sense that many solutions are not standardised, but require individual adaptation or new development for a certain customer or application. Thus, one can truly speak of socially responsible innovation here.

As there are many niches in the market that are not targeted by the big players, SMEs have real opportunities. At the same time, language technology, as a key-enabling technology, usually enters the markets in combination with other technologies as an essential component of novel products and services that can be arbitrarily complex, which has made it difficult for SMEs to identify customers in the past.

The question is how to transform innovation and research into new products, markets, growth, and, finally, new jobs. In recent years, drivers for innovations have often been applications and tools such as Skype, Facebook, or recommender systems that have been designed by smaller teams and start-ups. Important aspects of their success stories, besides the core and novel functionalities and feature sets for which there was an obvious need, often was their fast and viral outreach and uptake through social networks.

Large global platforms for novel end-user-services have also become the predominant innovation drivers for language technology solutions. These platforms can be web services such as Google Search that integrates the new Knowledge Graph concept network, speech-enabled search and also web translation. Combinations of hardware and operating systems such as iOS for Apple's mobile devices iPhone and iPad can also be considered platforms. Or it could be an open operating system such as Android which recently extended its current speech and language functionalities with a mobile assistant.

The trend towards widely used platforms will drastically facilitate the spreading of innovative language technologies. Actually, language technology has a good chance of becoming the essential feature for the success of the next generation of services. At closer inspection, the integration of sophisticated language technology in current platforms is rather limited, scratching only the surface of what will be possible in the near future.

Apart from new ways of sharing, development, and distribution, a generally innovative climate is needed. The availability of venture capital and meeting points like summits where research and decision makers from industry get together should be backed by public funding, and uptake. Flexible funded consortia that run over a longer period with changing partners, where research and innovation phases lead over to product development and marketing would also support innovation.

6.8 A EUROPEAN SERVICE PLATFORM FOR LANGUAGE TECHNOLOGIES

We argue for the creation of an ambitious large-scale sky-computing platform as a central motor for research and innovation in the next phase of IT evolution and a ubiquitous resource for the multilingual European society (an idea suggested by several experts from industry in META-

NET Vision Group meetings). The platform will be used for testing, show casing, proof-of-concept demonstration, avant-garde adoption, experimental and operational service composition, and fast and economical service delivery to enterprises and end-users.

The proposed creation of a powerful cloud or sky computing platform (see Section 3.6) for a wide range of services dealing with human language, knowledge and emotion will not only benefit the individual and corporate users of these technologies but also the providers.

Users will be able to receive customised integrated services without having to install, combine, support and maintain the software. They will have access to specialised solutions even if they do not use these regularly.

Language technology providers will have ample opportunity to offer services stand-alone or integrated with others.

Providers of language services rendered by human language professionals will be able to use the platform for enhancing their services by means of appropriate technology and for providing their services stand-alone or integrated into other application services.

Researchers will have a unique virtual laboratory for testing, combining, and benchmarking their novel technologies and for exposing them in realistic trials to real tasks and end users.

Providers of services that can be enabled or enhanced by text and speech processing will utilise the platform for testing the needed LT functionalities and for integrating them into their own solutions.

Citizens and corporate users will enjoy the benefits of language technology early and at no or reasonable costs through a large variety of generic and specialised services offered at a single source.

In order to allow for the gigantic range of foreseeable and currently not yet foreseeable solutions, the infrastructure will have to host all relevant simple services, including components, tools and data resources, as well as various layers or components of higher services that incorporate simpler ones. This is why META-SHARE will play an important role in the design of the overall platform (see section 6.6).

A top layer consists of **language processing** such as text filters, tokenisation, spell checking, hyphenation, lemmatising and parsing. At a slightly deeper level, services will be offered that realise some degree and form of **language understanding** including entity and event extraction, opinion mining and translation. Both basic language processing and understanding will be used by services that support **human communication** or realise some human-machine interaction. Part of this layer are question answering and dialogue systems as well as email response applications. Another component will bring in services for processing and storing **knowledge** gained by and used for understanding and communication. This part will include repositories of linked data and ontologies, as well as services for building, using and maintaining them. These in turn permit a certain range of rational capabilities often attributed to a notion of intelligence. The goal is not to model the entire human intelligence but rather to realise selected forms of **inference** that are needed for utilising and extending knowledge, for understanding and for successful communication. These forms of inference permit better decision support, pro-active planning and autonomous adaptation. A final part of services will be dedicated to **human emotion**. Since people are largely guided by their emotions and strongly affected by the emotions of others, truly user-centred IT need facilities for detecting and interpreting emotion and even for expressing emotional states in communication.

We consider the paradigm of federated cloud services or sky computing with its emerging standards such as OCCI, OVM and CDMI and toolkits such a OpenNebula as the appropriate approach for realising the ambitious infrastructure. All three priority areas of this SRA will be able to contribute to and at the same time draw im-

mense benefits from this platform. There are strong reasons for aiming at a single service platform for the three areas and for the different types of technologies. They share many basic components and they need to be combined for many valuable applications, including the selected showcase solution of the three areas.

6.8.1 Implementation of the Platform

The creation of the platform, for which a name has yet to be found, has to be supported by public funding. Because of the high requirements concerning performance, reliability, user support, scalability, persistence as well as data protection and conformance with privacy regulation, the platform needs to be established by a consortium with strong commercial partners and also be operated by this consortium or a commercial contractor. A similar platform with slightly different desiderata and functionalities is currently built under the name Helix-Nebula for the Earth Sciences with the help of the following commercial partners: Atos, Capgemini, CloudSigma, Interoute, Logica, Orange Business Services, SAP, SixSq, Telefonica, Terradue, Thales, The Server Labs and T-Systems. Partners are also the Cloud Security Alliance, the OpenNebula Project and the European Grid Infrastructure. These are working together with major research centres in the earth sciences to establish the targeted federated and secure high-performance computing cloud platform.

The intended platform for LT and neighbouring fields would be intended for a mix of commercial and non-commercial services. It would be cost-free for all providers of non-commercial services (cost-free and advertisement-free) including research systems, experimental services and freely shared resources but it would raise revenues by charging a proportional commission on all commercially provided services. In order to reduce dependence on individual companies and software products, the base technology should be supplied by open toolkits and standards such as OpenNebula and OCCI.

For each priority research theme, chances for successful showcasing and successful commercial innovation will increase tremendously if usable services could be offered on such a platform of required strength and reliability.

The platform will considerably lower the barrier for market entry for innovative technologies, especially for products and services offered by SMEs. Still, these stakeholders may not have the resources, expertise, and time to create the necessary interfaces to integrate their results into real-life services, let alone the overarching platform itself. There is still a gap between research prototypes and products that have been engineered and tested for robust applications. Moreover, many innovative developments require access to special kind of language resources such as recordings of spoken commands to smartphones, which are difficult to get for several reasons.

Thus the service platform will be an important instrument for supporting the entire innovation chain, but, in addition, interoperability standards, interfacing tools, middle-ware, and reference service architectures need to be developed and constantly adapted. Many of these may not be generic enough to serve all application areas, so that much of the work in resource service integration will have to take place in the respective priority theme research actions.

7

TOWARDS ROADMAPS AND A SHARED EUROPEAN PROGRAMME FOR MULTILINGUAL EUROPE 2020

7.1 NEXT STEPS

The current version of this Strategic Research Agenda is not the final one, several steps are foreseen before the document will be finalised. The step immediately following the publication of this first initial version is to collect further feedback from META-NET, META and the Language Technology community at large. Even though more than 130 persons have directly contributed to this Strategic Research Agenda (see Appendix B on 68 f.), we would like to collect as much further feedback, testimonials and additional contributions as possible. Among the reasons for sending in feedback can be, for example, significant gaps in the argumentation, interesting figures, data and numbers, convincing quotes and references, ideas for technology use cases, or steps and goals for the roadmap.

> Please send feedback to this SRA to georg.rehm@meta-net.eu with the subject line "META-NET SRA: feedback" or participate in our online discussion forum at http://www.meta-net.eu/forum.

The editors of this Strategic Research Agenda will process all feedback received by September 15, 2012, and include it into the final version of this document which is due for publication in November 2012.

7.2 TOWARDS ROADMAPS

Important components of the final version of this Strategic Research Agenda will be a small set of roadmaps that provide additional details with regard to the actual steps, order, priorities and dependencies of the research foreseen for a total of five areas. In addition to the three priority research themes (Sections 6.2, 6.3 and 6.4), roadmaps have to be prepared for the core language technologies and shared resources area (Section 6.6) and the European service platform for language technologies (Section 6.8).

7.3 TOWARDS A SHARED EUROPEAN PROGRAMME

The plans foreseen in this SRA can be successfully realised and implemented using a number of different measures and instruments, for example, through clusters of projects or a certain number of coordinated projects. Also an option is to set up a shared programme between the European Commission and the Member States as well as Associated Countries. First steps along those lines have been taken at META-NET's META-FORUM 2012 conference in Brussels, Belgium, on June 21, 2012, when representatives of several European funding agencies who participated in a panel discussion on this topic, unanimously expressed the urgent need for such a shared programme. A sizable portion of the research proposed in this SRA

under the umbrella of the three priority themes is to be carried out in the Horizon 2020 programme. The European service platform for language technologies is a very good fit for the Connecting Europe Facility programme (CEF) while large parts of the core technologies for language analysis and production are good candidates for support through national and regional programmes.

There are several options how to organise the research proposed in this strategic agenda. In June 2012 we have started discussing two possible instruments within META-NET that mainly aim at establishing a shared European programme – several other options still have to be screened. The two candidate instruments are an Article 185 Initiative (see Section 7.3.1) and a Contractual Public-Private Partnership (PPP, see Section 7.3.2).

7.3.1 Article 185 Initiative

To quote Article 185 of the Treaty of the Functioning of the European Union (TFEU): "In implementing the multiannual framework programme, the Union may make provision, in agreement with the Member States concerned, for participation in research and development programmes undertaken by several Member States [...]." Currently there are four joint programmes running as Article 185 Initiatives [34]: Ambient Assisted Living (AAL), Baltic Sea research (Bonus), a programme in the field of metrology (EMRP) and a programme for research performing SMEs and their partners (Eurostars).

A key idea behind Article 185 is to coordinate national programmes in order to reduce the fragmentation of research efforts carried out on the national or regional level. Among the goals to be achieved are to reach critical mass in certain research areas, to ensure better use of scarce resources and to find common answers and approaches to common needs and interests. Member states are given the opportunity to exchange good practice, to avoid unnecessary overlaps of efforts, to exchange information and expertise and to learn from each other.

The Seventh Framework Programme states that an Article 185 Initiative can be launched in areas to be identified in close association with the Member States on the basis of a series of criteria: relevance to EU objectives; the clear definition of the objective to be pursued and its relevance to the objectives of the Framework Programme; presence of a pre-existing basis (existing or envisaged research programmes); European added value; critical mass, with regard to the size and the number of programmes involved and the similarity of activities they cover; efficiency of Article 185 as the most appropriate means for achieving the objectives. Each Article 185 Initiative is set up individually through a decision of the European Parliament and of the European Council, following a proposal from the European Commission.

7.3.2 Contractual Public-Private Partnership

While many details of the upcoming programme Horizon 2020 are still under discussion, Contractual PPPs are currently emerging as the primary model to implement parts of the programme objectives with regard to sizeable, roadmap-based research and innovation efforts within the technology pillar of H2020, drawing also on resources beyond the EU support and related matching funds. The EC's proposal for Horizon 2020 states that "greater impact should also be achieved by combining Horizon 2020 and private sector funds within public-private partnerships in key areas where research and innovation could contribute to Europe's wider competitiveness goals and help tackle societal challenges" [35]. PPPs are an important mechanism for focusing research and innovation, ensuring stakeholders engagement and, above all, for improving the impact of EU support on Europe's competitiveness, growth and jobs creation. A public-private partnership is defined as "a partnership where private sector partners, the Union and, where appropriate, other partners, commit to jointly support the development and implementation of a research and innovation

programme or activities". Similar instruments are JTIs (Joint Technology Initiatives), ETPs (European Technology Platforms) and institutional PPPs which are a counterpart to Contractual PPPs.

For Contractual PPPs, a Contractual Agreement is foreseen between the EC and private and public partners that specifies the objectives of the partnership, commitments of the partners, target outputs and the activities that require support from Horizon 2020. PPPs are to be identified in an open and transparent way based on all of the following criteria: the added value of action at Union level; the scale of impact on industrial competitiveness, sustainable growth and socio-economic issues; the long-term commitment from all partners based on a shared vision and clearly defined objectives; the scale of the resources involved and the ability to leverage additional investments in research and innovation; a clear definition of roles for each of the partners and agreed key performance indicators over the period chosen (see [35], p. 21).

In contrast to an Article 185 Initiative, setting up a contractual PPP does not require a decision in the European Parliament. This is why a PPP for the priority research themes specified in this Strategic Research Agenda might be a promising avenue.

7.3.3 Conclusions

The research plans specified in this SRA are, among others, a good match for an Article 185 Initiative and also for a Contractual PPP. It remains to be discussed which instrument or maybe even set of carefully selected and compiled instruments is considered the most appropriate one to realise and implement the three priority research themes, the set of core technologies and shared resources and also the European service platform for language technology.

REFERENCES

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (Maschinelle Verarbeitung gesprochener und geschriebener Sprache)*. Prentice Hall, 2nd edition, 2009.

[2] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing (Grundlagen der statistischen Sprachverarbeitung)*. MIT Press, 1999.

[3] Language Technology World (LT World). http://www.lt-world.org.

[4] Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Battista Varile, Annie Zaenen, and Antonio Zampolli, editors. *Survey of the State of the Art in Human Language Technology (Sprachtechnologie: Überblick über den Stand der Kunst)*. Cambridge University Press, 1998.

[5] European Commission. A Digital Agenda for Europe, 2010. http://ec.europa.eu/information_society/digital-agenda/publications/.

[6] European Commission. Multilingualism: an Asset for Europe and a Shared Commitment, 2008. http://ec.europa.eu/languages/pdf/comm2008_en.pdf.

[7] The Council of the European Union. Council Resolution of 21 November 2008 on a European strategy for multilingualism, November 2008. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2008:320:0001:01:en:HTML.

[8] Directorate-General for Translation of the European Commission. Size of the Language Industry in the EU, 2009. http://ec.europa.eu/dgs/translation/publications/studies.

[9] European Commission. Languages mean business, 2011. http://ec.europa.eu/languages/languages-mean-business/.

[10] European Commission. Horizon 2020: The Framework Programme for Research and Innovation, 2012. http://ec.europa.eu/research/horizon2020/.

[11] European Commission. Connecting Europe Facility: Commission adopts plan for €50 billion boost to European networks, 2011. http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1200.

[12] European Commission. Languages, 2012. http://ec.europa.eu/languages/.

[13] The Language Rich Europe Consortium. *Towards a Language Rich Europe. Multilingual Essays on Language Policies and Practices*. British Council, July 2011. http://www.language-rich.eu/fileadmin/content/pdf/LRE_FINAL_WEB.pdf.

[14] Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*. Springer, 2012. This series comprises 30 volumes on the following 30 European languages: Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Spanish, Swedish. http://www.meta-net.eu/whitepapers.

[15] Gianni Lazzari. Human Language Technologies for Europe, 2006. http://cordis.europa.eu/documents/documentlibrary/90834371EN6.pdf.

[16] Andrew Joscelyne and Roselockwood. The EUROMAP Study. Benchmarking HLT progress in Europe, 2003. http://cst.dk/dandokcenter/FINAL_Euromap_rapport.pdf.

[17] Directorate-General of the UNESCO. Intersectoral Mid-term Strategy on Languages and Multilingualism, 2007. http://unesdoc.unesco.org/images/0015/001503/150335e.pdf.

[18] Directorate-General Information Society & Media of the European Commission. User Language Preferences Online, 2011. http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

[19] Daniel Ford and Josh Batson. Languages of the World (Wide Web), July 2011. http://googleresearch.blogspot.com/2011/07/languages-of-world-wide-web.html.

[20] Eric Fisher. Language communities of Twitter (European detail), October 2011. http://www.flickr.com/photos/walkingsf/6276642489/.

[21] Donald A. DePalma and Nataly Kelly. The Business Case for Machine Translation. How Organizations Justify and Adopt Automated Translation, August 2009. Common Sense Advisory. http://www.commonsenseadvisory.com/AbstractView.aspx?ArticleID=859.

[22] Franz Och. Breaking down the language barrier – six years in, April 2012. http://googleblog.blogspot.de/2012/04/breaking-down-language-barriersix-years.html.

[23] European Commission. Report on cross-border e-commerce in the EU, 2009. http://ec.europa.eu/consumers/strategy/docs/com_staff_wp2009_en.pdf.

[24] UN Department of Economic and Social Affairs Population Division. International Migration Report 2002, 2002. http://www.un.org/esa/population/publications/ittmig2002/2002ITTMIGTEXT22-11.pdf.

[25] Declaration of Principles – Building the Information Society: a global challenge in the new Millennium, December 2003. http://www.itu.int/wsis/docs/geneva/official/dop.html.

[26] Directorate-General of the UNESCO. Information for All Programme (AFP), 2011. http://www.unesco.org/new/en/communication-and-information/intergovernmental-programmes/information-for-all-programme-ifap/.

[27] Ford Motor Company. Fact Sheet: Ford SYNC Voice-Controlled Communications and Connectivity System, 2012. http://media.ford.com/article_display.cfm?article_id=33358.

[28] European Commission: Bureau of European Policy Advisors. Empowering people, driving change: Social Innovation in the European Union, May 2011. http://ec.europa.eu/bepa/pdf/publications_pdf/social_innovation.pdf.

[29] Global Industry Analysts. Spech Technology: A Global Strategic Business Report, March 2012. http://www.strategyr.com/Speech_Technology_Market_Report.asp.

[30] European Commission. European Interoperability Framework (EIF) for European public services, 2010. http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf.

[31] Moses – Statistical Machine Translation System, 2012. http://www.statmt.org/moses/.

[32] *Workshop on Language Technology for Decision Support at the Fourth Swedish Language Technology Conference*, 2012. http://permalink.gmane.org/gmane.science.linguistics.corpora/15911.

[33] Nicoletta Calzolari, Nuria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, and Claudia Soria. Language Resources for the Future – The Future of Language Resources, September 2011. http://www.flarenet.eu/sites/default/files/FLaReNet_Book.pdf.

[34] European Commission. Article 185 Initiatives, 2012. http://cordis.europa.eu/fp7/art185/.

[35] European Commission. Proposal for a Regulation of the European Parliament and of the Council establishing Horizon 2020 – The Framework Programme for Research and Innovation (2014-2020), 2011. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0809:FIN:en:PDF.

[36] Georg Rehm and Hans Uszkoreit. Multilingual Europe: A challenge for language tech. *MultiLingual*, 22(3):51–52, April/May 2011.

[37] Aljoscha Burchardt, Georg Rehm, and Felix Sasaki. The Future European Multilingual Information Society – Vision Paper for a Strategic Research Agenda, 2011. http://www.meta-net.eu/vision/reports/meta-net-vision-paper.pdf.

[38] Aljoscha Burchardt, Georg Rehm, and Felix Sasaki. LT 2020. Vision and Priority Themes for Language Technology Research in Europe until the Year 2020. Towards the META-NET Strategic Research Agenda, 2012. http://www.meta-net.eu/vision/reports/LT2020.pdf.

B

LIST OF KEY CONTRIBUTORS

The experts listed in the following contributed to this Strategic Research Agenda (55% from Language Technology User or Provider Industries, 49% from Language Technology Research, 2,3% from national or international institutions), which was edited by the META Technology Council.

1. Sophia Ananiadou, University of Manchester, UK
2. Toni Badia, Barcelona Media, Spain
3. Michaela Bartelt, Electronic Arts, Germany/USA
4. Christoph Bauer, ORF, Austria
5. Nozha Boujemaa, INRIA, France
6. Hervé Bourland, IDIAP, Switzerland
7. Antonio Branco, University of Lisbon, Portugal
8. Andrew Bredenkamp, acrolinx, Germany
9. Gerhard Budin, University of Vienna, Austria
10. Axel Buendia, Spir. Ops, France
11. Aljoscha Burchardt, DFKI, Germany
12. Will Burgett, Intel, USA
13. Johannes Bursch, Daimler AG, Germany
14. Nicoletta Calzolari, Consiglio Nazionale delle Ricerche, Italy
15. Nick Campbell, Trinity College Dublin, Ireland
16. Jean Carrive, INA, France
17. Khalid Choukri, ELDA, France
18. Philipp Ciminiano, University of Stuttgart, Germany
19. Ann Copestake, University of Cambridge, UK
20. Ido Dagan, Bar-Ilan University, Israel
21. Morena Danieli, Loquendo, Italy
22. Claude de Loupy, Syllabs, France
23. Maarten de Rijke, University of Amsterdam, The Netherlands
24. Marin Dimitrov, Ontotext, Bulgaria
25. Petar Djekic, SoundCloud, UK
26. Bill Dolan, Microsoft, USA
27. Christoph Dosch, Institut für Rundfunktechnik, Germany
28. Marcello Federico, FBK, Italy
29. David Filip, Moravia, Czech Republic
30. Dan Flickinger, Stanford University, USA
31. Gil Francopoulo, CNRS/LIMSI and IMMI, France
32. Piotr W. Fuglewicz, Micro, Poland
33. Robert Gaizauskas, University of Sheffield, UK
34. Martine Garnier-Rizet, CNRS/LIMSI and IMMI, France
35. Simon Garrett, British Telecom, UK
36. Stefan Geissler, Temis, Germany
37. Edouard Geoffrois, Ministry of Defense and French National Research Agency, France
38. Yota Georgakopolou, European Captioning Institute, UK
39. Serge Gladkoff, Logrus International and GALA Standards Director, USA/Russia
40. Daniel Grasmick, Lucy Software, Germany
41. Gregory Grefenstette, Exalead, France
42. Marko Grobelnik, Institut "Jožef Stefan", Slovenia
43. Joakim Gustafson, KTH Royal Institute of Technology, Sweden
44. Jan Hajic, Charles University Prague, Czech Republic
45. Paul Heisterkamp, Daimler AG, Germany
46. Mattias Heldner, KTH Royal Institute of Technology, Sweden
47. Manuel Herranz, PangeaMT, Spain
48. Theo Hoffenberg, Softissimo, France
49. Thomas Hofmann, Google, Switzerland/USA
50. Timo Honkela, Aalto University, Finland
51. Krzysztof Jassem, Poleng, Poland
52. Keith Jeffery, Science and Technology Facilities Council, Rutherford Appleton Lab., UK
53. Kristiina Jokinen, University of Helsinki, Finland
54. Rebecca Jonson, Artificial Solutions, Spain

55. John Judge, Dublin City University and CNGL, Ireland
56. Martin Kay, Stanford University, USA and Universität des Saarlandes, Germany
57. Christopher Kermorvant, A2iA, France
58. Simon King, University of Edinburgh, UK
59. Philipp Koehn, University of Edinburgh, UK
60. Maria Koutsombogera, ILSP, Greece
61. Steven Krauwer, University of Utrecht, The Netherlands
62. Verena Krawarik, APA, Austria
63. Stefan Kreckwitz, Across, Germany
64. Simon Krek, Institut "Jožef Stefan", Slovenia
65. Brigitte Krenn, OFAI, Austria
66. Michal Küfhaber, Skrivanek, Czech Republic
67. Jimmy Kunzmann, EML, Germany
68. Bernardo Magnini, FBK, Italy
69. Gudrun Magnusdottir, ESTeam, Sweden
70. Elisabeth Maier, CLS Communication, Switzerland
71. Joseph Mariani, CNRS/LIMSI and IMMI, France
72. Penny Marinou, Litterae Trans, Greece
73. Margaretha Mazura, EMF, UK/Belgium
74. Wolfgang Menzel, University of Hamburg, Germany
75. Roger Moore, University of Sheffield, UK
76. Sukumar Munshi, Across, Germany
77. Bart Noe, Jabbla, The Netherlands
78. Jan Odijk, University of Utrecht, The Netherlands
79. Stephan Oepen, University of Oslo, Norway
80. Karel Oliva, Czech Academy of Sciences, Czech Republic
81. Mehmed Özkan, Bogazici University, Turkey
82. Maja Pantic, Imperial College London, UK
83. Alexandre Passant, DERI, Ireland
84. Pavel Pecina, Dublin City University and CNGL, Ireland
85. Manfred Pinkal, Universität des Saarlandes, Germany
86. Stelios Piperidis, ILSP, Greece
87. László Podhorányi, Vodafone, Hungary
88. Jörg Porsiel, VW, Germany
89. Gabor Proszeky, Morphologic, Hungary
90. Artur Raczynski, European Patent Office, Germany
91. Georg Rehm, DFKI, Germany
92. Steve Renals, University of Edinburgh, UK
93. Peter Revsbech, Ordbogen, Denmark
94. Giuseppe Riccardi, University of Trento, Italy
95. Johann Roturier, Symantec, Ireland
96. Dimitris Sabatakakis, Systran, France
97. David Sadek, Institute Telecom, France
98. Sergi Sagàs, MediaPro, Spain
99. Felix Sasaki, W3C and DFKI, Germany
100. Jana Šatková, ACP Traductera, Czech Republic
101. Mirko Silvestrini, Rapidrad, Italy
102. Ruud Smeulders, Rabo Bank, The Netherlands
103. Svetlana Sokolova, ProMT, Russia
104. Juan Manuel Soto, Fonetic, Spain
105. Lucia Specia, University of Sheffield, UK
106. C. M. Sperberg-McQueen, BlackMesa Technologies, USA
107. Volker Steinbiss, RWTH Aachen and Accipio, Germany
108. Rudi Studer, KIT, Germany
109. Katerina Stuparicova, Charles University Prague, Czech Republic
110. Daniel Tapias, Sigma Tech, Spain
111. Alessandro Tescari, Pervoice, Italy
112. Lori Thicke, Translators without Borders and Lexcelera, France
113. Gregor Thurmair, Linguatec, Germany
114. Rudy Tirry, Lionbridge, Belgium
115. Hans Uszkoreit, DFKI and Universität des Saarlandes, Germany
116. Erik van der Goot, Joint Research Center, EC, Italy
117. Peggy van der Kreeft, Deutsche Welle, Germany
118. Jaap van der Meer, TAUS, The Netherlands
119. René van Erk, Wolters Kluwer, The Netherlands
120. Josef van Genabith, Dublin City University and CNGL, Ireland
121. Arjan van Hessen, Twente University and Telecats, The Netherlands
122. David van Leeuwen, TNO and Radboud University, The Netherlands
123. Andrejs Vasiljevs, Tilde, Latvia
124. Michel Vérel, VecSys, France
125. Claire Waast, EDF, France
126. Philippe Wacker, EMF, UK/Belgium
127. Wolfgang Wahlster, DFKI, Germany
128. Alex Waibel, KIT, Germany and CMU, Jibbigo, USA
129. Jakub Zavrel, Textkernel, The Netherlands
130. Elie Znaty, VecSys, France
131. Chenqing Zong, Chinese Academy of Sciences, China

ABOUT META-NET

META-NET is a Network of Excellence partially funded by the European Commission [36]. The network currently consists of 60 members in 34 European countries. META-NET forges the Multilingual Europe Technology Alliance (**META**), a growing community of currently more than 600 language technology professionals and organisations. META-NET fosters the technological foundations for a multilingual European information society that: 1. makes communication and cooperation possible across languages; 2. grants all Europeans equal access to information and knowledge regardless of their language; 3. builds upon and advances functionalities of networked information technology.

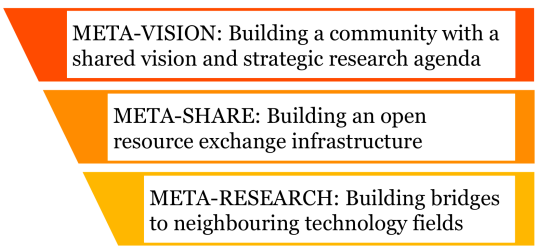
The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots.

Launched on 1 February 2010, META-NET is conducting various activities in its three lines of action META-VISION, META-SHARE and META-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. White Papers were produced for 30 languages, each one describing the status of one language with respect to its state in the digital era and existing technological support [14]. The shared technology vision was developed in three sectorial Vision Groups.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and services that are documented with metadata and organised in standardised categories. The resources can be accessed and uniformly searched. The available resources include free, open-source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. The action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.



office@meta-net.eu – http://www.meta-net.eu

MEMBERS OF META-NET

Austria	Zentrum für Translationswissenschaft, Universität Wien: Gerhard Budin
Belgium	Centre for Processing Speech and Images, University of Leuven: Dirk van Compernolle
	Computational Linguistics and Psycholinguistics Research Centre, University of Antwerp: Walter Daelemans
Bulgaria	Institute for Bulgarian Language, Bulgarian Academy of Sciences: Svetla Koeva
Croatia	Institute of Linguistics, Faculty of Humanities and Social Science, University of Zagreb: Marko Tadić
Cyprus	Language Centre, School of Humanities: Jack Burston
Czech Republic	Institute of Formal and Applied Linguistics, Charles University in Prague: Jan Hajič
Denmark	Centre for Language Technology, University of Copenhagen: Bolette Sandford Pedersen, Bente Maegaard
Estonia	Institute of Computer Science, University of Tartu: Tiit Roosmaa, Kadri Vider
Finland	Computational Cognitive Systems Research Group, Aalto University: Timo Honkela
	Department of Modern Languages, University of Helsinki: Kimmo Koskenniemi, Krister Lindén
France	Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur and Institute for Multilingual and Multimedia Information: Joseph Mariani
	Evaluations and Language Resources Distribution Agency: Khalid Choukri
	Laboratory of Computer Science, University of Le Mans: Holger Schwenk
	Laboratoire Informatique d'Avignon, University of Avignon: Georges Linares
Germany	Language Technology Lab, DFKI: Hans Uszkoreit, Georg Rehm
	Human Language Technology and Pattern Recognition, RWTH Aachen University: Hermann Ney
	Department of Computational Linguistics, Saarland University: Manfred Pinkal
	Institute for Natural Language Processing, University of Stuttgart: Jonas Kuhn, Hinrich Schütze
	Interactive Systems Lab, Karlsruhe Institute of Technology: Alex Waibel
Greece	R.C. "Athena", Institute for Language and Speech Processing: Stelios Piperidis
Hungary	Research Institute for Linguistics, Hungarian Academy of Sciences: Tamás Váradi
	Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics: Géza Németh, Gábor Olaszy

Iceland	School of Humanities, University of Iceland: Eiríkur Rögnvaldsson
Ireland	School of Computing, Dublin City University: Josef van Genabith
Italy	Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli": Nicoletta Calzolari
	Human Language Technology Research Unit, Fondazione Bruno Kessler: Bernardo Magnini
Latvia	Tilde: Andrejs Vasiļjevs
	Institute of Mathematics and Computer Science, University of Latvia: Inguna Skadiņa
Lithuania	Institute of the Lithuanian Language: Jolanta Zabarskaitė
Luxembourg	Arax Ltd.: Vartkes Goetcherian
Malta	Department Intelligent Computer Systems, University of Malta: Mike Rosner
Netherlands	Utrecht Institute of Linguistics, Utrecht University: Jan Odijk
	Computational Linguistics, University of Groningen: Gertjan van Noord
Norway	Department of Linguistic, Literary and Aesthetic Studies, University of Bergen: Koenraad De Smedt
	Department of Informatics, Language Technology Group, University of Oslo: Stephan Oepen
Poland	Institute of Computer Science, Polish Academy of Sciences: Adam Przepiórkowski, Maciej Ogrodniczuk
	University of Łódź: Barbara Lewandowska-Tomaszczyk, Piotr Pęzik
	Dept. of Comp. Linguistics and Artificial Intelligence, Adam Mickiewicz University: Zygmunt Vetulani
Portugal	University of Lisbon: António Branco, Amália Mendes
	Spoken Language Systems Laboratory, Institute for Systems Engineering and Computers: Isabel Trancoso
Romania	Faculty of Computer Science, University Alexandru Ioan Cuza of Iaşi: Dan Cristea
	Research Institute for Artificial Intelligence, Romanian Academy of Sciences: Dan Tufiş
Serbia	University of Belgrade, Faculty of Mathematics: Duško Vitas, Cvetana Krstev, Ivan Obradović
	Pupin Institute: Sanja Vranes
Slovakia	Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences: Radovan Garabík
Slovenia	Jožef Stefan Institute: Marko Grobelnik
Spain	Barcelona Media: Toni Badia, Maite Melero
	Aholab Signal Processing Laboratory, University of the Basque Country: Inma Hernaez Rioja
	Center for Language and Speech Technologies and Applications, Universitat Politècnica de Catalunya: Asunción Moreno
	Department of Signal Processing and Communications, University of Vigo: Carmen García Mateo
	Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: Núria Bel
Sweden	Department of Swedish, University of Gothenburg: Lars Borin

Switzerland	Idiap Research Institute: Hervé Bourlard
UK	School of Computer Science, University of Manchester: Sophia Ananiadou
	Institute for Language, Cognition and Computation, Center for Speech Technology Research, University of Edinburgh: Steve Renals
	Research Institute of Informatics and Language Processing, University of Wolverhampton: Ruslan Mitkov
	Department of Computer Science, University of Sheffield: Rob Gaizauskas

About 100 language technology experts – representatives of the countries and languages represented in META-NET – discussed and finalised the key results and messages of the White Paper Series at a META-NET meeting in Berlin, Germany, on October 21/22, 2011.

E

MILESTONES AND HISTORY OF THE STRATEGIC RESEARCH AGENDA

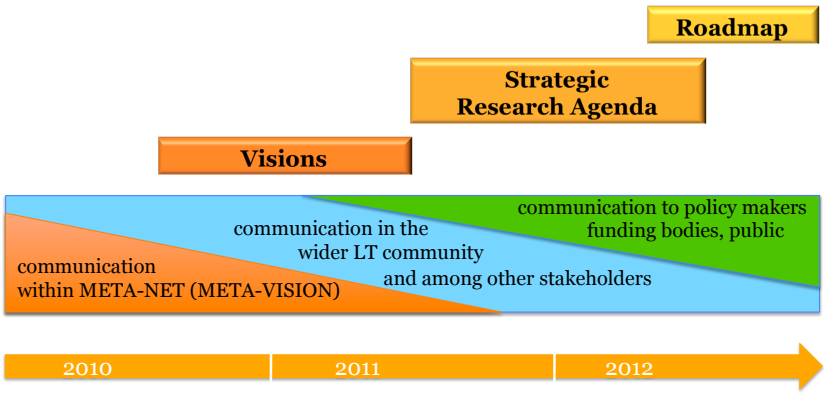
The META-VISION process within META-NET started in early 2010, its main aim was to produce this Strategic Research Agenda and the accompanying Roadmap document. Hundreds of representatives from academia, industry, official institutions, policy makers, politicians and journalists have contributed to this process. In this section we give an overview of the meetings at which the SRA or important components on the way towards the SRA have been presented and discussed (key meetings of the META-VISION process marked in bold typeface).

Important milestones within the long and complex process towards this Strategic Research Agenda include five documents: the three Vision Reports prepared by the three domain-specific Vision Groups (see below), a general Vision Paper, "The Future European Multilingual Information Society: Current State of the Discussion" [37], and the Priority Themes Paper in which the three technology visions are specified in a more concrete way, "LT 2020 – Vision and Priority Themes for Language Technology Research in Europe until the Year 2020. Towards the META-NET Strategic Research Agenda" [38]. All reports, papers and discussions that took place in the process have been reflected in the Strategic Research Agenda. The documents are available online at http://www.meta-net.eu/vision.

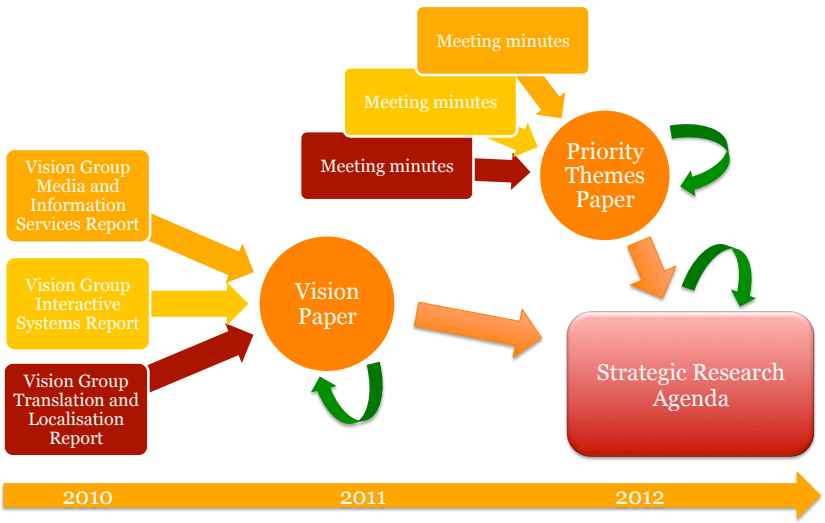
1. FLaReNet Forum, Barcelona, Spain, Feb. 11/12, 2010
2. Language Technology Days 2010, Luxembourg, March 22/23, 2010
3. EAMT 2010, Saint-Raphael, France, May 27/28, 2010
4. theMETAnk, Berlin, Germany, June 4/5, 2010
5. Translingual Europe 2010, Berlin, Germany, June 7, 2010
6. Localization World, Berlin, Germany, June 8/9, 2010
7. Multisaund Seminar, Istanbul, Turkey, June 16-18, 2010
8. **Vision Group "Text Translation and Localisation"** (1st meeting), Berlin, Germany, July 23, 2010
9. **Vision Group "Media and Information Services"** (1st meeting), Paris, France, Sep. 10, 2010
10. **Vision Group "Interactive Systems"** (1st meeting), Paris, France, Sep. 10, 2010
11. ICT 2010, Brussels, Belgium, September 27-29, 2010
12. **Vision Group "Text Translation and Localisation"** (2nd meeting), Brussels, Belgium, Sep. 29, 2010
13. **Vision Group "Interactive Systems"** (2nd meeting), Prague, Czech Republic, Oct. 5, 2010
14. Languages and the Media 2010, Berlin, Germany, October 7, 2010
15. HLT: The Baltic Perspective, Riga, Latvia, October 7/8, 2010
16. LISA Forum Europe, Budapest, Hungary, October 13, 2010
17. **Vision Group "Media and Information Services"** (2nd meeting), Barcelona, Spain, Oct. 15, 2010
18. EFNIL 2010, Thessaloniki, Greece, Nov. 3, 2010
19. Interact Presidential Summit, Moffett Field, USA, Nov. 8-9, 2010

20. **META Technology Council** (1st meeting), Brussels, Belgium, Nov. 16, 2010
21. Language question in research: English vs. national languages?, Finnish Parliament, Helsinki, Nov. 17, 2010
22. **META-FORUM 2010: "Challenges for Multilingual Europe"**, Brussels, Belgium, Nov. 17/18, 2010
23. Oriental-Cocosda 2010, Kathmandu, Nepal, Nov. 24-25, 2010
24. The International Workshop on Spoken Language Translation (IWSLT), Paris, France, Dec. 2/3, 2010
25. Meeting of the LT Berlin working group, Berlin, Germany, Dec. 9, 2010
26. Language Technology for Multilingual Applications, European Parliament, Luxembourg, Jan. 27, 2011
27. Opening of German/Austrian W3C Office at DFKI Berlin, Berlin, Germany, Feb. 10, 2011
28. Japanese Workshop for Machine Translation, Tokyo, Japan, Feb. 23, 2011
29. Meeting of Representatives of European Language Councils, Copenhagen, Denmark, March 08, 2011
30. TRALOGY, Paris, France, March 3/4, 2011
31. **Vision Group "Interactive Systems"** (3rd meeting), Rotterdam, The Netherlands, March 28, 2011
32. **Vision Group "Media and Information Services"** (3rd meeting), Vienna, Austria, April 1, 2011
33. Meeting of the LT Berlin working group, Berlin, Germany, April 4, 2011
34. W3C Workshop "Content on the multilingual web", Pisa, Italy, April 5, 2011
35. **Vision Group "Translation and Localisation"** (3rd meeting), Prague, Czech Republic, April 7/8, 2011
36. Attensity Forum 2011, Berlin, Germany, May 6, 2011
37. **META Technology Council** (2nd meeting), Venice, Italy, May 25, 2011
38. FLaReNet Forum, Venice, May 26-27, 2011
39. Multisaund Seminar, Bursa, Turkey, June 13-14, 2011
40. META-NET Workshop at ICANN 2011: Context in Machine Translation, Espoo, Finland, June 14, 2011
41. Speech Processing Conference, Tel Aviv, Israel, June 21-22, 2011
42. **META-FORUM 2011: "Solutions for Multilingual Europe"**, Budapest, Hungary, June 27/28, 2011
43. Media for All, London, June 29-July 1, 2011
44. EUROLAN 2011 Summer School, Cluj-Napoca, Romania, Aug. 28-Sep. 4, 2011
45. Interspeech 2011, Firenze, Italy, Aug. 28-31, 2011
46. RANLP 2011, Hissar, Bulgaria, Sep. 12-14, 2011
47. Multilingual Web Workshop, Limerick, Ireland, Sep. 21/22, 2011
48. ML4HMT Workshop at MT Summit, Xiamen, China, Sep. 19-23, 2011
49. Workshop Language Technology for a Multilingual Europe at GSCL 2011, Hamburg, Germany, Sep. 27, 2011
50. GSCL 2011: "Multilingual Resources and Multilingual Applications", Hamburg, Germany, Sep. 28-30, 2011
51. **META Technology Council** (3rd meeting), Berlin, Germany, Sep. 30, 2011
52. Workshop on IPR and Metadata by META-NORD, Helsinki, Finland, Sep. 30, 2011
53. META-NET Network Meeting and General Assembly, Berlin, Germany, Oct. 21/22, 2011
54. NPLD Assembly, Eskilstuna, Sweden, Oct. 25/26, 2011
55. EFNIL 2011, London, UK, Oct. 26, 2011
56. Oriental-Cocosda 2011, Hsinchu, Taiwan, Oct. 26-28, 2011
57. SIMC 2011 International Symposium on Multilingualism in the Cyberspace, Brasilia, Brasil, Nov. 7-9, 2011
58. IJCNLP 2011, Chiang Mai, Thailand, Nov. 9-13, 2011
59. ML4HMT-11 Workshop, Barcelona, Spain, Nov. 19, 2011
60. LTC 2011, Poznan, Poland, Nov. 25-27, 2011

61. GALA Conference, Monaco, March 26-29, 2012
62. EACL 2012, Avignon, France, April 23-27, 2012
63. CESAR Roadshow Event, Sofia, Bulgaria, May 2, 2012
64. LREC 2012, Istanbul, Turkey, May 21-27, 2012
65. CESAR Roadshow Event, Bratislava, Slovakia, June 7/8, 2012
66. Multilingual Web Workshop, Dublin, Ireland, June 11, 2012
67. **META Technology Council** (4th meeting), Brussels, Belgium, June 19, 2012
68. **META-FORUM 2012: "A Strategy for Multilingual Europe"**, Brussels, Belgium, June 20/21, 2012
69. CHAT 2012 Workshop, Madrid, Spain, June 22, 2012



The three phases of the META-VISION process



Steps towards the Strategic Research Agenda for Multilingual Europe 2020

F

ABBREVIATIONS AND ACRONYMS

AI Artificial Intelligence

API Application Programming Interface

CALL Computer-Assisted Language Learning

CAT Computer-Aided Translation

CEF Connecting Europe Facility

CMS Content Management System

EFNIL European Federation of National Institutions for Language

ETP European Technology Platform

GALA Globalization and Localization Association

GPS Global Positioning System

HQMT High-Quality Machine Translation

HLT Human Language Technology

HTML Hypertext Markup Language

IaaS Infrastructures as a Service

IR Information Retrieval

ISO International Organization for Standardization

ICT Information and Communication Technology

IT Information Technology

JTI Joint Technology Initiative

LR Language Resource

LSP Language Service Provider

LT Language Technology

META Multilingual Europe Technology Alliance

MT Machine Translation

NLP Natural Language Processing

NPLD Network to Promote Linguistic Diversity

PaaS Platforms as a Service

PHP PHP: Hypertext Preprocessor

PPP Public-Private Partnership

RSS RDF Site Summary; Really Simple Syndication

SME Small and Medium Enterprises

SaaS Software as a Service

SRA Strategic Research Agenda

TFEU Treaty of the Functioning of the European Union

TM Translation Memory

TMS Translation Management System

WWW World Wide Web

W3C World Wide Web Consortium

In everyday communication, Europe's citizens, business partners and politicians are inevitably confronted with language barriers. Language technology has the potential to overcome these barriers and to provide innovative interfaces to technologies and knowledge. This document presents a Strategic Research Agenda for Multilingual Europe 2020. The paper was prepared by META-NET, a Network of Excellence funded by the European Commission. META-NET consists of 60 research centres in 34 countries, who cooperate with stakeholders from economy, government agencies, research organisations, non-governmental organisations, language communities and European universities. META-NET's vision is high-quality language technology for all European languages.

"The research carried out in the area of language technology is of utmost importance for the consolidation of Portuguese as a language of global communication in the information society."
– Dr. Pedro Passos Coelho (Prime-Minister of Portugal)

"It is imperative that language technologies for Slovene are developed systematically if we want Slovene to flourish also in the future digital world."
– Dr. Danilo Türk (President of the Republic of Slovenia)

"For such small languages like Latvian keeping up with the ever increasing pace of time and technological development is crucial. The only way to ensure future existence of our language is to provide its users with equal opportunities as the users of larger languages enjoy. Therefore being on the forefront of modern technologies is our opportunity."
– Valdis Dombrovskis (Prime Minister of Latvia)

"Europe's inherent multilingualism and our scientific expertise are the perfect prerequisites for significantly advancing the challenge that language technology poses. META-NET opens up new opportunities for the development of ubiquitous multilingual technologies."
– Prof. Dr. Annette Schavan (German Minister of Education and Research)