Chapter 3

The META-NET strategic research agenda for language technology in Europe: An extended summary

Georg Rehm
DFKI GmbH

Recognising Europe's exceptional demand and opportunities for multilingual language technologies, 60 leading research centres in 34 European countries joined forces in META-NET, a European Network of Excellence. META-NET has developed a Strategic Research Agenda (SRA) for multilingual Europe – the complex planning and discussion process took more than two years to complete. While the complete SRA has been published elsewhere (Rehm & Uszkoreit 2013), this heavily condensed version provides an extended summary as an alternative mode of access and to enable interested parties to familiarise themselves with its key concepts in an efficient way.

1 Introduction

The multilingual setup of our European society imposes grand societal challenges on political, economic and social integration and inclusion, especially in the creation of the single digital market and unified information space targeted by the Digital Agenda (European Commission 2010). As many as 21 European languages are at risk of digital extinction (Rehm & Uszkoreit 2012). They could become victims of the digital age as they are under-represented online and under-resourced with respect to language technologies. Huge market opportunities remain untapped because of language barriers. If no action is taken, many European citizens will find that speaking their mother tongue leaves them at a social and economic disadvantage.

Georg Rehm. The META-NET strategic research agenda for language technology in Europe: An extended summary. In Georg Rehm, Felix Sasaki, Daniel Stein & Andreas Witt (eds.), *Language technologies for a multilingual Europe: TC3 III*, 19–41. Berlin: Language Science Press. DOI:10.5281/zenodo.1291926

Language technology is the missing piece of the puzzle that will bring us closer to a single digital market. It is the key enabler and solution to boosting future growth in Europe and strengthening our competitiveness. The key question is: Will Europe wholeheartedly decide to participate in this fast growing market?

Although we use computers to write, phones to chat and the web to search for knowledge, Information Technology (IT) does not yet have access to the meaning, purpose and sentiment behind our trillions of written and spoken words. Technology will bridge the rift separating IT and the human mind using sophisticated technologies for language understanding. Today's computers cannot understand texts and questions well enough to provide translations, summaries or reliable answers, but in less than ten years such services will be offered for many languages. Technological mastery of human language will enable a host of innovative IT products and services in commerce, administration, government, education, health care, entertainment, tourism and other sectors.

Recognising Europe's exceptional demand and opportunities, 60 leading research centres in 34 European countries joined forces in META-NET (http://www.meta-net.eu), a European Network of Excellence dedicated to the technological foundations of a multilingual, inclusive, innovative and reflective European society and partially supported through several projects funded by the European Commission (EC). META-NET assembled the Multilingual Europe Technology Alliance (META) with more than 700 organisations and experts representing multiple stakeholders. In addition, META-NET signed collaboration agreements and memoranda of understanding (see META-NET 2013) with more than 40 other projects and initiatives in the field such as CLARIN (Common Language Resources and Technology Infrastructure, http://www.clarin.eu) and FLaReNet (Fostering Language Resources Network, http://www.flarenet.eu).

Working together with numerous organisations and experts from a variety of fields, META-NET has developed a Strategic Research Agenda (SRA, Rehm & Uszkoreit 2013). Our recommendations for Multilingual Europe 2020, as specified in the SRA, are based on a thorough planning process involving more than one thousand experts.

We predict, in line with many other forecasts, that the next generation of IT will be able to handle human language, knowledge and emotion in competent and meaningful ways. These new competencies will enable an endless stream of novel services that will improve communication and understanding. Many services will help people learn about and understand things such as world history, technology, nature and the economy. Others will help us to better understand each other across language and knowledge boundaries. They will also drive many other services including programmes for commerce, localisation, and personal assistance.

Our ultimate goal is monolingual, crosslingual and multilingual technology support for all languages spoken by a significant population in Europe. To achieve this, we recommend focusing on three priority research topics connected to innovative application scenarios that will provide European research and development (R&D) in this field with the ability to compete with other markets and subsequently achieve benefits for European society and citizens as well as an array of opportunities for our economy and future growth. We are confident that upcoming EU funding programmes, specifically Horizon 2020 (European Commission 2012b) and Connecting Europe Facility (European Commission 2011a), combined with national and regional funding, can provide the necessary resources for accomplishing our joint vision.

A recent policy brief (Veugelers 2012) proposes that Europe specialises in new ICT (Information and Communications Technology) sectors as a means for post-crisis recovery. The European problem lies less in the generation of new ideas than in their successful commercialisation. The study identifies major obstacles: the lack of a single digital market, and the absence of ICT clusters and powerful platform providers. It suggests that the EU policy framework could overcome these barriers and leverage the growth potential of new ICT markets by extending research and infrastructure funding to pre-commercial projects, in particular those involving the creation of ICT clusters and platforms. This is exactly the goal we are trying to achieve. Our recommendations envisage five lines of action for large-scale research and innovation. First, there are three priority research themes: *Translingual Cloud*, *Social Intelligence and e-Participation* and *Socially Aware Interactive Assistants*. The other two themes focus upon *Core Technologies and Resources for Europe's Languages* and a *European Service platform for Language technologies.*

The objective of the priority research themes is to turn our joint vision into reality and allow Europe to benefit from a technological revolution that will overcome barriers of understanding between people of different languages, people and technology, and people and the digitised knowledge of mankind.

2 Multilingual Europe: Facts and opportunities

During the last 60 years, Europe has become a distinct political and economic structure. Culturally and linguistically it is rich and diverse. However, everyday communication between Europe's citizens, enterprises and politicians is inevitably confronted with language barriers. They are an invisible and increasingly problematic threat to economic growth (Economist 2012). The EU's institutions spend about *one billion Euros per year* on translation and interpretation to maintain their policy of multilingualism (European Commission 2012c) and the overall European market for translation, interpretation, software localisation and website globalisation was estimated at 5.7 billion Euros in 2008.

The only – unacceptable and rather un-European – alternative to a multilingual Europe 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. Given the 23 official EU languages plus 60 or more other languages spoken in Europe (European Commission 2012a), language learning alone cannot solve the problem. Without technological support, our linguistic diversity will be an insurmountable obstacle for the entire continent. Only about half of the 500 million people who live in the EU speak English. There is no such thing as a lingua franca shared by the vast majority of the population.

Less than 10% of the EU's population are willing or able to use online services in English, which is why multilingual technologies are badly needed to support and to move the EU online market from more than 20 language-specific sub-markets to one unified single digital market with more than 500 million users and consumers. The current situation with "many fragmented markets" is considered one of the main obstacles that seriously undermine Europe's efforts to exploit ICT fully (European Commission 2010).

Language technology is a key enabler for sustainable, cost-effective and socially beneficial solutions to overcome language barriers. It will offer European stakeholders tremendous advantages, not only within the European market, but also in trade relations with non-European countries, especially emerging economies.

In the late 1970s the EU realised the 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 (Joscelyne & Lockwood 2003; Lazzari 2006), the European Commission set up a department dedicated to language technology and machine translation a few years ago. Selective funding efforts have led to a number of valuable results. For example, the EC's translation services now use Moses, which has been mainly developed in European research projects. However, these 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.

Europe now has a well-developed research base. Through initiatives such as CLARIN and META-NET the community is well connected and engaged in a long term agenda that aims gradually to strengthen language technology's role. What is missing in Europe is awareness, political determination and political will that would take us to a leading position in this technology area through a concerted funding effort. This major dedicated push needs to include the political determination to modify and to adopt a shared, EU-wide language policy that foresees an important role for language technologies.

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 (European Commission 2008; 2012a). 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 (UNESCO 2007; 2008; 2011b; Vannini & Crosnier 2012).

Many Europeans find it difficult to interact with online services and participate in the digital economy. According to a recent study, only 57% of internet users in Europe purchase goods and services in languages that are not their native language. Fifty-five percent of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the web (European Commission 2011c). A few years ago, English might have been the lingua franca of the web 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 (Ford & Batson 2011). Already today, more than 55% of web-based content is not in English.

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, while 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 ca. 16.5 billion Euros in 2015 (European Commission 2009b; 2011b). Yet, this existing capacity is not enough to satisfy current and future needs, e. g., with regard to translation (DePalma & Kelly 2009). Already today, Google Translate translates the same volume per day that all human translators on the planet translate in one year (Och 2012).

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 will be impossible. The same applies to information services, document services, media industries, digital archives and language teaching. The most compelling solution for ensuring the breadth and depth of language usage in tomorrow's Europe is to use appropriate technology. Still, the quality and usability of current technologies is far from what is needed. Especially the smaller European languages suffer severely from under-representation in the digital realm.

Drawing on the insights gained so far, today's hybrid language technology mixing deep processing with statistical methods could 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. The three priority research themes are mainly aimed at Horizon 2020 (European Commission 2012b). The more infrastructural aspects, platform design and implementation and concrete language technology services are aimed at CEF (European Commission 2011a). An integral component of our strategic plans are the member states and associated countries: it is of utmost importance to set up, under the umbrella of the SRA, a coordinated initiative both on the national (member states, regions, associated countries) and international level (EC/EU), including research centres as well as small, medium and large enterprises who work on or with language technologies.

3 How can language technology help?

We believe that *Language Technology made in Europe for Europe* will significantly contribute to future European cross-border and cross-language communication, economic growth and social stability while establishing for Europe a worldwide, leading position in technology innovation, securing Europe's future as a worldwide trader and exporter of goods, services and information. There are many societal changes and challenges as well as economic and technological trends that confirm the urgent need to include sophisticated language technology in our European ICT infrastructure. Among these changes and challenges are language barriers (European Commission 2009a), an ageing population, people with disabilities, immigration and integration, personal information services and customer care, operation and cooperation on a global scale, preservation of cultural heritage, linguistic diversity (WSIS 2003; UNESCO 2011a), social media and e-participation as well as market awareness and customer acceptance.

Multilingualism has become the global norm rather than the exception (Vannini & Crosnier 2012). Future 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 and can 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.

4 Language technology 2012: Current state

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 the White Paper Series "Europe's Languages in the Digital Age" (Rehm & Uszkoreit 2012) that describes the current state of language technology support for 30 European languages (including all 23 official EU languages). This immense undertaking has been in preparation since mid 2010 and was published in the Summer of 2012. More than 200 experts participated to the 30 volumes as co-authors and contributors.

The differences in technology support between the various languages and areas are dramatic and alarming. In all of the four areas we examined (machine translation, speech processing, text analytics, language resources), 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 analytics and essential language resources. Others have basic resources but the implementation of semantic methods is still far away. 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.

The META-NET White Paper Series is fully available online at http://www.meta-net.eu/whitepapers. On this website we also present the press release “At least 21 European Languages in Danger of Digital Extinction” which was circulated on the occasion of the European Day of Languages 2012 (Sept. 26), and also its impact around the world. The echo generated by our press release shows that Europe is very passionate and concerned about its languages and that it is also very interested in the idea of establishing a solid language technology base for overcoming language barriers.

5 Language technology 2020: The META-NET technology vision

We believe that in the next IT revolution computers will master our languages. Just as they already understand measurements 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 they are today and will further enhance our work and life.

The broad area of COMMUNICATION AMONG PEOPLE will see a dramatically increased use of sophisticated language technology (LT). By the year 2020, with sufficient research effort on high-quality automatic translation and robust accurate speech recognition, reliable dialogue translation for face-to-face conversation and telecommunication will be possible for at least hundreds of languages, across multiple subject fields and text types, both spoken and written. Authoring software will check for appropriate style according to genre and purpose and help evaluate comprehensibility. It will flag potential errors, suggest corrections, and use authoring memories to suggest completions of started sentences or even whole paragraphs. By 2020 tele-meetings will be the norm for professional meetings. LT will be able to record, transcribe, and summarise them. Brainstorming will be facilitated by semantic lookup and structured display of relevant data, proposals, pictures, and maps. Business email will be embedded in semantic process models to automate standardised communication. Even before 2020, email communication will be semantically analysed, checked for sentiment indicators, and summarised in reports. Semantic integration into work processes, threading, and response management will be applied across channels, as will machine translation and analytics.

Human language will become the primary medium for COMMUNICATION BETWEEN PEOPLE AND TECHNOLOGY. The voice-control interfaces we see today for smartphones and search engines are just the modest start of overcoming the communication barrier between humankind and the non-human part of the world. Only a few years ago the idea of talking to a car to access key functions would have seemed absurd, yet it is now commonplace. Recently the concept of a personal digital assistant has increased in popularity. We will soon see much more sophisticated virtual personalities with expressive voices, faces, and gestures. They will become an interface to any information provided online. 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. By the year 2020 we could have a highly personalised, socially aware and interactive virtual assistant. Having been trained on the user's behaviour and communication space, it will proactively offer advice and it will be able to speak in the language and dialect of the user but also digest information in other natural and artificial languages and formats. The assistant will translate or interpret without the user even needing to request it. By 2020 there will be a competitive landscape of intelligent interfaces to all kinds of objects and services employing human language and other modes for effective communication.

In the context of the Semantic Web, Linked Open Data and the general semantification of the web as well as knowledge acquisition and ontology population, LT can perform many tasks in the PROCESSING OF KNOWLEDGE AND INFORMATION. It can sort, categorise, catalogue, and filter content and it can deliver the data for data mining in texts. LT can connect web documents with meaningful hyperlinks and it can produce summaries of larger text collections. Opinion mining and sentiment analysis can find out what people think about products, personalities, or problems and analyse their feelings about such topics. 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 predicted and planned use of language and knowledge technologies for social intelligence applications will involve text and speech analytics, translation, summarisation, opinion mining, sentiment analysis, and several other technologies. In 2020, LT will enable forms of knowledge evolution, transmission and 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.

The wide range of novel or improved applications in our shared vision represents only 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, and 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.

6 Language technology 2020: The META-NET priority research themes

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. 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:

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 dialects, sociolects, registers, jargons, genres and slangs.

Depth and meaning. Understanding language is a complex process. Human language is not only the key to knowledge and thought, it also cannot be interpreted without shared knowledge and active inference. Computational language proficiency needs semantic technologies.

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.

It is fortunate for research and economy that the only way to effectively tackle the three challenges involves submitting the evolving technology continuously to the growing demands and practical stress tests of real world applications. Only a continuous stream of technological innovation can provide the economic pull forces and the evolutionary environments for the realisation of the grand vision.

We propose five major action lines of research and innovation:

- Three Priority Research Themes along with application scenarios to drive research and innovation. These will demonstrate novel technologies in show-case solutions with high economic and societal impact. They will open up numerous new business opportunities for European language-technology and -service providers.
  1. **Translingual Cloud:** generic and specialised federated 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 and consumers to enable e-participation and more effective processes for preparing, selecting and evaluating collective decisions.
  3. **Socially Aware Interactive 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.
- The other two themes focus upon base technologies and a service platform:
  4. **Core technologies and resources for Europe's languages:** a steadily evolving system of shared, collectively maintained interoperable core technologies and resources for the languages of Europe and selected other languages. These will ensure that our languages will be sufficiently supported and represented in the next generations of IT.
  5. **A European service platform for language technologies** for supporting research and innovation by testing and showcasing research results, integrating various services, even including professional human services, will allow small to medium enterprise (SME) providers to offer component and end-user services, and share and utilise tools, components and data resources.

These priority themes have been designed with the aim of turning our 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 knowledge of mankind. The themes connect societal needs with LT applications.

6.1 Priority theme 1: Translingual cloud

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 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.

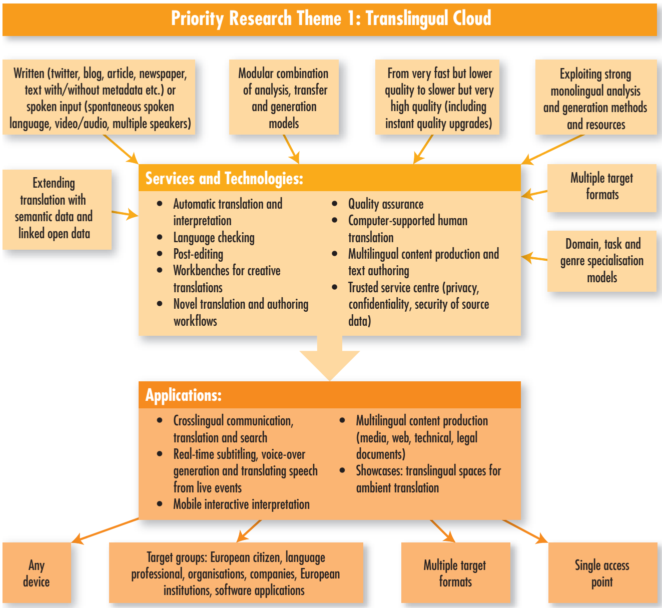
Priority Research Theme 1: Translingual Cloud
Written (twitter, blog, article, newspaper, text with/without metadata etc.) or spoken input (spontaneous spoken language, video/audio, multiple speakers)
Modular combination of analysis, transfer and generation models
From very fast but lower quality to slower but very high quality (including instant quality upgrades)
Exploiting strong monolingual analysis and generation methods and resources
Extending translation with semantic data and linked open data
Services and Technologies:
Automatic translation and interpretation
Language checking
Post-editing
Workbenches for creative translations
Novel translation and authoring workflows
Quality assurance
Computer-supported human translation
Multilingual content production and text authoring
Trusted service centre (privacy, confidentiality, security of source data)
Multiple target formats
Domain, task and genre specialisation models
Applications:
Crosslingual communication, translation and search
Real-time subtitling, voice-over generation and translating speech from live events
Mobile interactive interpretation
Multilingual content production (media, web, technical, legal documents)
Showcases: translingual spaces for ambient translation
Any device
Target groups: European citizen, language professional, organisations, companies, European institutions, software applications
Multiple target formats
Single access point

Figure 1: Priority Research Theme 1: Translingual Cloud

One key component of this service (see Figure 1) is a use and provision platform for providers of computer-supported top-quality human translation, multilingual text authoring and quality assurance by experts. Other important components are trusted service centres as certified service providers fulfilling highest standards for privacy, confidentiality and security of source data and translations and quality upscale models embedded into services permitting instant quality upgrades if the results of the requested service levels do not yet fulfil the quality requirements.

6.2 Priority theme 2: Social Intelligence and e-Participation

The central goal behind the second theme is to use information technology and the digital content of the web for improving effectiveness and efficiency of decision-making in business and society (see Figure 2). Social intelligence builds on improved text analytics methodologies but goes far beyond the analysis. One goal is the analysis of large volumes of social media, comments, blogs, forum postings etc. of citizens, customers, 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. As 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.

A key enabler will be technologies that can map large, heterogeneous, and, to a large extent, unstructured volumes of online content to actionable representations that support decision making and analytics tasks. Such mappings can range from the relatively shallow to the relatively deep, encompassing 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 or the resolution of co-reference

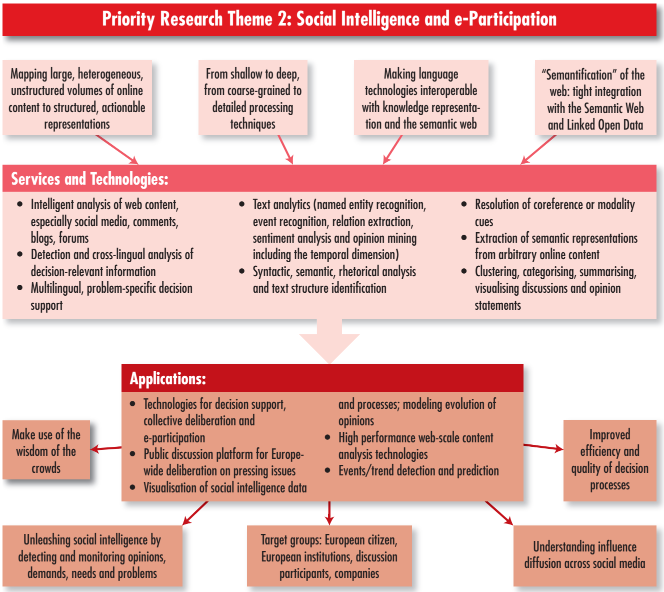
Priority Research Theme 2: Social Intelligence and e-Participation
Mapping large, heterogeneous, unstructured volumes of online content to structured, actionable representations
From shallow to deep, from coarse-grained to detailed processing techniques
Making language technologies interoperable with knowledge representation and the semantic web
"Semantification" of the web: tight integration with the Semantic Web and Linked Open Data
Services and Technologies:
Intelligent analysis of web content, especially social media, comments, blogs, forums
Detection and cross-lingual analysis of decision-relevant information
Multilingual, problem-specific decision support
Text analytics (named entity recognition, event recognition, relation extraction, sentiment analysis and opinion mining including the temporal dimension)
Syntactic, semantic, rhetorical analysis and text structure identification
Resolution of coreference or modality cues
Extraction of semantic representations from arbitrary online content
Clustering, categorising, summarising, visualising discussions and opinion statements
Applications:
Technologies for decision support, collective deliberation and e-participation
Public discussion platform for Europe-wide deliberation on pressing issues
Visualisation of social intelligence data
and processes; modeling evolution of opinions
High performance web-scale content analysis technologies
Events/trend detection and prediction
Make use of the wisdom of the crowds
Improved efficiency and quality of decision processes
Unleashing social intelligence by detecting and monitoring opinions, demands, needs and problems
Target groups: European citizen, European institutions, discussion participants, companies
Understanding influence diffusion across social media

Figure 2: Priority Research Theme 2: Social Intelligence and e-Participation

or modality cues within and across sentences. Technologies such as, e. g., information extraction, data mining, automatic linking and summarisation have to be made interoperable with knowledge representation and semantic web methods. Drawing expertise from related areas such as knowledge management, information sciences, or social sciences is a prerequisite to meet the challenge of modelling social intelligence (Généreux & Hamon 2012).

6.3 Priority theme 3: Socially aware interactive assistants

Socially aware interactive assistants are conversational agents (see Figure 3). Their socially-aware behaviour is a result of combining analysis methods for speech, non-verbal and semantic signals. They support people interacting with their environment, including human-computer, human-agent/robot, and computer-mediated human-human interaction. The assistants must be able to act in various environments, both indoor, outdoor and virtual environments, 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. It can assist in language training and education, and provide synthetic multimedia information analytics.

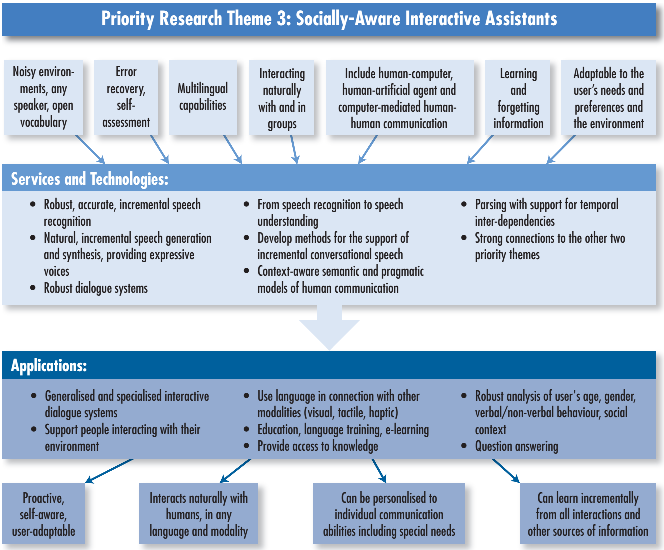
Priority Research Theme 3: Socially-Aware Interactive Assistants
Noisy environments, any speaker, open vocabulary
Error recovery, self-assessment
Multilingual capabilities
Interacting naturally with and in groups
Include human-computer, human-artificial agent and computer-mediated human-human communication
Learning and forgetting information
Adaptable to the user's needs and preferences and the environment
Services and Technologies:
• Robust, accurate, incremental speech recognition
• Natural, incremental speech generation and synthesis, providing expressive voices
• Robust dialogue systems
• From speech recognition to speech understanding
• Develop methods for the support of incremental conversational speech
• Context-aware semantic and pragmatic models of human communication
• Parsing with support for temporal inter-dependencies
• Strong connections to the other two priority themes
Applications:
• Generalised and specialised interactive dialogue systems
• Support people interacting with their environment
• Use language in connection with other modalities (visual, tactile, haptic)
• Education, language training, e-learning
• Provide access to knowledge
• Robust analysis of user's age, gender, verbal/non-verbal behaviour, social context
• Question answering
Proactive, self-aware, user-adaptable
Interacts naturally with humans, in any language and modality
Can be personalised to individual communication abilities including special needs
Can learn incrementally from all interactions and other sources of information

Figure 3: Priority Research Theme 3: Socially Aware Interactive Assistants

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 and robustness 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.

6.4 Theme 4: 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 5). To this group belong tools and technologies such as, among others, tokenisers, part-of-speech taggers, parsers, tools for building language models, information retrieval tools, machine learning toolkits, speech recognition and speech synthesis engines, and integrated architectures. 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, terminologies, dictionaries, ontologies and language models. Both tools and resources can be rather general or highly task- or domain-specific, tools can be language-independent, while datasets are, by definition, language-specific. There are also several types of resources, such as corpora for machine translation or spoken dialogue corpora specific to one or more of the priority themes.

A key component of this 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 their feature sets, performance, accuracy, etc. through dynamic push-pull effects. Conceptualising these technologies as a set of shared core technologies will have positive effects on their sustainability and interoperability. Also, many European languages other than English are heavily under-resourced (Rehm & Uszkoreit 2012).

The European academic and industrial technology community is fully aware of the need for sharing resources such as language data, language descriptions, tools and core technology components as a basis for the successful development and implementation of the priority themes. Initiatives such as FLaReNet (Calzolari et al. 2011) 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 (see http://www.meta-share.eu). All language resources and basic technologies will be created under the core technologies umbrella.

6.5 Theme 5: A European service platform for language technologies

We recommend the design and implementation of an ambitious large-scale platform as a central motor for research and innovation in the next phase of IT evolution and as a ubiquitous resource for the multilingual European society. The platform will be used for testing, showcasing, proof-of-concept demonstration, avant-garde adoption, experimental and operational service composition, and fast and economical service delivery to enterprises and end-users (see Figure 4). The creation of a cloud platform 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. Large-scale ICT infrastructures and innovation clusters such as this one are foreseen in the Digital Agenda for Europe (European Commission 2010: 24).

A top layer consists of LANGUAGE PROCESSING such as text filters, tokenisation, spell, grammar and style checking, hyphenation, lemmatising and parsing. At a 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 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. 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.

All three priority areas will be able to contribute to and at the same time draw immense 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 solutions of the three areas.

6.6 Languages to be supported

The SRA has a much broader scope in terms of languages to be supported than our study "Europe's Languages in the Digital Age" (Rehm & Uszkoreit 2012). The set of languages to be reflected with technologies include not only the 23 official

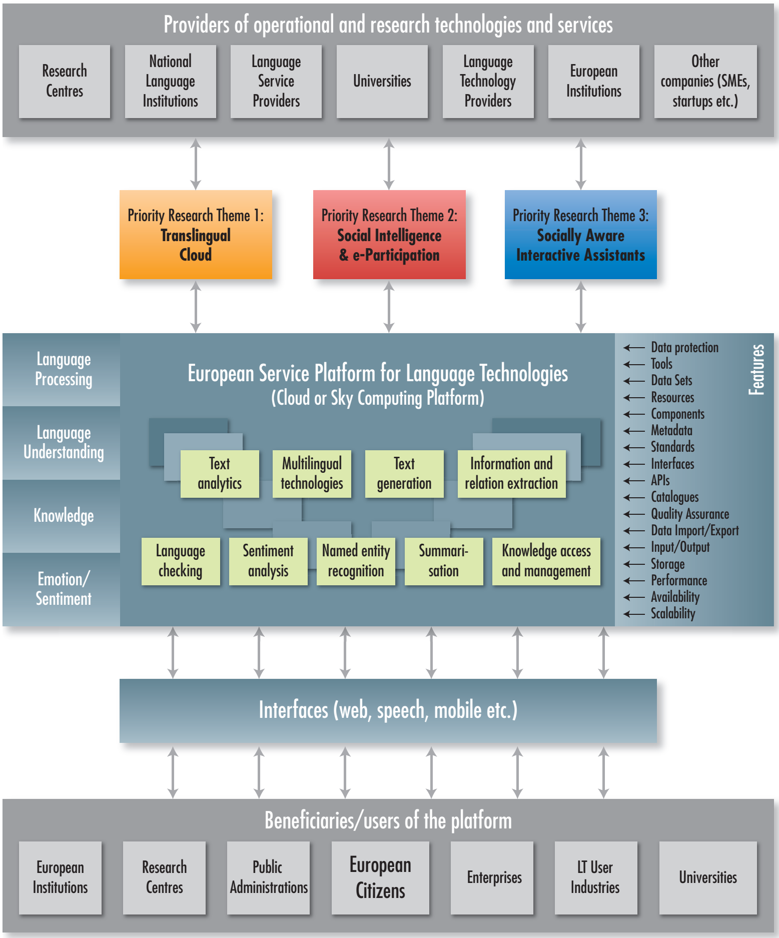
Providers of operational and research technologies and services
Research Centres
National Language Institutions
Language Service Providers
Universities
Language Technology Providers
European Institutions
Other companies (SMEs, startups etc.)
Priority Research Theme 1: Translingual Cloud
Priority Research Theme 2: Social Intelligence & e-Participation
Priority Research Theme 3: Socially Aware Interactive Assistants
Language Processing
Language Understanding
Knowledge
Emotion/ Sentiment
European Service Platform for Language Technologies
(Cloud or Sky Computing Platform)
Text analytics
Multilingual technologies
Text generation
Information and relation extraction
Language checking
Sentiment analysis
Named entity recognition
Summari-sation
Knowledge access and management
Features
Data protection
Tools
Data Sets
Resources
Components
Metadata
Standards
Interfaces
APIs
Catalogues
Quality Assurance
Data Import/Export
Input/Output
Storage
Performance
Availability
Scalability
Interfaces (web, speech, mobile etc.)
Beneficiaries/users of the platform
European Institutions
Research Centres
Public Administrations
European Citizens
Enterprises
LT User Industries
Universities

Figure 4: European Service Platform for Language Technologies

languages of the EU but also recognised and unrecognised regional languages and the languages of associated countries or non-member states. Equally important are the minority and immigrant languages that are in active use by a significant population in Europe (for Germany, these are, among others, Turkish and Russian; for the UK, these include Bengali, Urdu/Hindi and Punjabi). An important set of languages outside our continent are those of important political and trade partners such as Chinese, Japanese, Korean, Russian, and Thai. META-NET already has good working relationships with several of the respective official bodies, especially EFNIL (European Federation of National Institutions for Language), NPLD

(Network to Promote Linguistic Diversity), and also the Maaya World Network for Linguistic Diversity. The concrete composition of languages to be supported by this agenda's research programme up until the year 2020 and beyond depends on the composition of participating countries and regions and also on the specific nature of the funding instruments used and combined for realising the ambitous plan.

6.7 Structure and principles of research organisation

The three proposed priority research themes overlap in technologies and challenges. 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 this highly ambitious programme. The 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 Translingual Cloud will focus on cross-lingual technologies such as translation and interpretation; the Social Intelligence strand will take care of knowledge discovery, text analytics and related technologies; the research dedicated to Interactive Assistants will take on technologies such as speech and multimodal interfaces (see Figure 5).

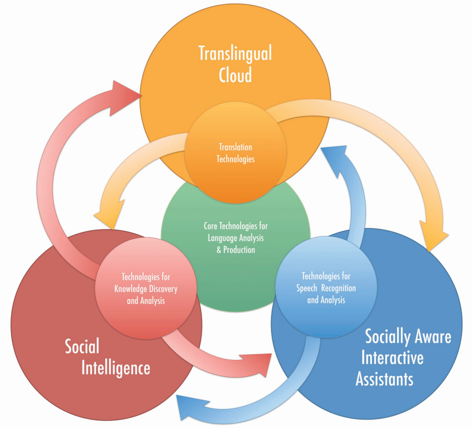
Translingual Cloud
Translation Technologies
Core Technologies for Language Analysis & Production
Technologies for Knowledge Discovery and Analysis
Technologies for Speech Recognition and Analysis
Social Intelligence
Socially Aware Interactive Assistants

Figure 5: Scientific cooperation among the three priority research themes

The final model for the organisation of collaboration will have to be guided by a thoughtful combination of the following basic approaches. The collaboration will be interdisciplinary, flexible, evolutionary and analytical. It will be staged in two major phases (2015–2017, 2018–2020). We will make heavy use of improving systems by bootstrapping earlier systems and prototypes and by collaborating very closely with relevant areas of service and technology industries.

7 Towards a shared European programme

In the Strategic Research Agenda META-NET recommends setting up a large, multi-year programme on language technologies to build the technological foundations for a truly multilingual Europe. The research strands and associated sets of applications we suggest to build in the next ten years are of utmost importance for Europe. Through these technologies we will be able to overcome language barriers in spoken and written communication, we will be able to carry out country- and language-border-crossing debates and we will enable new forms and means of communication. We are confident that the impact of our technologies will be so immense that they will be able to help establishing a sense of a European identity in the majority of European citizens. The research plan described in the SRA will generate a countless number of opportunities, it will significantly participate to Europe's future growth and will secure Europe's position in many global markets.

Due to the scope and duration of the suggested action, our preferred 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 (Bulgaria, Czech Republic, France, Hungary, The Netherlands, Slovenia) who participated in a panel discussion on this topic unanimously expressed the urgent need for setting up such a shared programme (META-NET 2012a).

The programme will include a carefully planned governance structure. Here, first steps have been taken as well: META-NET has an Executive Board with currently 12 members, the operations of the network and its bodies are specified in its Statutes (META-NET 2012b). Furthermore, a legal person for META-NET was established. This legal person, META-TRUST AISBL, is an international non-profit organisation under Belgian law (META-NET 2012c). These proven and established structures can be used as starting points for the governance structure of a future programme.

Acknowledgements

This extended summary of the *META-NET Strategic Research Agenda for Multilingual Europe 2020* (Rehm & Uszkoreit 2013) is, first and foremost, based on joint work with the SRA's other co-editor, Hans Uszkoreit. The work presented in this article would not have been possible without the dedication and commitment of our colleagues Aljoscha Burchardt, Kathrin Eichler, Tina Klüwer, Arle Lommel and Felix Sasaki (all DFKI), the 60 member organisations of the META-NET network of excellence, the ca. 70 members of the Vision Groups, the ca. 30 members of the META Technology Council, the more than 200 authors of and contributors to the META-NET Language White Paper Series (Rehm & Uszkoreit 2012) and the ca. 200 representatives from industry and research who contributed to the META-NET Strategic Research Agenda. The author would also like to thank the reviewers of this paper for their helpful comments.

META-NET is co-funded by the 7th Framework Programme of the European Commission through the following grant agreements: T4ME Net (no. 249 119), CESAR (no. 271 022), METANET4U (no. 270 893) and META-NORD (no. 270 899).

More information is available at http://www.meta-net.eu and via office@meta-net.eu.

References

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

DePalma, Donald A. & Nataly Kelly. 2009. *The Business Case for Machine Translation. How Organizations Justify and Adopt Automated Translation.* Common Sense Advisory. http://www.commonsenseadvisory.com.

Economist, Economist Intelligence Unit: The. 2012. *Competing across borders. How cultural and communication barriers affect business.* http://www.managementthinking.eiu.com/competing-across-borders.html.

European Commission. 2008. *Multilingualism: An Asset for Europe and a Shared Commitment.* http://ec.europa.eu/languages/pdf/comm2008_en.pdf.

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

European Commission. 2009b. *Size of the language industry in the EU.* http://ec.europa.eu/dgs/translation/publications/studies.

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

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

European Commission. 2011b. *Languages mean business*. http://ec.europa.eu/languages/languages-mean-business/.

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

European Commission. 2012a. *Europeans and their Languages*. European Commission. Special Eurobarometer 386/77.1. http://ec.europa.eu/languages/languages-of-europe/eurobarometer-survey_en.htm.

European Commission. 2012b. *Horizon 2020: The Framework Programme for Research and Innovation*. http://ec.europa.eu/research/horizon2020/.

European Commission. 2012c. *Languages*. http://ec.europa.eu/languages/.

Ford, Daniel & Josh Batson. 2011. *Languages of the world (wide web)*. http://googleresearch.blogspot.com/2011/07/languages-of-world-wide-web.html.

Généreux, Michel & Thierry Hamon (eds.). 2012. *Workshop on Language Technology for Decision Support at the Fourth Swedish Language Technology Conference*. http://permalink.gmane.org/gmane.science.linguistics.corpora/15911.

Joscelyne, Andrew & Rose Lockwood. 2003. *The EUROMAP Study. Benchmarking HLT progress in Europe*. http://cst.dk/dandokcenter/FINAL_Euromap_rapport.pdf.

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

META-NET. 2012a. *META-FORUM 2012: A Strategy for Multilingual Europe. Panel discussion "Plans for LT Research and Innovation in Member States and Regions"*. Videos available at http://www.meta-net.eu/events/meta-forum-2012/programme.

META-NET. 2012b. *META-NET Statutes. Version 1.1.* http://www.meta-net.eu/META-NET-Statutes.pdf.

META-NET. 2012c. *META-TRUST AISBL (Association internationale sans but lucratif)*. http://www.meta-trust.eu.

META-NET. 2013. *META-NET: Collaborations with other projects and initiatives*. http://www.meta-net.eu/collaborations.

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

Rehm, Georg & Hans Uszkoreit (eds.). 2012. *META-NET White Paper Series: Europe's Languages in the Digital Age.* Heidelberg / New York / Dordrecht / London: Springer. 31 volumes on 30 European languages. More details and full text available at http://www.meta-net.eu/whitepapers.

Rehm, Georg & Hans Uszkoreit (eds.). 2013. *The META-NET Strategic Research Agenda for Multilingual Europe.* Heidelberg / New York / Dordrecht / London: Springer. Presented by the META Technology Council. More details and full text available at http://www.meta-net.eu/sra and http://link.springer.com/book/10.1007/978-3-642-36349-8.

UNESCO. 2007. *Intersectoral Mid-term Strategy on Languages and Multilingualism.* Paris. http://unesdoc.unesco.org/images/0015/001503/150335e.pdf.

UNESCO. 2008. *UNESCO Information for All Programme: International Conference Linguistic and Cultural Diversity in Cyberspace: Final Document. Lena Resolution.*

UNESCO. 2011a. *Information for All Programme (AFP).* http://www.unesco.org/new/en/communication-and-information/intergovernmental-programmes/information-for-all-programme-ifap/.

UNESCO. 2011b. *UNESCO Information for All Programme: Second International Conference Linguistic and Cultural Diversity in Cyberspace: Final Document. Yakutsk Call for Action.* http://www.maayajo.org/IMG/pdf/Call_for_action_Yakutsk_EN-2.pdf.

Vannini, Laurent & Hervé Le Crosnier (eds.). 2012. *Net. Lang: Towards the Multilingual Cyberspace.* Paris: C&F éditions. The Maaya World Network for Linguistic Diversity. http://net-lang.net.

Veugelers, Reinhilde. 2012. *New ICT Sectors: Platforms for European Growth?* Issue 2012/14. http://www.bruegel.org/publications/.

WSIS. 2003. *World Summit on the Information Society: Declaration of Principles – Building the Information Society: A global challenge in the new Millennium.* http://www.itu.int/wsis/docs/geneva/official/dop.html.