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. They are an invisible and increasingly problematic threat to economic growth as several recent studies have shown [8].

The EU's institutions spend about *one billion Euros per year* on translation and interpretation to maintain their policy of multilingualism [9] and the overall European market for translation, interpretation, software localisation and website globalisation was estimated at 5.7 billion Euros in 2008. Are these expenses necessary? Are they even sufficient? Despite this high level of expenditure, only a fraction of the information is translated that is available to the whole population in countries with a single predominant language, such as the USA or China.

Language technology and linguistic research, as well as related fields such as the digital humanities, social sciences and psychology, can significantly contribute to overcoming linguistic barriers. Combined with intelligent devices and applications, a European language technology platform will help European citizens to talk and do business together even if they do not speak a mutual language.

The economy benefits from the European single market. 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 global level. The only (unacceptable) alternative to a multilingual Europe [10] 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, an area in which language technologies can play a key role.

Given the 23 official EU languages plus 60 or more other languages spoken in Europe [11], language learning on its own cannot solve the problem of cross-border communication or commerce [8]. Without technological support such as machine translation, our 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.

Less than 10% of the EU's population are willing or able to use online services in English which is why multilingual services based on language technologies are badly needed to support and to move the EU online market from more than 20 language-specific sub-markets to a unified single digital market with more than 500 million users and consumers. The main goal, 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 [8], robust and high-quality cross- and multilingual language technologies need to be developed urgently. In fact, the current situation with "many fragmented markets" is considered one of the main obstacles that seriously undermine Europe's efforts to exploit ICT fully [5]! A truly functioning single digital market can only be established once

G. Rehm and H. Uszkoreit (eds.), *META-NET Strategic Research Agenda for Multilingual Europe 2020*,
White Paper Series, DOI: 10.1007/978-3-642-36349-8_1, © The Author(s) 2013

the language barrier has fallen, something that can be achieved only through research, development and wide deployment of language technologies (see Figure 1). The single digital market functions poorly because multilingual Europe itself functions poorly.

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. One prerequisite to develop these solutions was 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" [12] this important step has now been taken (see also Chapter 4, p. 27 ff., and Appendix C, p. 80 for an overview of the timeline and history of this document).

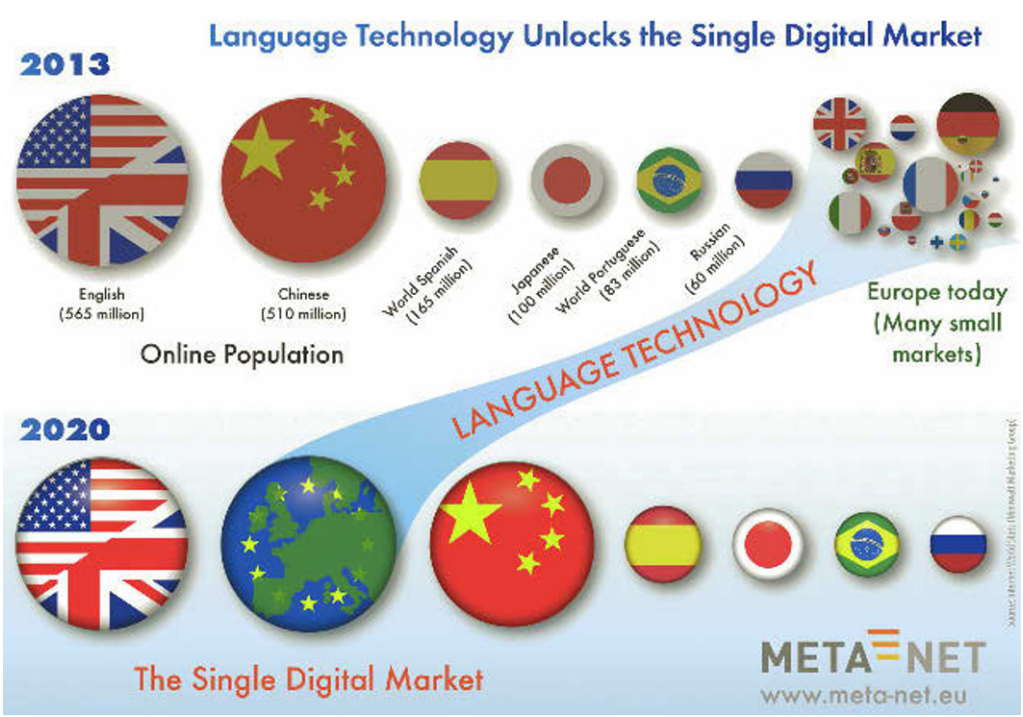
There are two main axes around which language technologies are needed and able to bring about the next IT revolution: *communication* and *data analysis*. Communication includes support for activities such as talking, conversing, carrying out dialogues and debates (both spoken and written), authoring and further processing (summarising, categorising etc.) of texts ranging from instant messages to complex documents, and also translation. Data analysis includes organising, structuring and understanding data, extracting information and relations between entities. The term data here refers to arbitrary types of unstructured data as well as any type of text. In the medium-to-long term we want to realise technologies for socially-aware and context-aware natural language understanding and generation, including translation.

In 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 [13, 14], the European Commission set up a department dedicated to language technology and machine translation a few years ago; in an internal reorganisation this department was recently integrated into a new unit called "Data Value Chain", part of Directorate G, "Media & Data", in the EC Directorate General for "Communications Networks, Content and Technology" (DG Connect). In the past ca. five years, the EU has been supporting 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 EC's translation services 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 2 depicts the languages that have been studied by Language Technology researchers in 2010, taking into account major conferences and journals. It illustrates how research has focussed primarily on English followed by Chinese, German, French, and a few other bigger languages. Many European languages were not studied at all, e. g., Slovak, Maltese, Lithuanian, Irish, Albanian, Croatian, Macedonian, Montenegrin, Romansh, Galician, Occitan, or Frisian.

Research activities have tended to be isolated and while they have delivered valuable results, they have had difficulty making a decisive impact on the market. In many cases research funded in Europe eventually bore fruit outside Europe; enterprises such as Google and Apple have been noteworthy beneficiaries. In fact, many of the predominant actors in the field today are based in the US.

Europe now has a well-developed research base. Through initiatives such as CLARIN and META-NET the re-

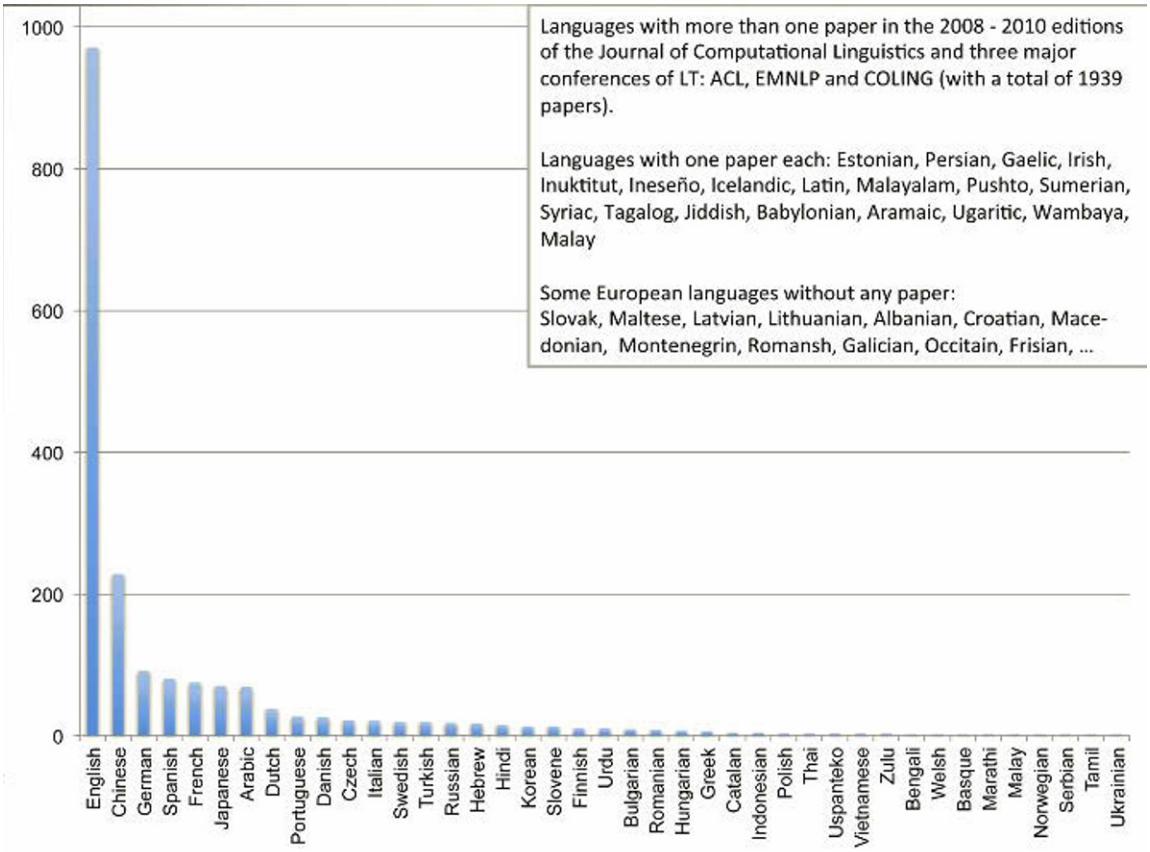


1: Language Technology unlocks the Single Digital Market

search community is well connected and engaged in a long term agenda that aims gradually to strengthen language technology's role. At the same time, our position is worse when compared to other multilingual societies. Despite having fewer 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 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.

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. 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 emerging business opportunities into competitive advantages. To this end, we have developed this Strategic Research Agenda (see Appendix C, p. 80).

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. The two core chapters of this document summarise our shared vision of the role of language technol-



2: Languages treated in research published in the 2008–2010 edition of the Journal of Computational Linguistics and the conferences of ACL, EMNLP and COLING (internal, unpublished study)

ogy in the year 2020 in non-technical terms (Chapter 5, p. 32 ff.) and outline three priority themes for large-scale research and innovation (Chapter 6, p. 41 ff.):

1. **Translingual 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 our 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 themes build the bridge between societal needs, applications, and roadmaps for the organisation of research, development and scientific innovation. They cover the main functions of language: storing, sharing and using information and knowledge, as well as improving social interaction among humans and enabling social interaction between humans and technology.

We also present 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 sketched strategy are novel modes of large-scale col-

lective research and interaction among the major stakeholder constituencies including 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 our strategy. Of central importance is a rapid 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) [15] and Horizon 2020 [16].

The three priority research themes are mainly aimed at Horizon 2020 (2014–2020). The more infrastructural aspects, platform design and implementation and concrete language technology services are aimed at CEF. 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 are to include public services, 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 level (EC/EU), including research centres as well as small, medium and large enterprises who work on or with language technologies. Only through an agreement and update of our national and international language policy frameworks, close cooperation between all stakeholders, and tightly coordinated collaboration can we realise the ambituous plan of researching, designing, developing and putting into practice a European platform [17] that supports all citizens of Europe, and beyond, by providing, among others, sophisticated services for communication across language barriers.