EAFT–ELETO Symposium

"National Languages and Terminology in Higher Education, Science & Technology"

Athens, Greece, 7 November 2013

ICT Enabling Language Diversity

Joseph Mariani
LIMSI-CNRS
&
Director
Institute for Multilingual and Multimedia Information (IMMI)
BP133, 91403 Orsay Cedex (France)
Joseph.Mariani@limsi.fr

Summary

The issues of multilingualism

The issues of multilingualism are twofold:

- First, to take care of preserving cultures and languages, i.e. to allow citizens to express themselves in their first language. This question takes on a particular depth in the context of the construction of Europe, given the strong linguistic diversity within a single political entity. A study conducted for the European Commission shows that 90% of European citizens questioned prefer to find websites in their native language rather than in a foreign language. One can also note that currently it is estimated that less than 30% of the web is in English, a proportion that has declined sharply from a rough estimate of 50% in 2000. 50% of European citizens speak only one language and when they speak a second one, it is not necessarily English. Only 3% of Japanese speak a foreign language. In India, less than 5% of people fluently speak English. Preserving languages and, through them, their corresponding culture responds to a strong demand from citizens.

- The second challenge is to enable communication among humans, usually in the framework of common democratic structures. We are facing it in the European Union, where, with the recent expansion, there are now 28 member countries and 24 official languages, representing 552 language pairs. If one considers all the European languages, one can count more than 100, which represents more than 10,000 pairs of languages to translate! The European Commission employs more than 2,500 translators who translate about two million pages per year. This covers only a fraction of the needs. To cover the totality would require 8,500 translators to process 6.8 million pages annually. Taking into account the EU linguistic diversity represents 30% of the budget of the European Parliament, or about 300 million euros per year, with the use of 500 translators and interpreters. The estimated total cost of multilingualism for the European Union is a little over one billion euros per year; but considering the number of Europeans, that represents only 2.2 euros per citizen per year, which ultimately is not prohibitive. The same study conducted for the EC showed that, on the economic side, only 33% of the EU citizens would buy goods over the internet in a foreign language while, on the cultural side, 80% of the EU citizens think that websites in their language should be translated to foreign languages. A similar situation exists within some nations, like India, but also internationally, with more than 6,000 major languages that are spoken, or 36 million pairs of languages to translate. And a simple statistic: at present YouTube, every minute, uploads 100 hours of new videos in all languages!

Needs related to multilingualism

At the European level, the needs related to multilingualism are very numerous: needs for the establishment of the European Digital Library (Europeana, which included, in 2013, 23 million documents in 26 languages), for which it is necessary to provide crosslingual and multilingual tools to enable access for all; for the realization of a multilingual platform for alert and information exchange planned by the European Security

Agency (ENISA) for the Member States; for the European Patent Office – The London Protocol has reduced the number of official languages to three (English, German and French) for reasons of cost, whereas, with more automated tools, more languages could be handled; for meetings of the European Commission, of the European Parliament or of the European Court of Justice, where English tends increasingly to become the only working language. In 1997, 45% of the source documents to be translated at the EC were in English and 40% in French, while in 2007, 72% of the documents were in English and 12% in French!

Such needs respond to a real democratic necessity, to be met more generally at the international level. Dubbing and subtitling of audiovisual works; writing technical manuals, in the aerospace or automotive industries, or instruction manuals for the consumers; live super-titling of works of performing art; translation of texts, videos, and radio or television programs that are innumerable, and in all languages; simultaneous interpreting in military or sanitary operations, which take place throughout the world, and at multiple meetings, conferences, or workshops; interpretation of courses, with the coming of the Massive Open Online Courses (MOOC). Think also of the situation related to scientific articles written in a mother tongue, which are diminished markedly due to the overvaluation of English by bibliometrics, risking the loss of specialized terminology for innovation and science in other languages. The ratio of papers written in English listed in the Web of Science went from 85% to 96% in the period 1980-2000, while te ratio of papers written in French went down from 4% to 1%! The ratio of cited papers in English was already 97% in 1990!

Add to this picture the many needs related to the accessibility of information by the visually or hearing impaired, requiring the translation of information from one medium to another (written to oral, oral to written, oral to gesture (sign language)), and more generally to the accessibility of information by people who do not speak fluently the language in which it was encoded, including, notably, migrants.

Findings

The extent of these needs shows very well that they cannot all be covered by existing or even future human resources of professions dealing with language processing.

Taking into account multilingualism is not a top priority in any economic sector. If we ask the CEO of a big company what is his/her priority, none will say it is multilingualism. But if we add up the priorities in each area where it is necessary to take it into account, then we reach a very large sum. This therefore requires, in our opinion, thought and political action to bring out this awareness and provide appropriate responses.

Even when multilingualism is seen as a necessity, its cost is too important. It is this gap that calls for the development of Language Technologies (LT), both for written and spoken language, and their utilization when their performance is up to the needs of target applications.

Languages that lack LT, for example in car GPS, Smartphone interaction, Internet search, or Emergency access, will be less and less used, while languages that benefit from crosslingual LT, such as Machine Translation or Speech Translation, will get more used.

It should be noted that currently, language technologies have not yet reached maturity for all languages, with strong imbalances among languages. For example, automated translation is not good enough to translate literary works or, in general, texts which require high quality translation. This must be said clearly. But on the other hand, it can help a human translator in his or her work and has a sufficient quality to give an approximate translation, of web pages for example, thus meeting the needs of the general public.

Language Technologies

Language technologies are said to be *monolingual* when they handle a single language, *multilingual* when the same technology processes several (individual) languages, or *crosslingual* when they allow for switching and transferring from one language to another.

Language technologies cover the processing of written language, whether monolingual (morphosyntactic and syntactic analysis; text understanding; text generation; automatic summarization; terminology extraction; information retrieval; systems that respond to questions, etc.) or crosslingual (automatic or computer-aided translation; crosslingual information retrieval, etc.).

For the processing of spoken language, there are also monolingual technologies (speech recognition and understanding; speech-to-text transcription (textual transcription of what has been said); speech synthesis; spoken dialogue; speaker recognition, etc.…) and crosslingual (identification of a spoken language, speech

translation, real-time interpretation, etc.).

Finally, it includes the processing of signed languages (recognition, synthesis and translation).

These technologies can be intermedia, ie translating from one medium to another, with numerous applications to enable accessibility for the disabled (Text-To-Speech synthesis for the visually impaired, automatic transcription (subtitles or supertitles), aids to lip reading, Sign Language processing for the hearing impaired, voice commands for the motor-impaired).

Numerous resulting applications are now in everyday use, such as, regarding written language processing, spelling and grammar checkers, monolingual and crosslingual search engines, online machine translation…, and, regarding spoken language processing, talking GPS systems, dictation systems, transcription and automatic indexing of audiovisual content... This list shows that many of these existing applications are related to linking spoken and written language (transcription of speech into text, speech synthesis from text). Spoken dialogue systems, including voice recognition and synthesis, are also growing, but in very specific applications: Voice command on mobile phones, Call centers, tourist or public transportation information, etc.

Language Resources and Evaluation

It is crucial for conducting research aiming at developing language technologies to provide a base that includes both language resources and evaluation methods for the technologies that are developed.

With regard to language resources, the data (corpus, lexicons, dictionaries, terminology databases, etc.) are both necessary for conducting research investigations in linguistics and for training automatic language processing systems that are based in most cases on statistical methods. The greater the amount of data, the better the statistical model and therefore the better the system performances. The interoperability of language resources also invites us to think more deeply on the standards to be put in place in order to organize, browse, and transmit data.

It is also necessary to have a means for evaluating these technologies in order to compare the performance of systems, using a common protocol with common test data, in the context of evaluation campaigns. This allows for comparing different approaches and having an indicator of the quality of the research and of the advances of technology, and of the adequacy of a technology to respond to the needs of an application. We now speak of "coopetition"- a mix of international competition and cooperation - and this has become a way to carry on technological research.

The digital divide and language coverage

There is currently a two-speed situation and a "digital divide" between languages for which technologies exist, and others. This is related to the "weight of languages". It should be noted that 95% of languages are spoken by only 6% of world population. Some linguists believe that 90% of languages will have disappeared within a century. We can therefore classify languages according to the data and automatic processing systems that exist for these languages: whether they are well, less or not at all "resourced", or indeed if they have only an oral tradition and no writing system at all. Only 1-2% of languages presently benefit from language technologies. The availability of data is crucial for the development of usable systems, often based on statistical approaches.

In order to resolve this digital divide, how can we take into account "minority" languages, regional languages, languages spoken by migrants, foreign or regional accents? Who bears the cost when these languages are of no economic or political interest, or are unrelated to armed conflicts or natural disasters that justify addressing them? How to ensure that citizens in a community of states are able to communicate among themselves? How to reduce the risk of conflicts and crises by allowing exchanges between people? This is now a major social and political issue, which is the subject of much debate.

Research efforts in the domain

To produce the language resources and technologies that are needed to address multilingualism, different initiatives can be identified:

- Those of big companies. It must be underlined that many large U.S. companies in the information technology sector, such as like Google, Microsoft, Apple, IBM, Amazon, eBay or Facebook, make a major effort in multilingualism and crosslingualism, some of them taking advantage of their access to huge amount

of data for developing better systems. The Google search engines work in 145 languages (national and regional), and Google has made available "free" tools for machine translation and crosslingual information retrieval online. Google Translate handles 72 languages and 5,112 language pairs on the Internet and on smartphones, including 17 languages with voice input, and 26 languages with voice output and several varieties of English, Chinese and Spanish. It also now allows for camera input. It is claimed to have 200 million users and to translate 1 million documents per day: more than what translators do worldwide in a year. And Google targets 100 languages, i.e. 10,000 language pairs, by 2015. The Google Book Search Library contains 30 million documents in 46 languages and in December 2010 Google provided statistics on the evolution of human language from a corpus of 500 billion words (including 361 billion words in English and 45 billion words in French and Spanish). Apple Siri provides spoken interaction with iPhones in 8 languages and 19 language varieties. Also Microsoft provides the MS Word spelling checker in 126 languages (233 if we consider regional variants) and the grammar checker in 6 languages (61 if we consider regional variants).

- National programs in some countries, with different objectives: to process an internal multilingualism (TDIL (Technology Development for Indian Languages) in India, developing Language Technologies for the 22 Indian official languages, NHN (National Human Language Technology Network) in South Africa, addressing the 11 official languages; to understand foreign languages for geopolitical reasons (GALE or EARS in the United States, funded by the Department of Defense (DARPA)); to ensure the use and promotion of a national or transnational language (TechnoLangue for French, STEVIN for Dutch/Flemish); or to maintain a place in an economic and cultural competition (Quaero in France).

- Efforts to support R & D programs of the European Commission. Research is state-of-the-art in Europe as it appears in the results of the international evaluation campaign. But development in LT is only conducted by SMEs, and there is not yet a EC program addressing LT to support Multilingualism in the EU as a political issue (only as a research issue for the time being). A study conducted by the EC T4ME project showed that 21 European languages are in danger of digital extinction. Should we abandon them? Should we let US companies take care of them? The EC probably doesn’t have enough forces to address by itself all LT for all EU official languages, and furthermore for all European languages, that would require more and better coordination with EU Member States and Regions.

Conclusion and Perspectives

Multilingualism is a fundamental dimension of the European Union, and is mandatory in order to preserve national cultures and allow for communication. But Multilingualism is very costly. Language Technologies can cut those costs and are therefore the only way to allow for Multilingualism, in Europe and worldwide.

In order to develop Language Technologies, Language Resources and Language Technology evaluation are needed, while they are presently available only for a very small set of languages. The other languages are therefore in danger of digital extinction.

It is thus proposed to launch a program aiming at developing LT covering all European languages through a joint coordinated effort between the European Commission, the Member States and the regions, ensuring the availability of Language Resources and of evaluated Language Technologies through a common platform. This model could be extended to include non-EU third parties.