META-NET. Official Languages of Spain in the Digital Age

Asunción Moreno[1], Núria Bel[2], Maite Melero[3], Carmen García-Mateo[4], Inma Hernáez[5], Sergio Oller[6], Aljoscha Burchardt[7], Kathrin Eichler[7], Georg Rehm[7], and Hans Uszkoreit[7]

[1] Universitat Politècnica de Catalunya. Centre TALP
asuncion.moreno@upc.edu,
[2] Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada
nuria.bel@upf.edu,
[3] Barcelona Media
maite.melero@barcelonamedia.org
[4] Universidade de Vigo. AtlantTIC Research Center
carmen.garcia@uvigo.es
[5] University of the Basque Country. Aholab Signal Processing Laboratory
inma.hernaez@ehu.es
[6] Universitat de Barcelona
sergioller@gmail.com
[7] DFKI GmbH, Berlin, Germany
{aljoscha.burchardt,kathrin.eichler,georg.rehm,hans.uszkoreit}@dfki.de

Abstract. In the framework of the Network of Excellence META-NET, a White Paper series has been created to chart the status of Language Technologies for 30 European languages. The exhaustive analysis reports on the state of a language including social and technological aspects and statistics about the availability of Language Resources and Tools. The White Papers are addressed to politicians, journalist and decision makers. In this paper, we show a comparative study of the findings on Language Technologies and Tools for the four official languages spoken in Spain.

Keywords: Language technologies, Basque, Catalan, Galician, Spanish

1 Introduction

META-NET [10] is a Network of Excellence funded by the European Commission through a cluster of coordinated projects: T4ME [12], CESAR [13], META-NORD [14], and METANET4U [15]. The network also has multiple unfunded members. The network currently consists of 60 members from 34 European countries [2]. META-NET forges META, the Multilingual Europe Technology Alliance, a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- grants all Europeans equal access to information and knowledge regardless of their language;
- 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 has already conducted 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 common Strategic Research Agenda (SRA) [1]. Its main focus is to build a coherent and cohesive LT community in Europe, by bringing together representatives from extremely diverse and highly fragmented groups of stakeholders. Of great importance is the White Paper series [4] produced for more than 30 languages, of which the main results for the official languages spoken in Spain are compiled in this paper.

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 Web services that are documented with high-quality metadata and organised in standardised categories [3]. The resources can be readily 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. In particular, 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 the members of the community.

As of November 2012, META-NET consists of 60 research centres from 34 European countries. META-NET is working with stakeholders from economy (software companies, technology providers and users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

In the framework of META-VISION, a White Paper series has been created to chart the status of Language Technologies for 30 European languages.

The exhaustive analysis reports on the state of each language including social and technological aspects, and statistics about the availability of Language Resources and Tools. In this paper we show a comparative study of the findings on Language Technologies and Tools for the four official languages spoken in Spain. Next section briefly describes the contents of the White Papers. Section 3 shows a detailed comparison of the language technologies, applications and solutions for Basque, Catalan, Galician and Spanish. Section 4 show the results of this analysis among the other analyzed European languages. Conclusions are included in section 6.

2 White Paper series

META-VISION has conducted an analysis of current language resources and technologies in a White Paper series. The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe [11]. More than 200 experts from all over Europe have contributed to the 31 volumes of the White Paper series and more than 8,000 printed copies were disseminated to key decision makers, politicians and journalists in September and October of 2012. Four White Papers for the four official languages spoken in Spain: Basque [7], Catalan [5] Galician [8] and Spanish [6] were produced within a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others.

Each volume has three main parts; first of all, there is a description of how the diversity of languages in Europe is a rich cultural heritage and, at the same time, a risk for the survival of the languages. The White Paper series show that Language technologies can play a key role helping multilingual societies to survive. The second part of the White Papers deals with the specificities of each language and how it is promoted at national and international level. The third part provides a general description of the Language Technologies, with a presentation of the State of the Art, and also the programmes, projects and efforts at national and international levels done to improve the Language Technologies and Resources for each particular language. This part ends with a summary of availability of tools and resources and a cross-language comparison between the intended language and the other 29 European languages included in the White Paper series.

The analysis of the White Paper series shows that the availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of additional research. In the specific case of Spain, we find one official language, Spanish, and three co-official languages: Catalan, Galician and Basque. Spanish

Fig. 1: Compilation of White Papers

is, in addition, an official language in Europe, most of the countries in Latin America and is broadly spoken in the US. Preserving multilingualism in Spain has not been an easy task. It is the result of a complex process to intentionally preserve cultural and linguistic identity within and among the various regions and peoples of Spain. Similar to the use of English in the European case, direct communication between citizens of different language areas of Spain, often need to use Spanish as a lingua franca.

3 State of Language Technology in the four official languages of Spain

A Table with a summary of the current state of language technology (LT) support for each of the languages was elaborated. For each intended language, the

data were taken from a survey among a representative number of researchers and industrial people in the LT community. Two main areas were considered: Language Technology including tools, technologies and applications; and Language Resources including resources, data and knowledge data bases. A number of representative Language Technologies were chosen for the study: Speech Recognition, Speech Synthesis, Grammatical analysis, Semantic analysis, Text generation, and Machine Translation. For the Language Resources study, the following groups were chosen: Text corpora, Speech corpora, Parallel corpora, Lexical resources and Grammars.

Experts were asked to rate in a scale from 0 (very low) to 6 (very high) the existing tools and resources for each group according to seven criteria: Quantity, Availability, Quality, Coverage, Maturity, Sustainability, and Adaptability.

Figures 2 and 3 show the mean scores obtained for each language in each Language Technology for the criteria Quality and Coverage. Figures 4 and 5 show the mean scores obtained for each language in each Language Resources group for the criteria Quality and Coverage

The key results of the White Paper series for the official languages of Spain can be summed up as follows:

- Speech processing currently seems to be slightly more mature than the processing of written text. In fact, for Spanish and Catalan, speech technology has already been successfully integrated into many everyday applications, from spoken dialogue systems and voice-based interfaces to mobile phones and car navigation systems.
- Semantics is more difficult to process than syntax; text semantics is more difficult to process than word and sentence semantics. The more semantics a tool takes into account, the more difficult is to find the right data; more efforts to support deep processing are needed.
- Research has successfully led to the design of medium to high quality software for basic text analysis, such as tools for morphological analysis and syntactic parsing. But advanced technologies that require deep linguistic processing and semantic knowledge are still in their infancy.
- As to resources, there are interesting resources in Spanish and Catalan, but they are not easily available for research. There are a number of corpora annotated with syntactic, semantic and discourse structure mark-up, but again, there are not nearly enough language corpora containing the right sort of content to meet the growing need for more deep linguistic and semantic information.
- In particular, there is a lack of the sort of parallel corpora that form the basis for statistical and hybrid approaches to machine translation. Parallel corpora exist, between Spanish and English, and between Spanish and other languages from Spain. However, parallel corpora between the four languages and other foreign languages are mostly missing.
- Many of these tools, resources and data formats do not meet industry standards and cannot be sustained effectively. A concerted programme is required to standardise data formats and APIs.

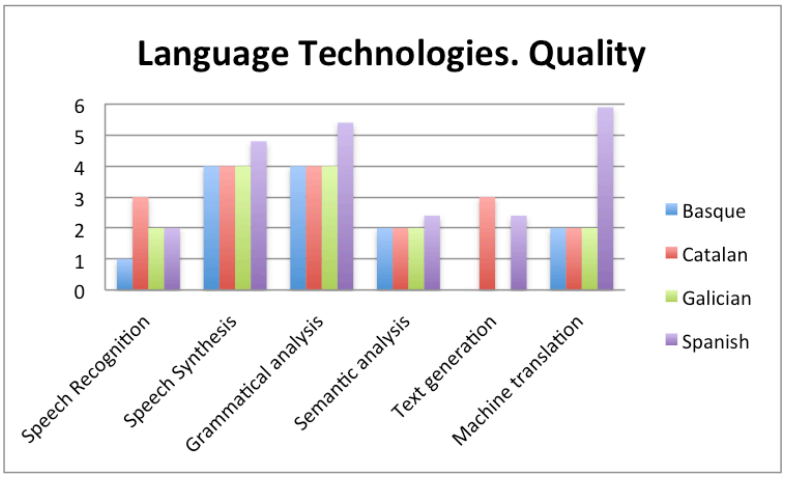

Fig. 2: Quality scores for Language Technologies

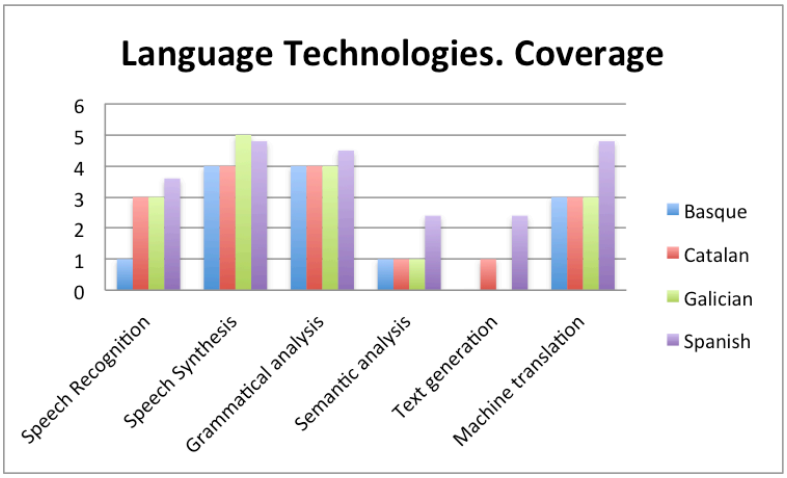

Fig. 3: Coverage scores for Language Technologies

– An unclear legal situation restricts making use of digital texts, such as those published online by newspapers, for empirical linguistic and language technology research, for example, to train statistical language models. Together with politicians and policy makers, researchers should try to establish laws or regulations that enable them to use publicly available texts for language-related R&D activities.
– The cooperation between the Language Technology community and those involved with the Semantic Web and the closely related Linked Open Data movement should be intensified with the goal of establishing a collaboratively maintained, machine-readable knowledge base that can be used both

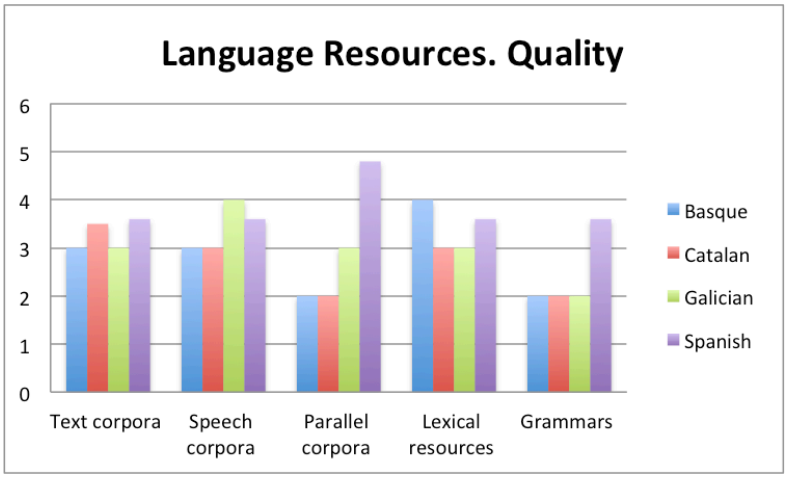

Fig. 4: Quality scores for Language Resources

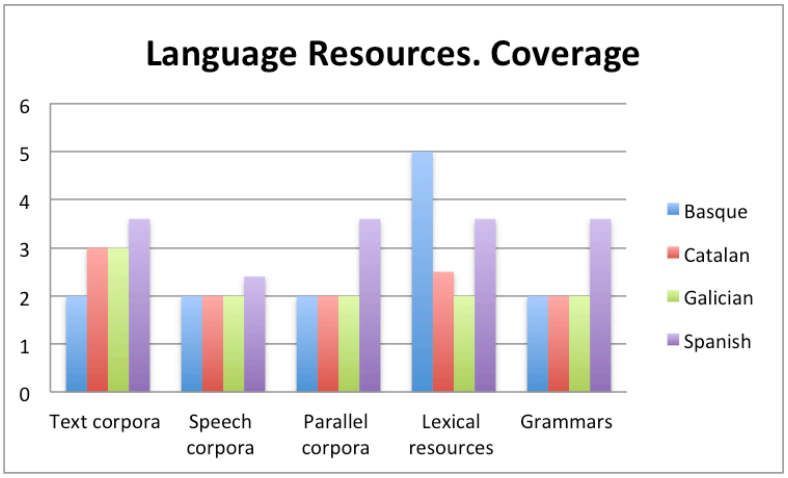

Fig. 5: Coverage scores for Language Resources

in web-based in-formation systems and as semantic knowledge bases in LT applications – ideally, this endeavour should be addressed in a multilingual way on the European scale.

To conclude, in a number of specific areas of language research, we have software with limited functionality available today. Obviously, further research efforts are required to meet the current deficit in processing texts on a deeper semantic level and to address the lack of resources such as parallel corpora for machine translation.

3.1 Cross-language comparison

The current state of LT support varies considerably from one language to another. In order to compare the situation between languages, this section presents an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The languages were categorised using the following five-point scale:

1. Excellent support
2. Good support
3. Moderate support
4. Fragmentary support
5. Weak or no support

LT support was measured according to the following criteria:

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

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.

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 analysis shows that, thanks to large-scale LT funding in recent decades, Spanish is better equipped than most other languages. It compares well with most large languages, such as French and German. But LT resources and tools for Spanish clearly do not yet reach the quality and coverage of comparable resources and tools for the English language, which is in the lead in almost all LT areas. And there are still plenty of gaps in English language resources with regard to high quality applications.

The results also show that, thanks to LT funding programmes from the Local and Central Governments in recent decades, the Basque, Catalan, and Galician languages are equipped as most of other European languages. They compare well with languages with a similar number of speakers such as Hungarian, Greek, and European Portuguese despite these are official languages of EU countries. But LT resources and tools for these languages clearly do not yet reach the quality and coverage of comparable resources and tools for the Spanish language, which is in a good position in almost all LT areas.

For speech processing, current technologies perform well enough to be successfully integrated into a number of industrial applications such as Interactive

Voice Response systems and constrained domain dictation systems. Machine Translation systems get a good performance, especially between the language pairs Spanish-English, Spanish-Catalan and English-Catalan. However, to build more sophisticated applications, such as machine translation, there is a clear need for resources and technologies that cover a wider range of linguistic aspects and allow a deep semantic analysis of the input text. By improving the quality and coverage of these basic resources and technologies, we shall be able to open up new opportunities for tackling a vast range of advanced application areas, including high-quality machine translation.

4 Conclusions

The results of the study in the series of white papers show that there are serious differences in the support of European languages in the field of language technologies. While for some languages and certain applications, there is a good coverage in both software and language resources, for other languages, especially for minority languages, there are serious deficiencies in some areas.

In the case of Basque, Catalan, and Galician, the results of the study concerning the support of existing language technologies, are cautiously optimistic. There is a very active research community that has been working on research projects funded by the Local Governments and the Central Government. They have produced and distributed a number of language resources. Compared with other European languages, the resources and technology support is at a similar level to Greek, Hungarian or Norwegian being all official languages of European countries.

The range of products and applications available in Spanish greatly exceed those produced in the other three languages and the position of Spanish is comparable to French or German.

Language technologies are increasingly complex and require a large amount of data and resources to advance significantly. There are needs to coordinate the various sources of regional, national and European funding programs to circumvent language technologies support differences between the European languages.

5 Acknowledments

The authors of this document are grateful to the authors of the White Paper on German [9] for permission to re-use selected language-independent materials from their document. The authors acknowledge the support offered by the EU project T4ME [7], and the EU project Metanet4u [10]. Authors want to acknowledge their contribution to the coauthors and contributors of the White Paper series: for the Galician version, to the co-author Monsterrat Arza Rodríguez, and furthermore, to thank Dr. Xavier G. Guinovart (University of Vigo), Dr. Eduardo R. Banga (University of Vigo), Dr. Xosé Luis Regueira (University of Santiago de Compostela) and Mr. José Ramom Pichel (Imaxin Software) for their contributions to the Galician white paper. Material available at the web

pages of 'Consello da Cultura Galega Proxecto LOIA" and 'Secretaría Xeral de Política Lingüística Xunta de Galicia" was also been used. For the Catalan White Paper, we thank to the coauthors Eva Revilla, Emilia Garcia and Dr. Sisco Vallverdú, and to the contributors Dr. Lluís Padró, Dr. José B. Mariño, Dr. José R. Fonollosa and Dr. Climent Nadeu (all from Universitat Politècnica de Catalunya), Dra. Toni Martí (Universitat de Barcelona), Joan Soler (Institut d'Estudis Catalans), and Mercè Lorente (IULA-UPF). For the Spanish White Paper, authors thank his contribution to Dr. Toni Badia (UPF). For the Basque White Paper, authors thank his contribution to E. Navas (UPV/EHU), I. Odriozola (UPV/EHU) , K. Sarasola (UPV/EHU) , A. Daz de Ilarraza (UPV/EHU) , I. Leturia (Elhuyar Foundation), A. Daz de Lezana (Basque Government), B. Oihartzabal (CNRS), J. Salaberria (CNRS).

References

1. "Strategic Research Agenda for Multilingual Europe 2020", META Technology Council (eds.), 2012, META-NET, available at http://www.meta-net.eu/sra
2. Rehm G., and Uszkoreit H. Multilingual Europe: A challenge for language tech. MultiLingual, 22(3):5152, April/May 2011.
3. Federmann C., Giannopoulou I., Girardi C., Hamon O., Mavroeidis D., Minutoli S., Schrder M. META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey. 2012
4. META-NET White Paper Series: Europe's Languages in the Digital Age Georg Rehm and Hans Uszkoreit Eds., Spinger. 2012
5. Moreno A., Bel, N., Revilla E., Garcia, E., Vallverdú S. The Catalan Language in the Digital Age. Springer (2012)
6. Melero M., Badia T., Moreno A. The Spanish Language in the Digital Age. Springer (2012)
7. Hernáez I., Navas E., Odriozola I., Sarasola K., Daz de Ilarraza A., Leturia I., Daz de Lezana A., Oihartzabal B., Salaberria J. The Basque Language in the Digital Age. Springer (2012)
8. García-Mateo C., Arza, M. The Galician Language in the Digital Age. Springer (2012)
9. Burchardt A., Egg M., Eichler K., Krenn B., Kreutel J., Leßmöllmann A., Rehm G., Stede M., Uszkoreit H., Volk M. Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age. Springer 2012.
10. http://www.meta-net.eu
11. http://www.meta-net.eu/whitepapers
12. http://t4me.dfki.de/
13. http://www.meta-net.eu/projects/cesar
14. http://www.meta-nord.eu/
15. http://metanet4u.eu