Language technology challenges of a ‘small’ language (Catalan)

Maite Melero(1), Gemma Boleda(2), Montse Cuadros(2), Cristina España-Bonet(2), Lluís Padró(2), Martí Quixal(1), Carlos Rodríguez(1), Roser Saurí(1)

(1)Fundació Barcelona Media – Centre d’Innovació
(2)Centre de Recerca TALP
Dept. de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

Abstract

In this paper, we present a brief snapshot of the state of affairs in computational processing of Catalan and the initiatives that are starting to take place in an effort to bring the field a step forward, by making a better and more efficient use of the already existing resources and tools, by bridging the gap between research and market, and by establishing periodical meeting points for the community. In particular, we present the results of the First Workshop on the Computational Processing of Catalan, which succeeded in putting together a fair representation of the research in the area, and received attention from both the industry and the administration. Aside from facilitating communication among researchers and between developers and users, the Workshop provided the organizers with valuable information about existing resources, tools, developers and providers. This information has allowed us to go a step further by setting up a “harvesting” procedure which will hopefully build the seed of a portal-catalogue-observatory of language resources and technologies in Catalan.

1. Introduction

The impact of language technology in everyday life is getting ever broader. This includes a full array of devices and services, such as automatic call centres, on-line machine translation services, language checkers, mobile phone applications, car navigation systems, etc. In this context, speakers of ´small´ languages (lesser spoken languages or languages without a State behind them)[1] may find themselves at a comparative disadvantage to speakers of bigger languages, given that companies often offer their products only in languages with large markets, such as English or Spanish.

With the laws of the market against them, the actors involved in developing technology for small languages need to sharpen their wits to make the best of what they have.

With this idea in mind, a group of people involved in the computational processing of Catalan organized last March the *I Jornada del Processament Computacional del Català* (First Workshop on the Computational Processing of Catalan) in Barcelona, which assembled most of the relevant people in the field (at least from the academic world)[2].

Conceived as an opportunity for exchanging technical and scientific knowledge, and for promoting collaboration among the research groups involved, the Workshop succeeded in creating awareness of the challenges that the community is facing and of the need for a joint initiative to address them, which should include researchers but also industry and administration.

We present a brief snapshot of the state of affairs in computational processing of Catalan and the initiatives that are starting to take place in an effort to bring the field a step forward, by making a better and more efficient use of the already existing resources and tools, by bridging the gap between research and market, and by establishing periodical meeting points for the community of researchers, developers and consumers of linguistic technology in Catalan.

We believe that the challenges identified here are shared by other ‘small’ languages, which could potentially benefit from the initiatives that we propose.

2. Language technologies in Catalan: state of affairs

Natural language technology for Catalan is in better shape than could be expected, considering demographic factors and the degree of political influence of the language. Catalan has certainly benefitted from the fact that some of the most important research groups working on a much bigger language (Spanish) are located in Catalan speaking territories and have found opportunities to work also on Catalan on the side.

From the point of view of resources, Catalan boasts, among other assets, a high-quality manually annotated corpus of 55M words that was assembled for over a decade thanks to the foresight of Dr. Joaquim Rafel of the Institut d’Estudis Catalans (Rafel, 1994). The publication of a daily paper in two versions, Spanish and Catalan is also a major source of parallel texts for these two lan-

[1] In our view there are two groups of languages which enter the category of “small” (others may use the term “minor” with the same meaning): languages with relatively few number of speakers, even if endorsed by a state (Latvian, Estonian, but also Finnish, Norwegian, or Dutch); and languages without a national state (Catalan, Basque, Galician, but also Quechua or Kurd). The term “minorized” (as opposed to minority) referring to the latter has gained popularity mostly in the Iberian literature where it is considered more politically correct. Moreover we are not specifically addressing the problems of “languages with fewer resources”, but rather the problems that face languages with strong competitors for the language technology market share.

[2] http://sites.google.com/site/jornadacatala/Home

guages.
As for research groups doing some work on this language, there are around a dozen in Catalonia and three more in Valencia. To the extent of our knowledge, nobody else is working on natural language processing for Catalan in the rest of the Catalan speaking territories, although there are some organizations working on language resources (terminology, glossaries, etc.) as well as companies using language products (IB3, El Periòdic Andorrà, etc.). It is worth noting that, due to the interdisciplinary nature of language technology, research groups come evenly from the technical world (computer science and engineering departments) and humanities (language and translation departments). Most of the groups are based in universities and do either theoretical or applied research, with at least one of them specializing in technology transfer, although all the groups deal with this issue in some way or another. Funding for the projects comes from the local government, the state government, or the UE. Other sources of funding are contracts with companies or with the administration.
Most research going on for Catalan is devoted to written text: ten out of fourteen groups focus only on written data; three work with both modalities and there is just one group working exclusively on speech. In this last modality, the basic areas covered are speech synthesis and analysis. As for written language, research concentrates on the creation of both general processing tools (taggers, parsers etc.), as well as resources (corpora, lexica, grammars, etc.). All groups are more or less involved in both activities, but groups in technical departments tend more toward tool development, whereas groups in the humanities are more prone to resource building. Machine translation is the star application: nine out of fourteen groups do some work in that area. Next comes language checking and automatic text summarization (3 groups), and then a range of applications such as information extraction, automatic text generation, or paraphrases detection.
Among the companies involved in language technology for Catalan, a handful of them offers machine translation (AutomaticTrans, Transducens, Translendium), others provide text correction (Barcelona Media, Maxigramar, Thera), information extraction and retrieval (Barcelona Media, Inbenta, Thera), speech synthesis (Barcelona Media, Loquendo, Telefónica I+D, Verbio), and speech analysis (Telefónica I+D, Verbio). Similarly, there are a number of public bodies working on the creation and management of textual and lexical resources (Assessorament Lingüístic de la Corporació Catalana de Mitjans Audiovisuals, Institut d'Estudis Catalans, TERMCAT), and on the creation and exploitation of different tools and resources, such as search engines, translation memories, style guides, etc. (Servei Lingüístic de la UOC, Servei Lingüístic de la UB, Comissió Jurídica Assessora de la Generalitat de Catalunya, among others).
Furthermore, there are some entities, associations, and non-profit organizations that do their bit for normalizing Catalan in the technological domain, most notably Softcatalà and VilaWeb.
Thanks to this collective effort, most of the basic tools and resources for language processing actually exist for Catalan, although they are not always well-publicized or available to the research community, industry, or general public.

3. Language technology needs and challenges faced by Catalan

Official bilingualism in the Catalan territories entails that –contrary to other bilingual countries, such as Belgium or Canada- most residents are actually bilingual. All Catalan speakers in practice are also fluent in Spanish, which can explain the relatively low social demand for technological products in Catalan. This situation results, in turn, in a lack of interest from the industry or the business world in general to invest in Catalan, Spanish being a much more profitable market.
As seen, research and development is often confined to the academic world or to altruistic initiatives, with little connection with the real world with the honourable exception of Softcatalà. Oftentimes, the technology already exists, has been developed in some laboratory but needs the final push that would turn it into a real product.
There is a general lack of communication among the actors involved in the field of language processing in Catalan. The companies developing products do not have information of what kind of technology has already been developed by research groups; the administration does not have clear guidelines on what should be publicly supported, and what not; the research groups themselves do not always know what resources or tools are already available and often duplicate efforts; and, finally, there is a big gap between developers and final users. The developers community knows little or nothing about user needs, be it a final user or a service company.
While good coordination and collaboration policies among the different actors (here, research groups, administration, and companies) are necessary for any endeavour, this is a must in the case of technology development for a small language like Catalan. Language resources are expensive, funding is limited, and the community cannot afford duplication and dispersion of efforts.
We also think that the vicious circle of market indifference and disinterest from industry could potentially be broken if language products in Catalan were able to reach the market and proved to be good selling items by themselves. Particularly, if they were better than their counterpart in Spanish or English, the consumer would spontaneously choose them over the rest.
In addition, open source initiatives, such as Apertium (Forcada, 2006), a machine translation system originally developed for Catalan-Spanish by Universitat d'Alacant, could likely bridge the gap between developers and consumers thanks to the community of users that usually builds around this type of applications. OS user communities share applications and are quick to adapt them to their needs and interests. This type of community can successfully bring together researchers, developers, service companies, and final users.

Finally, it is also worth mentioning that in the research world itself, languages that are big in resources, such as English, act as attractors for research efforts and only get bigger thanks to the activity of thousands of non-English speaking researchers. In order to counterbalance this situation, it is essential to favour the presence of linguistic resources of smaller languages (e.g. annotated corpora) in international computational contests.

4. Addressing the challenges through a grassroots initiative: the First Workshop on the Computational Processing of Catalan

The First Workshop on the Computational Processing of Catalan took place in March 2009, promoted by researchers from two centres who perceived the need to enhance the communication and collaboration among the different groups involved, and wanted to find ways to make an efficient use of the existing resources, while at the same time increase the level of visibility of the field to the world.

The Workshop was very well received by the community. Not only did it succeeded in putting together a fair representation of the research in the area, but it received also attention from the industry and was endorsed by the administration. Aside from the oral presentations from the 14 research groups attending the event, around 30 different tools, applications, and resources were presented during a demo and poster session to a diverse audience including members of the academic community, people from the industry, representatives of the administration and general public.

The participation numbers, classified according to the registration data is shown in **Table 1**.

Category	*Regist.*	*%*
Research centre	109	64,1
Industry	23	13,5
Administration	15	8,8
Other	23	13,5
Total	170	100

Table 1. Workshop's audience according to origin.

Table 2 contains the full list of the groups that presented their research, together with a summary of their lines of research and web site.

Research centers	**Lines of research**	**URL**
Voice and Language Group (Barcelona Media Innovation Centre)	Information extraction; Opinion mining; Grammar checking; Machine translation; Speech synthesis; Sign language; E-learning.	http://www.barcelonamedia.org/linies/7/ca
Research Group on Natural Language Processing (Universitat Politècnica de Catalunya)	Linguistic analyzers; Acquisition, integration and exploitation of lexico-semantic knowledge; Semantic analysis (WSD, SRL, NLU); Machine learning; Automatic summarization; Machine translation.	http://www.lsi.upc.edu/nlp
NLP group at the Department of Information and Communication Technologies (TALN) (Universitat Pompeu Fabra)	Multilingual natural language generation; Automatic summarization; Computational lexicology.	http://www.recerca.upf.edu/taln
Linguistic Services (Universitat Oberta de Catalunya)	Machine translation; Computer assisted translation and correction	http://www.uoc.edu/serveilinguistic
FlexSem, Departament de Filologia Francesa i Romànica (Universitat Autònoma de Barcelona)	Prosody; Lexical semantics; Computational lexica; Linguistic analyzers	http://grupsderecerca.uab.cat/flexsem/
Computer Linguistics Centre (CLIC) (Universitat de Barcelona)	Corpora processing and annotation; Morphological and syntactic analysis; Co-reference resolution; Paraphrases	http://clic.ub.edu
Institut Universitari de Lingüística Aplicada (IULA) (Universitat Pompeu Fabra)	Information extraction; Computational lexicology; Terminology; Forensic linguistics; Demolinguistics; Sociolinguistics	http://www.iula.upf.edu
Media Technologies Department (La Salle – Universitat Ramon Llull)	Speech synthesis; Avatars; Sign language; Speech recognition; Emotion recognition; Oral corpora	http://www.salle.url.edu/portal/departaments/home-depts-DTM
Linguistic Applications Inter-University Research	Computational grammars (dependency, HPSG); Lexical resources;	http://grial.uab.es

Group (GRIAL)	Corpora (annotation and exploitation); Lexical information acquisition; Translation memories	
Pattern Recognition and Human Language Technology, Instituto Tecnológico de Informática (Universidad Politécnica de Valencia)	Machine translation (text-to-text and speech-to-speech); Computer assisted translation; Speech recognition; Handwriting recognition; Biometrics; Interactive image mining.	http://prhlt.iti.es
Center for Language and Speech Technologies and Applications (TALP) (Universitat Politècnica de Catalunya)	Speech recognition; Text-to-speech; Machine translation (text-to-text and speech-to-speech); Linguistic resources building (text and speech); Multimodal interfaces.	http://gps-tsc.upc.es/veu
Transmedia Catalonia (Universitat Autònoma de Barcelona)	Audiovisual translation (dubbing, subtitling, voice over); Media accessibility (audiodescriptions, subtitling); Accessibility for teaching purposes	http://grupsderecerca.uab.cat/transmediacatalonia/en
International Translation Virtual Institute (IVITRA) (Universitat d'Alacant)	Old Catalan corpus building; Parallel corpora; Corpus processing	http://www.ivitra.ua.es
Transducens group, Department of Software and Computing Systems (Universitat d'Alacant)	Computer assisted learning; Digital libraries (indexation and marking languages); Automatic rule inference; Machine translation; Phonetic annotation; Alignment tools for parallel texts; Translation memories generation	http://transducens.dlsi.ua.es

Table 2. List of research groups participating in the *I Jornada del Processament Computacional del Català*

The success of the Workshop and the interest it arose, as well as the quantity and quality of the groups, resources, and applications were an evidence of the vitality of technology for Catalan, but the debate that ensued also revealed its weaknesses and the challenges it needs to address.

Where to go from here? How to strengthen the cooperation among the different actors, including research, developers, service companies, administration, and users? Two concrete proposals arose from the meeting, namely the creation of a web portal that brings together all the information concerning the computational processing of the Catalan language (a catalogue-observatory of tools, resources, publications, and actors, along the lines of the Portuguese *Linguateca* (Santos, 2009)); and the creation of an association able to voice the needs and concerns of the community.

As a first step, a distribution list was created with all the participants to the workshop as initial members with the idea to use it as a seed for the future association. Through the list different working groups have been organised to deal with the different initiatives: (1) create the portal-catalogue; (2) constitute a formal association and search for funding sources to set up a minimal infrastructure; (3) elaborate a survey of interests and needs of companies and final-users of language technology in Catalan, and (4) turn the workshop into a periodic event, with an increased participation of the industry.

5. Making language resources available: Getting ready for the harvesting day!

According to (Binnenpoorte et al., 2002) establishing a roadmap for Human Language Technologies for a given language requires, in the first place, (i) to define the minimum requirements for an adequate digital language infrastructure for that language, e, g. in the form of a Blark matrix[3]; secondly, (ii) to have an accurate understanding of the current situation of HLT in that language; and last but not at all least! (iii) to guarantee that, at any rate, what is required is available (and remains so). Visibility and availability of resources is therefore the key issue for success.

On one hand, the volume of digital data grows exponentially every day. On the other hand, the increased maturity in language technologies results in more tools and applications. Hence, the current problem for most languages is not the lack of resources but their invisibility and fragmentation.

In an attempt to address this critical issue, the UE has sponsored several initiatives to inventory linguistic resources, such as the well-established ELRA[4] catalogue or the CLARIN project[5], focused on making HLT available

[3] Basic LAnguage Resource Kit: http://www.blark.org/

[4] European Language Resources Association: http://www.elra.info/

[5] Common LAnguage Resources and technology Infra-

to Humanities scholars.

Two ambitious initiatives have been recently launched in a similar spirit: FLaReNet[6] and META-NET[7]. Both networks have the mission to minimize the impact of the linguistic diversity in a digital and multilingual Europe: the former through the preparation of strategies and recommendations, and the promotion of standards for HLT; and the latter through a strategic alliance among the main actors of the field (research, industrial providers, users, policy makers and funding agencies) and through the creation of an Open Resource Exchange.

These trends in the European HLT world fit nicely with the aspirations that the Catalan LT community expressed in the context of the Workshop, and give indications on the path to follow. The information gathered in preparation for the Workshop and during the Workshop gives us an accurate understanding of the current situation (cf. Binnenporte's second requirement above) and is a valuable first step towards building the much demanded resources and tools catalogue.

However, beyond the effort of building the catalogue, there is the maintenance issue. Citing (Villegas and Parra, 2010) who reviewed about 800 resources during 2009 in the framework of the FLaReNet project: "The compilation of information for this first survey was harder than expected because of the lack of documentation for most of the resources surveyed. Besides, the availability of the resource itself is problematic: Sometimes a resource found in one of the catalogues/repositories is no longer available or simply impossible to be found; sometimes it is only possible to find a paper reporting on some aspects of it; and, finally, sometimes the information is distributed among different websites, documents or papers at conferences. This made it really difficult to carry out an efficient and consistent study, as the information found is not always coherent (e.g. not every corpus specifies the number of words it has) and sometimes it even differs from the one found in different catalogues/repositories."

Being a 'small' language community, we are sharply aware of the high costs required to maintain such a catalogue up to date. Maintenance efforts need to be efficiently distributed. Resource and tool providers must be made aware of the importance of guaranteeing the visibility of their own resources but they must be endowed with tools to make that easy.

In the FLaReNet Forum (The Future of Language Resources) that took place in Barcelona on February 2010, Marta Villegas and Carla Parra from the Spanish FLaReNet site announced the "Harvesting Day"[8]. They propose to start a decentralized effort of resource description and to launch an automatic, periodical information gathering routine. Each developer must enhance and guarantee the visibility of his/her resources by minimally describing them with a Basic Metadata Description. This information will have to be made automatically harvestable in a server by different robots. In order to easy the setting up of this initiative they will provide an online form that will automatically create the required XML for harvesting the information about resources and tools. Self-executable packages for setting up harvestable servers will also be provided.

In this way a provider just needs to fill in the online form, save the BAMDES XML file that describes the resources and place it in a server with the self-executable package.

The automatic harvesting of metadata will then be possible and will supply with results the main catalogues and observatories, enhancing and guaranteeing the visibility of resources and tools and ensuring that the information available about them is always up-to-date, as the harvesting will take place periodically.

Using the information gathered during the Workshop on Computational Processing of Catalan, a first pilot on resource harvesting will shortly be conducted, which will collect data about tools and resources in Catalan and incorporate it to an embryonic portal-catalogue-observatory of Human Language Technologies for the Catalan language.

6. Conclusion

In this paper, we discuss the challenges faced by language technologies in Catalan and present the results of the First Workshop on the Computational Processing of Catalan. The Workshop succeeded in creating community awareness, provided us with a snapshot of the state of affairs in language technologies in Catalan and triggered a series of actions aimed at further enhancing cohesion and communication among the different actors in the Computational Processing of Catalan.

Most interestingly, we have been able to launch a concrete proposal to tackle the problem of invisibility and fragmentation of resources, namely a "harvesting" procedure which will hopefully build the seed of a portal-catalogue-observatory of language resources in Catalan.

7. Acknowledgements

The First Workshop on the Computational Processing of Catalan was sponsored by the Generalitat de Catalunya, Barcelona Media Centre d'Innovació and the Universitat Politècnica de Catalunya.

We also wish to thank the research groups and particular researchers who presented their tools and resources at the Workshop, as well as all the participants to the event.

A special thanks to Marta Villegas for helping in setting up the "harvesting" procedure for HLT in Catalan.

8. References

Binnenpoorte, Diana, Catia Cucchiarini, Elisabeth D'Halleweyn, Janienke Sturm and Folkert De Vriend (2002). Towards a roadmap for Human Language Technologies: Dutch-Flemish Experience.

structure: http://www.clarin.eu/

[6] Fostering Language Resources Network: http://www.flarenet.eu/

[7] Multilingual Europe Technology Alliance: http://www.meta-net.eu/

[8] The date is set to June 21st, 2010

Forcada, Mikel L. (2006) Open-source machine translation: an opportunity for minor languages in *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)* (organised in conjunction with LREC 2006 (22-28.05.2006))

Rafel, Joaquim (1994). Un corpus general de referència de la llengua catalana. Caplletra. Vol. 17, p. 219-250

Santos, Diana (2009). Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. Linguamática 1.1, pp. 25-59

Villegas Marta and Parra, Carla (2010). The Metadata Harvesting Day. Position paper at the 2nd FLaReNet Forum http://www.flarenet.eu/sites/default/files/S1_Villegas-Parra_Position_Paper.pdf