Promoting a less resourced language via Language Technology for translation, content-management and learning

The Basque language is one of the oldest alive in Europe, although it has suffered continuous regression over the last centuries. However, many citizens and local or regional governments have been promoting its recovery since the 1970s. Now Basque holds partial co-official language status in the Basque regions of Spain but it has no official standing in the Northern Basque Country in France, neither in the European institutions. Today, there are about 700,000 Basque speakers, around 25% of the total population of the Basque Country, but they are not evenly distributed, and its use in industry and especially in Information and Communication Technology is still not widespread. A language that seeks to survive in the modern information society has to be present also in such fields and this requires language technology products. Basque, as other minority, languages has to make a great effort to face this challenge (Williams et al., 2001).

In this context, BerbaTek is a strategic research project with a duration of three years (2009-2011). Its consortium is made up of the Elhuyar Foundation, the IXA and Aholab research groups of the University of the Basque Country, and the technology centres Vicomtech and Robotiker. The BerbaTek project is partly funded by the Departments of Industry and Culture of the Government of the Basque Country. The members of the consortium have been collaborating since 2002 in other two previous projects: Hizking (Diaz de Ilarraza et al., 2003) and Anhitz (Arrieta et al., 2008). In those projects basic foundations, tools and applications were created for Basque Language, having developed several applications that are effective tools to promote the use of Basque. From our experience we defend that research and development for less resourced languages should be faced following these four points: (1) high standardization, (2) open-source, reusing language foundations, tools, and applications, and (4) incremental design and development. We think that any HLT project related with a less privileged language should follow those guidelines, but from our experience we know that in most cases they do not. We believe that if Basque is now in a fairly good position in HLT is because those guidelines have been applied, even though in some cases it was easier to create "toy" resources or easily obtainable tools.

The main aim of the BerbaTek project is the research and development of language-, speech- or multimedia-technologies so that they can provide the technological basis to support the economic sector of the language industries in the Basque Country, which consists of the translation, teaching and content industries. This sector needs to be able to avail itself of advanced technologies so that it can be competitive in a globalised, interconnected and multilingual environment.

In addition to research into these technologies, the BerbaTek project has the following aims:

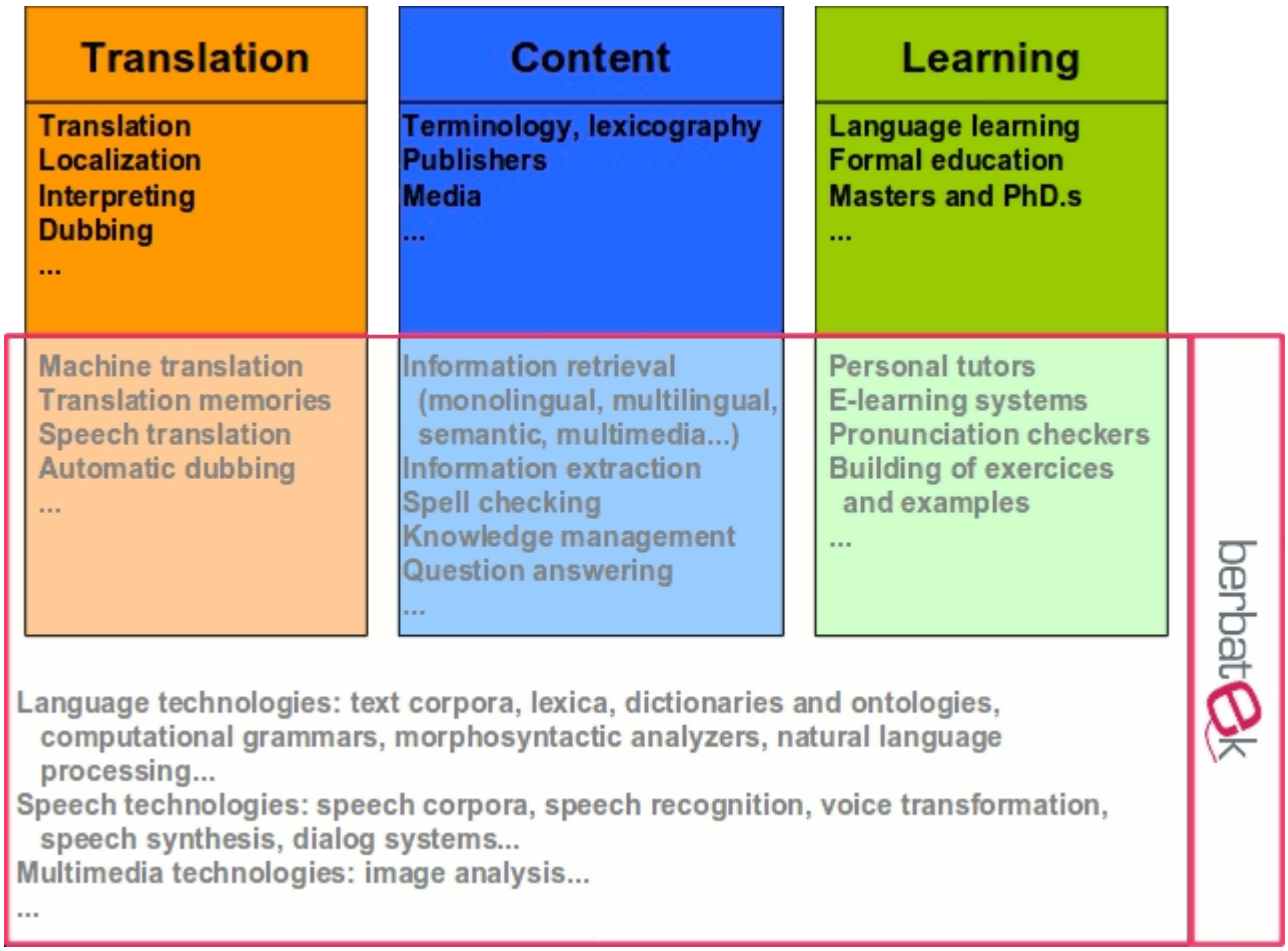
- To take a new significant step forward in the strengthening of the language industries by incorporating the results and devices into real market scenarios (home, electronics and industry), to make the marketing of them viable.
- To set up new technology-based firms and consolidate the existing ones on the basis of the results of the project.
- To generate knowledge and train a qualified critical mass that will enable the future of R+D within this strategic line to be tackled in an autonomous way. To achieve this, training courses, postgraduate Masters, specialised seminars and PhD courses in the sphere of language technology are run.

To meet these aims, the members that are part of the BerbaTek project have at their disposal a scientific and research community with the necessary internationally recognised critical mass, in addition to preferential relations with many companies in the language industry.

The BerbaTek project is geared towards applications. Without neglecting basic research, it is endeavouring to present experimental applications which can subsequently be developed and turned into products by companies. In any case, development geared towards applications and which has been found to be viable generates highly important basic as well as applied lines of research.

The importance of generating knowledge in the area of language technologies for voice and multimedia lies in their potential for application mainly in the language industry sector:

- Translation: interpretation, dubbing, localization, human translation.
- Content industry: Internet, audiovisual sector, multimedia, off- and on-line publishing, etc.
- Training: language learning, technical and professional education, ongoing training, etc.



Although there is no reason why they should not have applications in many other industrial sectors, such as telecommunications (speech, messaging, voice and data integration...), telesales and telesupport (e-commerce, call centres, after-sales services, telebanking and all the forms of business that take place through the telecommunications networks...), business processes (knowledge acquisition, publishing, localization and association of corporate information, office automation...), interactive elements (control systems, navigation and steering, both for global systems and for small pieces of equipment and domestic products, including home automated systems), leisure and entertainment (games, computer stories, adventures and virtual trips, etc.) or public administration (citizens' service, e-Government...).

What is more, as the BerbaTek project is the first to include the different areas comprising the economic sector of the language industries as such, it is set to facilitate the structuring of the sector. The development of language technologies will also have a positive impact on the globalisation process of Basque companies of all kinds, as they will enable the leap to the multilingual scenario in which translation and language teaching are key factors to be addressed more effectively. Also the fact that the Basque Country is a multilingual community offers know-how and an unbeatable test bench for the development not only of language technologies, but also of the Language Industry, and offers a good starting point with a view to positioning the Basque Autonomous Community among top countries in language and voice technologies.

Throughout the BerbaTek project, we are creating some demos to show the potential of the integration of language-, speech- and multimedia-technologies, when it comes to creating applications for the areas of language industries, that is, for translation, contents and teaching. These are the demos we are building:

- Automatic dubbing of documentaries into Basque using subtitles in Spanish (with possible automatic creation of the Spanish subtitles from the Spanish audio, by means of ASR).
- Multimedia and multilingual semantic web search engine on science and technology content, including posterior navigation through related content or similar images.
- Personal tutor in language learning through a speech-driven avatar, with automatically created grammar and comprehension exercises, writing aids (dictionaries, writing numbers, spelling...) and automatic evaluation of pronunciation.

References

Arrieta K., Arantza Diaz de Ilarraza, Inma Hernáez, Urtza Iturraspe, Igor Leturia, Eva Navas, Kepa Sarasola 2008 AnHitz, development and integration of language, speech and visual technologies for Basque Second International Symposium on Universal Communication OSAKA, pp. 338-344, 530-0005, JAPAN. Published by IEEE Computer Society. ISBN: 978-0-7695-3433-6 http://doi.ieeecomputersociety.org/10.1109/ISUC.2008.43

Díaz de Ilarraza A., Gurrutxaga A., Sarasola K., A. Gurrutxaga, I. Hernaez, N. Lopez de Gereñu 2003 HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities *Workshop on NLP of Minority Languages and Small Languages. TALN 2003. Nantes*

Williams B., K. Sarasola, D. Ó´Cróinin, B. Petek. 2001. Speech and Language Technology for Minority Languages. *Proceedings of Eurospeech 2001*.