The Digital Language Diversity Project

Claudia Soria, Irene Russo

Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Pisa, Italy

Abstract

In this paper we introduce *The Digital Language Diversity Project*, a three-year project funded under EC Erasmus+ programme started in September 2015. The project addresses the problem of the scarce use and usability of EU regional and minority languages over digital devices by developing a training programme for adult speakers of regional and minority languages to empower them with the know-how for creating and sharing digital content. Availability of digital content and technical support to collect it are essential prerequisites for the development of language-based digital applications, which in turn will boost digital usage of these languages.

Keywords: Less-resourced languages, Language Technology, digital language vitality, digital language diversity

1. Introduction

Linguistic diversity is one of the richest heritages of Europe. It has been estimated that some 250 languages are currently spoken in Europe, but only 24 are official languages of the EU and just a handful more are nationally recognised as official. Some 80-90 languages fall into the category of regional and minority languages (hereafter RMLs). Preservation of this extraordinary diversity is crucial to secure the construction of Europe on grounds of mutual respect and equal opportunities for all citizens.

The status and vitality of RMLs is of major concern: in order to raise their chance of survival in the future, the range of opportunities to use them must be increased. Digital media and tools represent only one of these contexts of use, yet they are fundamental to secure survival for these languages (Crystal, 2010). As citizens' life makes an increasingly extensive use of digital devices, a language's digital presence is of utmost importance to be perceived as fitting the needs of modern world. Eisenlohr (Eisenlohr, 2004), for instance, argues that a presence in new technologies facilitates better appreciation of a language, by establishing a positive association with modernity and relevance to current lifestyles.

Europe's RMLs have only minimal digital representation (Rehm et al., 2014), (LT-Innovate.eu, 2013), either because they are not "digitally ready", i.e., they don't enjoy the range of tools and technical support available for other major languages, or simply because their perceived profile is so low that speakers turn to other languages when accessing the digital world. As a result, the amount of information and services that are available in less widely spoken languages is reduced, thus creating inequality at several different levels: of linguistic rights and digital opportunities for all languages and all citizens; inequality of information and access to services; uneven access to technological development, and uneven opportunities for language survival.

To increase the digital representation of smaller (i.e. regional, minority, or minoritised languages) their use and usability over the Internet and through digital devices needs to be supported by Language Technologies. These, however, appear to be lacking, even for major languages and EU official languages: the research carried out by the META-NET Network of Excellence and culminated in the publications of 30 *Language White Papers* (Rehm and Uszkoreit, 2012) argues how 29 European languages are at risk of digital extinction because of lack of sufficient support in terms of language technologies. Consequently, we should focus even more on fostering technology development for smaller and/or less-resourced languages and also on language preservation through digital means.

Since lesser-used languages are of little economic interest to the major players and developers of language-based digital applications, provision of state-of-the-art language-based applications, which would enable and foster their use over digital media and devices, is severely limited (Mariani, 2015). At the same time, the current data-driven paradigm of development of Language Technologies makes production of digital data a major bottleneck, as the development of language-based applications crucially depends on the availability of large quantities of good-quality open data (Soria et al., 2014). The long-term aim of DLDP is to contribute to breaking this vicious circle by increasing the amount of content available for less widely used languages, and therefore paving the way to advancement in the provision of state-of-the-art products and services allowing use of RMLs on digital devices.

2. Background

The Digital Language Diversity Project (hereinafter DLDP), is a three-year project started in September 2015 and funded by the European Commission under Erasmus+ programme as a strategic partnership in the adult education sector. The specific mission of DLDP is to advance the sustainability of Europe's regional and minority languages in the digital world by empowering their speakers with the knowledge and abilities to create and share content on digital devices. The DLDP projects complements some past EU projects, most notably the META-NET[1] and FLaReNet[2] projects, by implementing their respective recommendations regarding actions aimed at increasing the availability of digital content in as many as possible languages, raising awareness among policy makers and Language Technology providers, and devising and adopting appropriate strategies for making larger use of ICT to enhance language learning, promotion and vitality of all languages.

In this respect, DLDP's innovation lies in its specific focus

[1] www.meta-net.eu

[2] www.flarenet.eu

on regional and minority language, mainly targeting with its training programme speakers accessing ICT to promote and preserve their languages. It is vital in the revitalisation effort that regional and minority language communities (many of which are also endangered) can access ICT for the purposes of language training (in formal, informal and vocational education, and promotion).
There are several initiatives very closely related to our project, some underway and some completed, focusing on relevant parts of the larger problem of preservation of Europe's linguistic diversity, such as awareness raising and educational projects (INNET project)[3] together with use and availability of ICT-based tools for revitalization purposes (FinUgRevita[4], DigiSami[5], Taalweb Frysk[6], or Hacio'r Iaith[7]) . In their respect, the main innovation and complementary character of DLDP is twofold: 1) the concrete, hands-on approach at providing practical and actionable examples for regional and minority language speakers about how to produce content using digital tools and 2) a focus on the production of content rather than the development of tools (even if these are two closely linked aspects), providing recommendations that can be applied to as many minority languages as possible beyond those representing our case-study. From this point of view, DLDP shares many similarities with *Activismo Digital de Lenguas Indigenas*[8], a network of Latin American digital activists targeting indigenous languages of Latin America.

3. Objectives and Activities

The DLDP Consortium is composed by five partners: a research department in computational linguistics (CNR-ILC, Italy, DLDP Coordinator), an NGO working for the protection and promotion of Europe's regional minority and endangered languages (ELEN, Brittany, France), an association for the safeguard of Karelian language and culture (Karjalan Kielen Seura, Finland), a foundation devoted to the promotion of Basque in science and technology (Elhuyar Fundazioa, the Basque Country, Spain) and the Department of Northern European and Baltic Languages and Cultures of Johannes Gutenberg University, Mainz (Germany).
Such an initiative can only be realized in transnational cooperation, because RML speaker communities tend to act in isolation and often lack an overall vision of similar experiences: a concerted approach is needed to avoid fragmentation of initiatives, capitalizing on shared experiences and helping cut costs. Moreover, a supra-national dimension is the only way speakers' communities can receive support where and when regional and national governments fail to adopt policies in favour of linguistic diversity.
In the short term, the immediate objectives of DLD project are:

- a Europe-wide applicable training programme taking into account the digital fitness of languages in focus targeted to RML speakers to guide them towards effective production of digital content and language learning materials in their languages;
- strong, clear and actionable recommendations - named "Digital Language Survival Kit" - about what is needed for a language "to go digital": which are the challenges and difficulties, which areas need to be addressed first, which tools are available;
- a roadmap, aimed at stakeholders and policy makers, detailing the institutional and technological challenges as well as the proposed solutions for paving the way to a more widespread use of all European languages over digital devices.

The major activities of the project will depend on a preliminary assessment of digital fitness that takes into account availability and accessibility of digital content in regional and minority languages and usability of those languages over digital media and tools. This assessment will be carried out by means of a survey for evaluating current use and usability of
Peer-learning and development of a training programme targeted to adult learners will be shaped on that; such activities, finalised to the achievement of common recommendations of best practices and of a blueprint for future actions for ensuring digital representation to regional and minority languages, will be carried out through networking, information and counselling activities, either during project's events and virtually via the project interactive web site.
Finally, the identification of various categories of stakeholders (individual speakers and speakers' communities; SMEs, Digital Content Providers and ICT developers; Policy Makers) will inform the structure and content of the *Roadmap to Digital Language Diversity*, a collection of concrete and realistic recommendations to ensure Europe's regional and minority languages an appropriate digital presence.

3.1. Training Programme

A DLD Training Programme will be created to support the long-term maintenance of Europe's regional and minority languages through fostering their digital presence. This innovative training programme will be available to speakers of regional and minority languages who want to learn why and how to increase the presence of his/her language online, and how to practically do it: which tools and techniques are available, which media are more suitable, which aspects to address first.
The programme will be modular, each module addressing one or more particular level of Digital Capacity, as defined by the tool developed by the *Digital Language Survival Kit* (see next section).
Although the precise structure of the Training Programme will be defined only later, at least the following topics are foreseen. Each topic will be recommended for one or more specific levels of digital capacity:

- overcoming intellectual barriers: why is it important for a language to be digital?

[3] http://www.innet-project.eu/
[4] http://www.ieas-szeged.hu/finugrevita/
[5] http://www.helsinki.fi/digisami/
[6] http://www.fryske-akademy.nl/nl/taalweb/
[7] http://haciaith.com/what-is-hacior-iaith-english/
[8] https://rising.globalvoices.org/lenguas/

- overcoming technical barriers;
- creation of textual material;
- creation of audio material: podcasts, web radio, YouTube channels;
- RMLs and Social Media: Facebook pages and groups, Twitter. How to build and enlarge a social community;
- bringing others on our side: software and interfaces' localization projects;
- edutainment: ebooks, videogames, etc.

3.2. The Digital Language Survival Kit

The Digital Language Survival Kit will be an instrument allowing regional and minority languages speakers and communities to self assess the degree of digital fitness of their language, by pinpointing current gaps and areas where action can and needs to be taken, learning about what concrete actions and initiatives can be put in place depending on the particular digital fitness level identified. For instance, a minimal degree of digital fitness will require a level of "digital survival capacity" (in increasing order of necessity): ensuring connectivity; development and adoption of a standardized encoding; development of a standardized orthography, some basic language resources (at least a corpus, a spell checker, and a lexicon).

Higher levels of digital fitness will require other types of measures, such as creating or enriching a Wikipedia in the language, having localized version of important sites, main operating systems and social media interfaces.

The Kit will serve as a companion of the training programme and will have a similar modular structure. In the framework of the DLD Project, the Kit will be fully developed for Basque, Breton, Karelian and Sardinian; its model and structure, however, will be designed so as to be applicable to as many languages as possible, thus ensuring circulation and adoption beyond the languages investigated in the project and after the project's lifetime.

3.3. A Roadmap to Digital Language Diversity

The aim of the Roadmap for Digital Language Diversity is to ensure that proper and adequate actions are taken in order to ensure an appropriate digital presence to Europe's regional and minority languages.

The Roadmap is intended to complement other previous and ongoing initiatives, such as the NPLD European Roadmap for Linguistic Diversity[9], the META-NET Strategic Agenda[10], and the FLaReNet Blueprint for Actions and Infrastructures[11]. Its innovative character lies in its specific focus on the particular needs and challenges of regional and minority languages.

The Roadmap will detail the major practical lines of action to be undertaken by different categories of stakeholders: individual speakers and speakers' communities, SMEs, Digital Content Providers and ICT developers, and Policy Makers. Concrete and realistic recommendations will be provided to prepare the ground for a EU-wide directive concerning the attainment of equal digital opportunities for speakers of all languages, in order to stop underrepresentation of some languages and create strong pressure on local policies in member countries.

These recommendations are therefore to be intended as a contribution to concrete, tangible and far-reaching measures for strengthening Europe's linguistic diversity.

[9] http://www.npld.eu/uploads/publications/313.pdf

[10] http://www.meta-net.eu/sra

[11] http://www.flarenet.eu/sites/default/files/D8.2b.pdf

4. Conclusion

As a cornerstone and valuable cultural heritage of Europe, linguistic diversity needs effective measures to ensure its safeguard and promotion. Any sustainable policy in this respect cannot ignore the digital world, as a prominent context of use of the languages. To date, this linguistic diversity is only partially reflected in the digital world: a small subset of the world's languages (about 6%, according to estimates) are allowed to access the digital sphere.

The wealth of EU regional and minority languages is severely underrepresented on digital media, and almost completely excluded from digital services which are usually available in EU national languages only. Speakers of EU regional and minority languages, therefore, experience unequal digital opportunities with respect to speakers of major languages. It is of foremost importance, therefore, that more and more opportunities are created for RML speakers to use their languages on digital media and tools.

The mission of DLDP is to advance the sustainability of Europe's regional and minority languages in the digital world by empowering their speakers with the knowledge and abilities to create and share content on digital devices. This, in turn, will create the necessary conditions for software developers to advance in the provision of state-of-the-art products and services allowing use of regional and minority languages on digital devices. It will also help to raise the profile of these languages decisively, especially in the eyes of the young generation, tomorrow's speakers.

5. References

D. Crystal. 2010. *Language Death*. Cambridge University Press.

P. Eisenlohr. 2004. Language revitalization and new technologies: Cultures and electronic mediation and the refiguring of communities. *Annual Review of Anthropology*, 18(3):339–361.

LT-Innovate.eu. 2013. Lt2013: Status and potential of the european language technology markets. LT-Innovate Report, March.

J. Mariani. 2015. How Language Technologies Can Facilitate Multilingualism. In *Proceedings of 3rd International Conference on Linguistic and Cultural Diversity in Cyberspace*, pages 48–60.

G. Rehm and H. Uszkoreit, editors. 2012. *META-NET White Paper Series: Europes Languages in the Digital Age*. .

G. Rehm, H. Uszkoreit, I. Dagan, V. Goetcherian, M. Dogan, C. Mermer, T. Varadi, S. Kirchmeier-Andersen, G. Stickel, M. Prys Jones, S. Oeter, and S. Gramstad. 2014. An Update and Extension of the META-NET

Study “Europe’s Languages in the Digital Age”. In *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*, Reykjavik, Iceland, May.

C. Soria, N. Calzolari, M. Monachini, V. Quochi, N. Bel, K. Choukri, J. Mariani, J. Odijk, and S. Piperidis. 2014. The language resource strategic agenda: the flarenet synthesis of community recommendations. *Language Resources and Evaluation*, 48(4):753–775.