Fostering digital representation of EU regional and minority languages: the Digital Language Diversity Project

Claudia Soria[a], Irene Russo[a], Valeria Quochi[a], Davyth Hicks[b], Antton Gurrutxaga[c], Anneli Sarhimaa[d], Matti Tuomisto[e]

[a]Istituto di Linguistica Computazionale “A. Zampolli”, Consiglio Nazionale delle Ricerche, Italy
[b]European Language Equality Network, France
[c]Elhuyar Fundazioa, Spain
[d]Johannes Gutenberg Universität Mainz - Forschungs - und Lehrbereich Sprachen Nordeuropas und des Baltikums, Germany
[e]Karjalan Kielen Seura, Finland

Abstract

Poor digital representation of minority languages further prevents their usability on digital media and devices. The Digital Language Diversity Project, a three-year project funded under the Erasmus+ programme, aims at addressing the problem of low digital representation of EU regional and minority languages by giving their speakers the intellectual an practical skills to create, share, and reuse online digital content. Availability of digital content and technical support to use it are essential prerequisites for the development of language-based digital applications, which in turn can boost digital usage of these languages. In this paper we introduce the project, its aims, objectives and current activities for sustaining digital usability of minority languages through adult education.

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

1. The ‘low digital representation’ problem

Europe’s regional and minority languages (RMLs) are poorly represented digitally (Rehm et al., 2014), (LT-Innovate.eu, 2013).

A number of factors can be invoked to explain it. One is the low profile enjoyed by many regional and minority languages, which often are not officially recognised and rarely fully supported (as it is the case of the totality of the regional languages of France, for instance). Low prestige and a weak socio-political profile make it so that speakers turn to other languages when accessing the digital world. The presence of RMLs over digital media and their usability through digital devices is usually limited to instances of digital activism and/or by means of cultural initiatives focused on the preservation of cultural heritage.

Another reason is the fact that virtually no European citizen is monolingual in a regional or minority language: everyone can always make use of an official EU language instead of a minority one, thus making EU regional and minority languages not essential for communication purposes. This makes EU RMLs of little economic interest for companies developing language-based digital applications, since virtually no prospective customer would be unable to communicate if these languages were not supported. As a consequence, 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).

In addition, for a language to be used digitally, it has to be “digitally ready”, i.e. it must enjoy the range of tools and technical support available for other major languages. This is not always the case, see for instance the recent battle for the adoption of a keyboard better supporting French regional languages[1]. The majority of EU RMLs is affected by the problem of weak technological support, with the notable exceptions of Basque, Catalan, Galician, Welsh and to a lesser extent, Frisian. The digital readiness of a language is inextricably linked to its digital presence: whenever a language is technologically supported and thus widely digitally usable, its digital representation expands. Digital data become easily and readily available to be exploited to develop new and better applications, which in turn will foster even wider use. This relationship between digital readiness and digital usability turns into a vicious circle for RMLs: development of language-based applications crucially depends on the availability of large quantities of good-quality open data (Soria et al., 2014), but this data can only become available if RMLs can start to be widely used digitally, and this requires the support of technology.

The consequences are not only technological, and not only on the short term. It has been argued that lack of digital usability represents a severe threat for languages. The META-NET research carried out by the META-NET Network of Excellence and culminated in the publications of 30 *Language White Papers* (Rehm and Uszkoreit, 2012) has clearly shown how 29 European languages are at risk of digital extinction because of lack of sufficient support in terms of language technologies. Obviously, the situation for regional and minority languages cannot be but worse, given their almost complete lack of technological support. Since everyday life makes an increasing extensive use of digital devices that involve language use, the usability of a language over digital devices is a sign for that language of being modern, relevant to current lifestyles and capable of facing the needs of the XXI century. A positive correlation between presence in new technologies and better appreciation of a language has been repeatedly observed in the literature, see for instance Eisenlohr (2004), Crystal (2010).

[1] http://www.afnor.org/liste-des-actualites/actualites/2015/novembre-2015/respect-de-l-ecriture-francaise-vers-un-nouveau-modele-de-clavier-informatique

2. The DLDP Training Programme: tackling the problem from the speakers' perspective

If linguistic diversity is to be preserved as an important heritage of Europe, with the potential of enforcing the construction of Europe on grounds of mutual respect and equal opportunities for all citizens, EU regional and minority languages need to secure their presence online and to start to be widely and fully used digitally. This means not only for cultural preservation purposes, but also to communicate in every context and to carry out the range of activities that are possible for widely-used languages, such as buy tickets, read ebooks, or translate a page in another language.

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.

As we have argued, language-based technological support can be better provided if digital content in regional and minority languages becomes widely and easily available.

The long-term aim of the *Digital Language Diversity Project* (hereinafter DLDP) is to contribute to breaking the "low digital representation - low digital readiness" vicious circle by empowering speakers of RMLs with the intellectual and practical skills that will put them in the position of creating and sharing digital content, at the same time motivating them to achieve this goal.

The project 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[2].

Given the educational approach of the Erasmus+ funding framework, the core of the project is represented by a Training Programme that will be made available online under the form of MOOC modules. Through the Training Programme, speakers of regional and minority languages will learn why and how to increase the presence of their language online, and how to practically do it: which tools and techniques are available, which media are more suitable, which aspects are to be addressed more urgently.

Each module will be ranked so as to be suitable for variable levels of digital readiness of different languages/language communities (see 2.2.) and for different types of user categories (see 2.1.). Through a mixture of educational material and guidelines for practical activities, the Training Programme wants to teach basic strategies to increase the presence of minority languages online. It will be structured along the following lines:

- help in overcoming intellectual barriers: explaining speakers why is it important for a language to be digital and motivating them to collaborate;
- help in creation of textual contents;
- help in creation of audio materials such as podcasts, web radio, YouTube channels;
- help in basic Social Media management: focusing on the relevance of Facebook pages and groups and Twitter accounts managed in minority languages for the creation of a social community;
- bringing others on our side: software and interfaces' localization projects;
- edutainment: ebooks, videogames, etc.

2.1. Tailoring the Training Programme to different media users

Taking into account media user typology elaborated by Brandtzæg (2010), we plan to address in a different manner the following categories of users:

- Entertainment users: they use Internet radio or TV, listen music, chat, and play online games. They are willing to provide content and produce materials for minority language but in a playful way (e.g. *game-with-a-purpose*). They have a lively social life and are good at promoting minority languages uses among groups of peers on social media;
- Instrumental users: interested in goal-oriented activities such as searching for information about goods or net-banking, e-commerce. When speakers of a minority language they can act as beta testers of translated interfaces;
- Advanced users: they display a very varied and broad Internet behaviour but mainly instrumental activities. They can be representatives of speakers communities promoting local activities, sharing their skills and thus becoming drivers of a digital expansion of their language.

2.2. Tailoring the Training Programme to different situations: the Digital Language Vitality Scale

Despite being a general problem affecting every regional and minority language, poor digital representation is obviously not the same for all of them. Similarly, the extent to which different languages can be used over digital media and devices (i.e., their *digital usability*) varies from language to language: on the one hand there are languages such as Karelian that appear to be hardly used on the Internet; on the other, there are languages such as Basque, Catalan, or Breton, for which digital use is stronger and more widespread. A training programme must take this variability into account, in order to deliver appropriate measures for the different conditions and needs of languages with respect to their digital usability. Therefore, it was decided to develop a tool for measuring the degree of *digital vitality* of languages, which in turn is defined as the extent to which a language is present, used and usable over the Internet through digital devices (PCs as well as mobile phones, smartphones, tablets, satellite navigators, Internet TV, etc.). The Digital Language Vitality measuring tool being developed by the DLDP project consists of a graded scale and a set of associated indicators. The Digital Language Vitality Scale is graded from 1 to 7, with 1 representing the

[2] Detailed information about the funding programme and the DLDP Consortium is available from the project website: `http://www.dldp.eu`

'pre-digital' level and 7 characterising a 'digitally thriving' language, one for which most if not all current digital uses are possible. The scale is inspired to ethnolinguistic vitality assessment (such as GIDS, (Fishman, 2001)), updated by (Lewis and Simons, 2010) as EGIDS, and the UNESCO "nine factors" (Brenzinger et al., 2003)), and is based on previous work in this area such as (Kornai, 2013) and (Gibson, 2015)[3]. The indicators associated with the scale are proxies representing both digital representation (presence) of a language and digital use. They are clustered into three groups: a first group of indicators refers to *digital usability* of a language, for instance, the existence of Internet connection or the availability of standardised fonts for writing the language. A second group of indicators is related to the *quality and amount of digital use* of a language: if and how much a language is used for texting and emailing, on websites, blogs, if there are e-books, Wikipedias, if the language is used on social media. The last group of indicators correlates with the *digital prestige* of a language, and are a sign of a language that not only is indeed used on digital media and devices, but it is so in a full-fledged way, enjoying the widest possible ranges of uses and applications (e.g. localised digital services, machine translation, edutainment products and services).

3. Assessing current digital use and usability of EU regional and minority languages

During the time frame of the DLDP, the Digital Language Diversity Scale measuring tool will be applied to a limited number of case studies, representing very different degrees of digital language representation and use. Four EU regional/minority languages will be investigated in detail so as to precisely assess their position on the Digital Language Vitality Scale: Sardinian (`srd`), Karelian (`krl`), Basque (`eus`) and Breton (`bre`)[4]. The investigation will be performed by means of a survey that is currently being developed at the time of writing. The survey is developed on the basis of previous work carried out in the area of ethnolinguistic vitality, such as the ELDIA Barometer (Åkermark et al., 2013), and other inquiries addressing specifically digital use of languages and availability and usability of digital resources and media [5].

The DLDP survey consists of a general part collecting basic information on the informant (age, sex, proficiency level in the language, frequency of use, etc.). The second part is focused on gathering information about his/her personal digital use of the language and about any known digital resource and services that make use of the language. We decided to give preference to questions that could give us information not easily retrievable in other ways. For instance, we deliberately left out questions addressing the existence of localised services or interfaces in the particular language, since this information is easily available and would make the questionnaire unnecessarily long. It is planned that the survey will be circulated from Spring 2016. In addition to the data collected and analysed for the four languages representing the case study, we are encouraging wide adoption and dissemination of the survey to EU regional and minority languages beyond the four investigated. We are planning to do this with the help of our extensive network of contacts (see section 5. below): Advisors and Twin projects will be encouraged to uptake the survey and disseminate it in their respective networks if interested in collecting data about other languages. In this way, we aim at making the DLDP project a hub of data concerning digital use and usability of European regional and minority languages.

4. Sustaining digital language vitality: the Digital Language Survival Kits and the Roadmap

The assessment tools and self-educational materials described in the previous sections will be instrumental to the development of a sustainable tool to help regional and minority language communities support digital representation of their languages, by setting the appropriate actions and measures for improving their language digital language vitality level. This instrument - named "Digital Language Survival Kits" - is conceived as a set of "emergency packs" indicating the actions to be undertaken for improving the digital language vitality level, but also which are the challenges and difficulties, which areas need to be addressed first, which tools are available. The Digital Language Survival Kits will thus complement and support the content provided by the Training Programme.

The Kits can be conceived as actionable guidelines (as the emergency metaphor intends to suggest) for regional and minority language speakers and communities in order to identify current gaps and areas where action can and needs to be taken, and learn about concrete actions and initiatives that can be put in place depending on the particular digital vitality level identified. As such, the two tools - the Digital Language Survival Kits and the Digital Language Vitality Scale (cf. 2.2.)- are respectively the diagnostic and therapeutic phases of the same intervention measure. For instance, a minimal degree of digital vitality will require a level of "digital survival capacity": to ensure connectivity, to develop and adopt a standardized encoding, to develop a standardized orthography, some basic language resources (at least a corpus, a spell checker, and a lexicon). Higher levels of digital vitality instead 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, and developing advanced language resources and tools (e.g. a Wordnet, multilingual corpora, or MT applications).

In the framework of the DLDP 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.

Finally, DLDP will deliver a number of recommendations

[3] The Digital Language Vitality Scale is a work still in progress. Full details will be made available from the project website and subsequent dedicated publications.

[4] Language codes follow the ISO 639-3 standard

[5] One important model in this respect was the survey recently carried out by Wikimedia France: https://docs.google.com/forms/d/1QJ4DQ6RvU0b-Fqpgz1KiooYOME1uz9MKsfKPdUg4VhM/viewform?c=0&w=1

specifically addressed at language stakeholders and policy makers, the *Roadmap for Digital Language Diversity*. Its aim 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 intention here is 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 under-representation of some languages and create strong pressure on local policies in member countries.

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

The Roadmap is intended to complement other previous and ongoing initiatives, such as the NPLD European Roadmap for Linguistic Diversity[6], the META-NET Strategic Agenda[7], and the FLaReNet Blueprint for Actions and Infrastructures[8]. Its innovative character lies in its specific focus on the particular needs and challenges of regional and minority languages.

5. Networking for creating a community

DLDP is strongly committed to creating a community that values and supports the notion of digital linguistic diversity. To this end, right from the beginning we devoted strong efforts to create a broad network of professionals, researchers, digital activists who are variously involved and committed to enlarging digital linguistic diversity.

An Advisory Board [9] has been set up to bring together the most notable and/or active personalities in the field, with the twofold goal of getting advice and suggestions on the activities and goals of the project and also of enlarging the dissemination possibilities of the project outcomes and message.

DLDP Advisors are distinguished scholars in the field of digital language revitalization, digital language activists, NLP professionals and policy makers. We are striving as well to provide broad geographical coverage, so as to be able to bring to DLDP the experience and point of view of a wide variety of initiatives. To date, the Advisory Board is composed by 22 members, the majority of which representing Europe, Africa, Latin America and North America. In parallel with the structure of the Advisory Board, the DLDP is actively establishing liaisons and partnerships with projects and initiatives that share the same commitments and are inspired by similar aims. Currently, the DLDP partners are *Lenguas Indígenas*, a network of digital activists in Latin America[10]; *Conradh na Gaeilge*, an organisation promoting the use of Irish in every aspect of life in Ireland[11]; *Indigenous Tweets*, Kevin Scannell's website that records minority language Twitter messages and users[12]; *CIDLeS - Centro Interdisciplinar de Documentaçao Linguística e Social*, a non-profit institution involved in the documentation, study and dissemination of European endangered and minority languages[13]; *UTI - Uned Technolegau Iaith*, a self-funded research unit that develops language resources for the Welsh language, the Celtic languages, and for multilingual situations in general[14]; and *Anveatsă Armãneashti*, a project for the preservation and promotion of the Aromanian language[15].

6. Conclusions

Europe's linguistic diversity is a unique heritage that deserves effective measures to ensure its safeguard and promotion. Since the digital world has become over time an important context where languages are used, it cannot be ignored any longer by any sustainable policy for protecting language diversity. However, the digital world is much less linguistically diverse that the non virtual one, as only a tiny fraction of the world's languages (about 6%, according to estimates) has access to 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, are bound to resort to major languages when living their digital lives, and this has dramatic consequences on the prestige of minority languages: perception of their marginal role and limited applicability is reinforced, and their attractiveness is diminished. The young generation, representing the speakers of tomorrow and those who should pass these languages on to new generations, can be only but convinced of the uselessness of minority languages for modern life. 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. From this point of view, DLDP fully embraces a bottom-up approach to language revitalization by addressing the speakers' cognitive and practical skills as the cornerstone of effective revitalization initiatives. Creation of content will increase the digital representation of these languages, and allow them to be first-class citizen of the language data economy, thus creating 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.

Acknowledgements

The DLDP is a project funded under the Erasmus+ programme (Grant Agreement no. 2015-1-IT02-KA204-015090). The opinions expressed reflect only the authors'

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

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

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

[9] http://www.dldp.eu/content/advisors

[10] https://rising.globalvoices.org/lenguas

[11] https://cnag.ie

[12] http://indigenoustweets.com/

[13] http://www.cidles.eu

[14] http://techiaith.bangor.ac.uk/

[15] http://anveatsaarmaneashti.com

view and the Erasmus+ National Agency and the Commission are not responsible for any use that may be made of the information contained.

References

Åkermark, S. S., Laakso, J., Sarhimaa, A., Toivanen, R., Kühhirt, E., and Djerf, K. (2013). ELDIA eulavibar toolkit: Practical guide to the EuLaViBar tool, with reference to the ELDIA comparative report. Permalink: http://phaidra.univie.ac.at/o:301101.

Brandtzæg, P. B. (2010). Towards a unified media-user typology (MUT): A meta-analysis and review of the research literature on media-user typologies. *Comput. Hum. Behav.*, 26(5):940–956, September.

Brenzinger, M., Yamamoto, A., Aikawa, N., Koundiouba, D., Minasyan, A., Dwyer, A., Grinevald, C., Krauss, M., Miyaoka, O., Sakiyama, O., Smeets, R., and Zepeda, O. (2003). Language vitality and endangerment. Ad Hoc Expert Group Meeting on Endangered Languages, March.

Crystal, D. (2010). *Language Death*. Cambridge University Press.

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

Fishman, J. A. , editor. (2001). *Can threatened languages be saved?* Multilingual Matters.

Gibson, M. (2015). A Framework for Measuring the Presence of Minority Languages in Cyberspace. In *Proceedings of 3rd International Conference on Linguistic and Cultural Diversity in Cyberspace*, pages 61–70.

Kornai, A. (2013). Digital language death. *PLoS ONE*, 8(10).

Lewis, M. P. and Simons, G. F. (2010). Assessing endangerment: Expanding fishmans gids. *Revue Roumaine de linguistique*, 2(55):103–120.

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

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

Rehm, G. et al., editors. (2012). *META-NET White Paper Series: Europes Languages in the Digital Age*. Springer.

Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M., Mermer, C., Varadi, T., Kirchmeier-Andersen, S., Stickel, G., Jones, M. P., Oeter, S., and Gramstad, S. (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.

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