# The DLDP Survey on Digital Use and Usability of EU Regional and Minority Languages

**Claudia Soria, Valeria Quochi, Irene Russo**
CNR-ILC
Pisa, Italy
{firstname.lastname}@ilc.cnr.it

### Abstract

The digital development of regional and minority languages requires careful planning to be effective and should be preceded by the identification of the current and actual extent to which those languages are used digitally, the type and frequency of their digital use, the opportunity for their use, and the main obstacles currently preventing it. This paper reports about the design, the results and the key findings of an exploratory survey launched by the Digital Language Diversity Project about the digital use and usability of regional and minority languages on digital media and devices. The aim of the survey - the first of this kind - was to investigate the real usage, needs and expectations of European minority language speakers regarding digital opportunities, with a strong focus on electronic communication. The survey is restricted to four languages (Basque, Breton, Karelian and Sardinian) at different stages of digital development, which offers a starting point to develop strategies for assessing digital vitality of these languages and overcoming specific difficulties such as, for instance, the lack of official data.

**Keywords:** minority languages, digital survival, electronic communication

## 1. Background and Motivation

In this paper we present the results of the first survey about the actual usage of four European minority languages and the related needs of their speakers in terms of digital opportunities. The survey is part of the work carried out by the Digital Language Diversity Project (DLDP) (Soria et al., 2016)[1], a three-year Erasmus+ project started in September 2015.

The goal of the DLDP is to help minority language speakers' communities in the acquisition of intellectual and practical skills to create, share, and reuse online digital content in their languages. At the same time we want to define general guidelines and best practices for the promotion of minority languages with poor digital representation, a fact that further prevents their usability on digital media and devices.

One of the underlying assumptions of the Digital Language Diversity Project is that the sustainability and preservation of regional and minority languages is closely tied to their being perceived by their speakers as being fully-fledged languages that can be used in any context, the digital one included. Unfortunately, this is far from being a reality not only for regional and minority languages, but for the majority of the world languages. In most cases, the technical or infrastructural impediments for the digital use of European regional and minority languages are modest and fairly easily solvable. Marginalisation and minoritisation of those languages mostly derives from the concurrency of the national and global languages for which digital content and services are more easily available, which further discourages regional and minority language speakers from using those languages digitally. In order to break this vicious circle and make those languages digitally appealing and usable to an extent that can compete with other major languages, it is necessary to approach the problem in terms of "digital language planning".

In order to be able to plan for digital development, we first need to identify the current and actual extent to which RML are used digitally, the type and frequency of their digital use, the opportunity for their use, and the main obstacles currently preventing it so as to get a clear understanding of the different factors that may affect the digital use of RMLs. Some reports carried out for individual languages and specific media are available, like the Language White Papers (Uszkoreit and Rehm, 2012) published by the META-NET Network that has clearly shown how 30 European languages are at risk of digital extinction because of lack of sufficient support in terms of language technologies. The META-NET work, initially for each of the EU official languages, was then extended to cover as well some regional languages such as Basque (HernÃ¡ez et al., 2012), Catalan (Moreno et al., 2012), Galician (GarcÃa-Mateo and Arza, 2012), and Welsh (Evas, 2013). The reports mostly assessed the status of those languages in terms of language technology support. A general survey covering all regional and minority languages of the EU, the different types of digital media and services available, as well as inquiring about the *attitudes* and *desires* to make a digital use of the language is still lacking. The DLDP effort can therefore be seen as a first step towards the design of a survey about digital use of minority languages in both professional and informal contexts, specifically tailored on RMLs in the digital world and structured around a crucial question: is it possible for regional or minority language speakers to have a digital life in those languages? The paper is organised as follows: a description of the methodology underlying the design of the survey; an analysis of the results collected, with a separate section for each language; a summary of the key findings and an indication of the work planned for the future.

[1] http://www.dldp.eu

## 2. Survey Design and Methodology

The focus of the DLDP survey is on four European minority languages at different stages of digital development (Basque, Breton, Karelian and Sardinian). This allows us to explore and compare: (a) the behavior, perception and actual desires of different speakers' communities relative to the digital online use of their native languages; (b) the extent of the availability of digital content and language technology in 'digitally-different' languages; and (c) speakers' awareness of the latter.
All this has the final objective of devising and suggesting ad-hoc strategies for promoting the active digital usage of these languages and the development of language-based digital applications, which will substantially help in the preservation or revitalization of RMLs.

The questionnaire was developed around a set of topics, among which:
**RML knowledge**: perceived degree of fluency in the language inquired and values attributed to the language, whether mainly affective, identitarian, or instrumental;
**Activism**: whether the respondent qualifies as a language activist[2];
**Extent and frequency of use of the language**: is the language mainly used in oral or written form? In informal contexts only or in institutional ones as well? How often is the language used in those contexts?
**E-communication**: is the language ever used for writing e-mails, texting, chatting or other instant messaging? If yes, how often? If not, why?
**Digital use**: is the language ever used for surfing the Net, reading, writing/ commenting blogs, forums or websites? If yes, how? Only passively or actively as well? If not, why?
**Technological support**: if a specific keyboard is needed to type in the language, is it available?
**Digital media**: are digital media such as websites, blogs, Internet TV, audio and video streaming, e-books, etc. available in the language?
**Wikipedia**: is a Wikipedia available in the language and if yes, is it read and/or edited?
**Social media**: how often is the language used on social media (Facebook, Twitter, Instagram etc.)?
**Localised software and interfaces**: are operating systems, software and social media interfaces localised in the language? If yes, is the localised version used?
**Language resources**: are language resources and tools (such as online dictionaries, spell checkers, machine translation interfaces) available for the language?

This list of topics was further developed into a set of questions that were evaluated and discussed with the DLDP Advisors, distinguished scholars in the field of digital language revitalization, digital language activists, NLP professionals and policy makers[3]. DLDP Advisory Board brings 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.

### 2.1. Methodology

A basic questionnaire was developed in English and was then translated into Basque, Breton, Karelian and Sardinian, and made available as a Google form.[4]

The survey was carried out between July and September 2016, exclusively online. Informants were mostly recruited via the language associations involved in the projects and institutional contacts. The survey was also advertised on social media and through personal contacts. The profile of the typical respondent thus belongs to a person who makes at least a minimal digital use of the language in his/her everyday life, a fact which has to be taken into account when reading the results.
The addressees of the survey were individuals who reported to be speakers of one of the four case-study languages of the project and who are digitally literate.
As our main purpose was to understand the extent to which a language was used on the Internet and over a number of digital media and services, we were satisfied with digitally literate speakers only, as the reasons for not using the language digitally for those with no/low digital skills are evident.
We received a total of 1.301 replies, 749 from men, and 538 from women (breakdowns by age groups and sex are shown in Table 1). We are aware of the fact that the population inquired does not represent a balanced sample, which would have required the availability of data about the composition of a minority language population in terms of age, gender and other demographic variables.
Finally, the responses were normalised and analysed with the help of native speakers of each target language involved in the project.
In the next section, we describe the data collected for the four case studies of the DLDP project.

## 3. Survey Results

### 3.1. Basque: a digitally fit language asking for more opportunities

Basque is an isolate, non-Indo-European language spoken by about 900.000 speakers[5] living mainly in the Basque country and surrounding geographical areas situated in the northern part of Spain and the neighbouring south-western part of France. It has official status in the Spanish Basque Autonomous Community, but not in France.

[2] Language activists tend to be intentionally more assertive in their use of the language and, as a consequence, they can't represent average speakers.

[3] http://www.dldp.eu/en/content/advisors

[4] For Karelian, coherently with the co-existence of three recognised varieties and considering the possibility that speakers of one might not be able or willing to participate if the text was written in another variety, the questionnaire was translated into Olonets (or Livvi Karelian), South Karelian and North Karelian. However, for the analysis and the presentation of the results, the responses have been normalised to North Karelian, the variety most supported by the DLDP partner, KKS.

[5] source: NPLD.eu

| Age Group | Gender | Basque | Breton | Karelian | Sardinian |
|---|---|---|---|---|---|
| <20 | Female | 9 | 9 | 0 | 10 |
| | Male | 11 | 5 | 1 | 6 |
| | N/A | 1 | 0 | 0 | 0 |
| | *All* | *21* | *14* | *1* | *16* |
| 20-29 | Female | 37 | 17 | 5 | 24 |
| | Male | 35 | 18 | 6 | 49 |
| | N/A | 2 | 1 | 1 | 1 |
| | *All* | *74* | *36* | *12* | *74* |
| 30-39 | Female | 59 | 21 | 6 | 38 |
| | Male | 67 | 23 | 3 | 67 |
| | N/A | 1 | 1 | 1 | 1 |
| | *All* | *127* | *45* | *10* | *106* |
| 40-49 | Female | 54 | 13 | 17 | 46 |
| | Male | 55 | 26 | 12 | 72 |
| | N/A | 0 | 1 | 0 | 1 |
| | *All* | *109* | *40* | *29* | *119* |
| 50-59 | Female | 36 | 10 | 16 | 44 |
| | Male | 49 | 29 | 15 | 65 |
| | N/A | 1 | 0 | 0 | 0 |
| | *All* | *86* | *39* | *31* | *109* |
| 60-69 | Female | 2 | 5 | 34 | 15 |
| | Male | 7 | 19 | 24 | 54 |
| | N/A | 0 | 0 | 1 | 0 |
| | *All* | 9 | *24* | *59* | *69* |
| >70 | Female | 0 | 2 | 3 | 6 |
| | Male | 1 | 2 | 11 | 17 |
| | N/A | 0 | 0 | 0 | 0 |
| | *All* | *1* | *4* | *14* | *23* |
| *All* | | *427* | *202* | *156* | *516* |

Table 1: Number of respondents with breakdown by age groups and sex

From the responses to the questionnaire, Basque emerges as a language that is regularly and actively used for e-communication, on the Internet and on social media by 97.2% of respondents, in particular for chatting and other forms of instant messaging, but also for e-mail. This is in line with the political status of the language, officially used in public administration and schools.

Even if no particular technological barrier is reported to impede the use of the language and localised digital services and interfaces are available for the Basque language, more effort is needed to sustain the language at the professional and institutional level, and more entertainment services and products are required in Basque targeted to young people.

For example, despite of knowing about the existence of localised operating systems and interfaces, some of the respondents are not using Basque in their devices, applications or softwares. A third part of the respondents feel that using Spanish tools is easier also because the way of finding and installing software in Basque is not as easy as it is in other languages, and, as a result of that, the user has to do an extra effort.

Some of the respondents are claiming for a site, where all available resources in Basque are listed and ready to be downloaded by users. In addition to that, it has been mentioned a need of spreading the information about the existing services/interfaces in Basque, especially among the young people. A more detailed description of the results is given in (Hernaiz and Berger, 2017).

### 3.2. Breton: strong awareness of the importance of digital presence

Breton is a Celtic language of France spoken mainly in western Brittany by around 200.000 speakers[6], and the number is said to be decreasing. Hopefully this trend will change thanks to the growing inclusion of the language in educational contexts. However, at the time of writing, Breton still has neither official status nor specific protection in France.

Overall, the survey reveals a strong desire of Breton speakers to use their language digitally, and their awareness of the importance of Internet presence for its revitalization. Nevertheless, although almost all respondents say they use Breton on the Internet, the seem to do so mostly passively (e.g. for reading rather than for writing).

For e-communication instead active use is widespread, for e-mail in particular, where the language appears to be used regularly by 88.5% of respondents. The same holds for active use of Breton on social media, with Facebook being the most used network, perhaps thanks to the localised interface available.

[6] source: NPLD.eu

Most of the respondents are aware of the existence of a Wikipedia in Breton, with a 19% of them even contributing to it by editing existing articles or writing new ones (8%).

While the digital basics are firmly in place, the relative lack (or lack of awareness) of advanced services, apps and localised software stands out. At the same time, the respondents manifest a strong desire in this direction. For instance, automatic translation is almost completely lacking except for the online translation of Breton to French offered by *Ofis ar Brezhoneg*[7]; Google Translate is not available for Breton, yet. Indeed, if popular apps and key software interfaces are not provided in Breton soon, unable to compete with French apps, the language will inevitably appear less appealing to the younger generations. A more detailed description of the survey results can be found in (Hicks, 2017).

### 3.3. Karelian: motivated speakers want to be visible online

Karelian, a Uralic language of the Finnic branch, is a non-territorial language spoken in Russia (the Republic of Karelia and Tver oblast) and Finland. Two main dialects are distinguished: Olonets Karelian (also called Livvi) and Karelian proper, the latter of which is divided further into North Karelian (also called Viena Karelian) and South Karelian. The latest estimate suggest that in Finland around 11,000 people can speak Karelian well to fluently, with another 20,000 have some knowledge of the language. The language has official status in both countries. The digital portrait of the Karelian language reveals that Karelian speakers have a high linguistic self-esteem and are willing to use their language online, a crucial issue for the revitalization of a non-territorial language. However, there are a lot of necessary online and digital resources missing, and speakers often lack information about those that are available. For example, the existence of online dictionaries and of a Karelian Wikipedia [8] could benefit many Karelian speakers and the community if only they were aware of the existence of such resources.

In terms of social networks the use of Karelian is very heavily restricted to Facebook that anyway has users from all demographic groups and it is very well suited for extensive written communication, making it easy to connect with Karelians from different parts of the country and even across the border. The situation is not likely to change considerably anytime soon, although the amount of people using Karelian on Facebook might increase and there might be an increased Karelian presence on Twitter and Instagram, if more young people get interested in the language and start using it and if the Karelian revitalization efforts can get a better visibility in Finnish political discussion.

Among the reasons people are not using Karelian online we find that they don't know how to change their keyboard settings for this purpose. This is an important point to communicate to the community, since if writing Karelian is perceived as something that is difficult, it can easily make people less eager of using it. The interested reader can find additional information in (Salonen, 2017).

### 3.4. Sardinian: a digital language without self-awareness

Sardinian is a Romance language spoken by more than one million speakers in the island of Sardinia, Italy. The language is officially recognised but despite this, its presence in education and media is very limited. The responses to the questionnaire about Sardinian show high consideration of the language by its native speakers and a strong desire to use it in their everyday digital life. However, we observe a strong imbalance between the use of the language on social media and for e-communication on the one hand, and on the other the availability or awareness of other types of media and services in Sardinian. The language appears to be frequently used online both actively - in particular for texting, chatting, interacting on social media and blogs - and passively - for surfing and reading the Internet -. As for social media, it appears to be particularly vital on Facebook, by large the most used network, which even has a localised interface available. Between 73% and 87% of the respondents claim to use Sardinian digitally everyday but mostly in informal, private-life contexts. This fact which shows that, despite the official recognition granted in 1998, the language is still relatively little used on public sites and in official, formal contexts.

On the side of Internet media and services, we observe a different situation. Speakers tend to be unaware or not to use media and services in Sardinian even when available. For instance, almost half of the respondents is not aware that a Wikipedia is available and active for Sardinian, while almost the other half is only a passive user of it and many prefer the Italian version because more informative. Also, while online newspapers and news are relatively widely available, as is entertainment and - thanks to a previous investment by the Region - some Public Administration services, more advanced media such as smartphone apps, Internet TV, audio and video streaming are still missing. An interesting observation is that, despite there are no technical obstacles for using Sardinian online (i.e. typing is possible using the standard keyboard), many respondents nevertheless lament a lack of competence in written Sardinian that prevents them to write it with the necessary confidence. A more detailed description of the survey results can be found in (Russo and Soria, 2017).

Overall, the survey results clearly show the importance of encouraging the speaker community about using their language online as much as possible. The existence of a considerable number of language resources such as dictionaries, spell checkers, and even an automatic translation system is a good sign of the potential for this language to become a fully digital language, provided that the speakers are supported and encouraged to overcome the psychological barriers that are yet holding them back from considering Sardinian as a language in its full rights.

## 4. Key findings

From the exploratory study carried out, it appears that the four languages investigated place themselves at very dif-

[7] `http://www.fr.brezhoneg.bzh/42-traducteur-automatique.htm`

[8] The Karelian wikipedia, now online, was not publicly accessible at the time of the survey.

ferent levels of digital use and usability. However, a few themes emerge consistently:

- Regional and minority language speakers have a strong desire to use their languages digitally, in all the sociolinguistic domains and for all the purposes where major languages are used
- Social media has a huge potential as a domain that drives language revitalisation, but this sociolinguistic space is still restricted
- There is a clear demand for increased regional language usage in the public domain
- Minoritised language speakers need to be supported and encouraged regarding their ability to use their languages digitally and for their importance as digital content creators
- The lack of structural support for these languages is a serious issue that needs to be addressed. The digital development of these languages is not sustainable when it has to rely on the work of a handful of activists and volunteers

Table 5. presents a synoptic overview of the main findings regarding use of the languages over the main media and services addressed by the survey. Besides giving an approximate idea of the range, amount and frequency of digital uses, the survey also opens a window about the reasons why a language is not used. These are briefly summarised in Table **??**. These obstacles and limitations fall into three main categories:

1. technological barriers, such as the unavailability of a specific keyboard or spell checkers that would ease the writing;
2. linguistic barriers: lack of competence in the written language is often seen as a main problem that restrains people from using their language in written form. For a language such as Sardinian, where a standard is available but not yet widespread, this is a strongly felt issue;
3. psychological barriers: fear of being misunderstood, teased or of looking offensive. This category also comprises

## 5. Further work

To the best of our knowledge, this survey is the first of this kind and in theory it is applicable to any language with minor adaptations. Since one of the aims of the DLDP is to keep collecting data about digital language diversity, we encourage the broad community to adopt, adapt and localise the survey to gather data about other languages, and we are willing to offer support in this direction. To this end we make the model for the survey available upon request under a CC-BY-4.0 license.

For those who are interested in the original, raw data of the survey, they are deposited in the ILC4CLARIN Repository and made available under a CC-BY 4.0 license (Soria et al., 2017)[9].

Finally, the DLDP team is currently finalising a Digital Language Vitality Scale with measurable indicators for practically assessing the level of digital development of any language, with the questionnaire being a possible source of information, especially when official data are inexistent or not available. In the near future these two instruments will be used to better assess the four languages and derive four complete case studies that will hopefully inspire other similar exercises for many other RMLs.

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

We thank our project partners, Klara Ceberio Berger, Antton Gurrutxaga Hernaiz, Davyth Hicks and Tuomo Salonen for their precious contribution to the set up and dissemination of the survey campaign first and then for the help in the analysis of the data. We are also grateful to all our advisors who offered their comments and suggestions, which helped us greatly improve our study. Any liability for the content of this paper, is of course the authors' alone.

## Bibliographical References

Evas, J. (2013). *Y Gymraeg yn yr Oes Ddigidol – The Welsh Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). META-NET. Available online at `http://www.meta-net.eu/whitepapers`.

GarcÃa-Mateo, C. and Arza, M. (2012). *O idioma galego na era dixital – The Galician Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Hernaiz, A. G. and Berger, K. C. (2017). Basque – a digital language? Reports on Digital Language Diversity in Europe. Work carried out within the Digital Language Diversity Project (www.dldp.eu), EC Erasmus+ Programme (2015-1-IT02-KA204-015090).

HernÃ¡ez, I., Navas, E., Odriozola, I., Sarasola, K., Diaz de Ilarraza, A., Leturia, I., Diaz de Lezana, A., Oihartzabal, B., and Salaberria, J. (2012). *Euskara Aro Digitalean – The Basque Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Hicks, D. (2017). Breton – a digital language? Reports on Digital Language Diversity in Europe. Digital Language Diversity Project (DLDP). Work carried out within the Digital Language Diversity Project (www.dldp.eu), EC Erasmus+ Programme (2015-1-IT02-KA204-015090).

[9] `http://hdl.handle.net/20.500.11752/ILC-77`

| Dimension | Basque | Breton | Karelian | Sardinian |
|---|---|---|---|---|
| e-communication | 95.8% | 88.5% | 56.4% | 71.8% |
| Available media (As reported by at least 50% of respondents) | All but Internet TV | Websites, streaming audio, Internet TV, streaming video, blogs and forums | Websites | Websites, blogs and forums |
| Wikipedia use | 81.2% | 68.3% | - | 43.7% |
| Social Media use (Active use by more that 50% of respondents) | Facebook, Twitter | - (Facebook use not reaching treshold) | - (Facebook use not reaching treshold) | Facebook |
| Available digital services (As reported by at least 50% of respondents) | All but e-commerce, advertising | Online newspapers, Entertainment, online news and search engines | Online newspapers, online news | Online news |

Table 2: Overview of key findings about digital use

Moreno, A., Bel, N., Revilla, E., Garcia, E., and Valverdú, S. (2012). *La llengua catalana a l'era digital – The Catalan Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at http://www.meta-net.eu/whitepapers.

Russo, I. and Soria, C. (2017). Sardinian – a digital language? Reports on Digital Language Diversity in Europe. Digital Language Diversity Project (DLDP). Work carried out within the Digital Language Diversity Project (www.dldp.eu), EC Erasmus+ Programme (2015-1-IT02-KA204-015090).

Salonen, T. (2017). Karelian – a digital language? Reports on Digital Language Diversity in Europe. Digital Language Diversity Project (DLDP). Work carried out within the Digital Language Diversity Project (www.dldp.eu), EC Erasmus+ Programme (2015-1-IT02-KA204-015090).

Soria, C., Russo, I., Quochi, V., Hicks, D., Gurrutxaga, A., Sarhimaa, A., and Tuomisto, M. (2016). Fostering digital representation of eu regional and minority languages: the digital language diversity project. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'16)*, pages 3256–3260, Portoroz, Slovenia, May. European Language Resource Association (ELRA).

Hans Uszkoreit et al., editors. (2012). *METANET White Paper Series*. Springer.

## Language Resource References

Claudia Soria and Valeria Quochi and Irene Russo and Anneli Sarhimaa and Eleonore Kruse and Davith Hicks and Tuomo Salonen and Antton Gurrutxaga Hernaiz and Klara Ceberio Berger. (2017). *Digital Language Diversity Project Survey Data*. Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche (ILC-CNR), distributed via CLARIN, 1.0, ISLRN http://hdl.handle.net/20.500.11752/ILC-77.