WHAT IS DIGITAL LANGUAGE DIVERSITY AND WHY SHOULD WE CARE?

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

Introduction

The relationship between language and the Internet is a growing area of policy interest and academic study, see for instance (MAAYA 2012), (Paolillo et al. 2005), (Pimienta 2001), (Kornai 2013), (Pimienta et al. 2009), (Rehm and Uszkoreit 2012).

The emerging picture is one where language profoundly affects a person's experience of the Internet. It determines how much – if any – information you can access on Wikipedia. It orients a person's choices and decisions by shaping the results of a search engine, depending on the language used. It determines the range of services that can be available over the Internet, and therefore the amount of everyday tasks (such as buying a ticket, reviewing opinions about hotel and restaurants, purchasing books or other goods, etc.) that can be carried out virtually. Far from infinite, the Internet, it seems, is only as big as one's language.

Should this hold true, it would be at odds with the original spirit of the Internet, which - according to the words of Tim Berners-Lee - would be a place "to cross barriers and connect cultures".

But it is safe to argue that the extent to which a language can be used over the Internet not only affects a person's experience and choice of opportunities; it also affect the language itself.

If a language is poorly or not supported to be used over digital devices, for instance if the keyboard of the PC is not equipped with the characters and diacritics necessary to write in the language, or if there is no spell checker for a language, then its usability becomes severely affected, and it might will never be used online. The language could become "digitally endangered", and its value and profile could be lessened, especially in the eyes of the new generations.

These considerations call for closer examination of a number of related issues. First, the "digital language diversity", i.e. the linguistic diversity of the Internet. Second, it is important to reflect on the conditions that make it possible for a language to be used over digital devices, and about what can be done in order to grant this possibility to languages other than so-called "major" ones.

Linguistic Diversity

According to linguists there are between 6.000 and 7.000 spoken languages (Lewis et al. 2013), and perhaps as many sign languages. The impressive language diversity of the world is reported to concentrate in some areas more than in others: for instance, Papua New Guinea (home to 830 languages over 400.000 km2), Indonesia (722 languages for 240M people), Nigeria (more than 500 languages), India (22 official languages, 400 languages, more than 4000 dialects). These areas of incredible concentration of different languages are called *language hotspots*: regions having not only the highest levels of linguistic diversity, but also the highest levels of endangerment, and often the least-studied languages (Harrison 2010a). The highest linguistic diversity tends to be located in areas of lesser economic development, that have endured little or no globalisation, have relatively well preserved the natural environment. This has been interpreted as a signal of the fact that linguistic diversity represents the normal or natural condition, while the monolingualism frequently observed especially in the Western countries is an artifact or a side effect of socio-political forces.

Although only recently, there is a growth of scholarly and public discourse about the value of linguistic diversity. The arguments in favor can be classified either as aesthetic, cognitive, anthropological or ecological.

From an aesthetic point of view, languages can be seen as living monuments of the peculiarly human way of forming societies, of communicating and transmitting experience.

The cognitive strand of argumentation maintains that from the point of view of the study of the human species, the variety of languages and its related variability of forms and structures (sound systems, syntactic patterns and morphological constructions) offer a unique view on the functioning of the human brain and the humanly peculiar language faculty.

In anthropological terms, language diversity is interpreted as one of the many responses of the human species to the extreme variability of its surrounding environment: the variety of the way in which human beings have adapted and responded to the various climates and challenges is uniquely embodied in languages. Along this argument, some authors argue that linguistic diversity embodies the resilience weaponry of the human species against the environment, by codifying the knowledge about surviving techniques, plants, animals, crops, preparation and use of medicinal food, as well as traditional methods of farming, fishing, and hunting, of land use and resource management. This enormous wealth of knowledge that was accumulated over the centuries may turn useful again and we cannot afford to lose it by eradicating language diversity. David K. Harrison, a linguist and advocate of linguistic diversity, expresses this view in a very powerful way: "What hubris allows us, cocooned comfortably in our cyber-world, to think that we have nothing to learn from people who a generation ago were hunter-gatherers? What they know - which we've forgotten or never knew - may some day save us. We hear their voices, now muted, sharing knowledge in 7000 different ways of speaking. Let's listen while we still can." (Harrison 2010b).

Another argument in favor of linguistic diversity is the ecological one. Language diversity tends to correlate with biodiversity, they underpin and mutually reinforce one another: it appears that those places with high species diversity (tropical forests in particular) tend to show higher linguistic diversity, while areas low in species diversity, such as deserts and tundra, also show lower linguistic diversity (Loh and Harmon 2014, Nettle and Romaine 2000, Loh and Harmon 2005). Exactly as it happens for biodiversity, language diversity is threatened with regions where this loss is more acute and faster (Loh and Harmon 2014), (Harmon and Loh 2010). Both are facing an extinction crisis, and both crises are consequences of similar processes. According to Sutherland (Sutherland 2003), the loss of languages goes at a faster pace than the loss of species. The reasons behind the loss of linguistic diversity are mostly concerned with social or economic reasons (commerce, migration, globalization of trade and media, but also unfavorable national policies and the prestige associated with one or more dominant languages); more rarely they are associated with natural phenomena such as a population's extinction.

Regardless of the point of view wherefrom language diversity is approached, there is consensus about its severe endangerment: all the different strands of argumentation converge in advocating for sustaining language diversity, either as a collective commitment to the sustainability of our planet and of humankind or out of reverence and respect for the value of languages per se.

The Digital Language Divide

The digital world has become an important battlefield for protection of linguistic diversity. Digital media and tools represent only one of the various possible contexts of language 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 (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.

In order to establish a sustainable policy for safeguarding and promoting linguistic diversity, the digital world cannot be ignored any longer. In a world that is increasingly being dominated by ICT, no language can afford to miss the digital opportunity if it aspires to be a vital language. As Mark Turin aptly says, "in our digital age, the keyboard, screen and web will play a decisive role in shaping the future linguistic diversity of our species" (Turin 2013). Languages are living entities that need to be used on a daily basis by humans in order to survive. With so much of our lives happening on the Internet and through digital devices, the digital space represents a context that cannot be ignored. Speakers of major languages can access apparently unlimited amounts of Web content, easily perform searches, interact, communicate through social media and voice-based applications. They can enjoy interactive ebooks, have fun with word games for mobiles, engage in multi-player video-games, or take advantage from innovative language learning facilities for other widely spoken languages.

According to a 2013 survey (LTInnovate 2013), in 2012 digital content has grown to 2.837 zettabytes, up almost 50% from 2011, on its way to 8.5ZB by 2015. The community of social network users in Western Europe was set to reach 174.2 million people in 2013, which is about 62% of Internet users. A massive 800 million people are Facebook users, of which 170 are from highly linguistically diverse countries such as Brazil, India, Indonesia, and Mexico. The number of Twitter's active users is estimated around 200 millions. LinkedIn has 115 million users, and Google+ as many as 180 millions[1].

These numbers, as imperfect as they may be, give a flair of the depth and breadth of Internet. But what can be said about its linguistic diversity? How the enormity of Internet users behave, from a linguistic point of view? Which languages do they use? In other words, does the Internet reflect the linguistic diversity of the planet?

A study by W3Techs[2] shows that at the time of writing of this article, 55.9% of all content online is in English. Aside from English, Spanish and Portuguese, only five other EU languages (German, French, Italian, Polish and Dutch), out of 60 or more spoken in the Union, are published on more than 1% of the top million sites (LTInnovate 2013).

With reference to domain names, a majority of domains (78%) are registered in Europe or North America: a finding that reinforces the dominance of those two regions in terms of Internet content production. Asia, in contrast, is home to 13% of the world's domains while Latin America (4%), Oceania (3%), and the Middle East and Africa combined have even smaller shares of the world's websites (2%). Globally, there are about 10 Internet users for every registered domain. The United States is home to almost a third of all registered domains, and has about one website for every three Internet users.

From the Wikipedia point of view, Wikipedia articles in 44 language versions of the encyclopedia are highly unevenly distributed. Slightly more than half of the global total of 3,336,473 articles are about places, events and people roughly concentrated in the European area, occupying only about 2.5% of the world's land area: the majority of content produced in Wikipedia is about a relatively small part of our planet.

The Internet, therefore, appear to be far from being linguistically diverse. English is still the language most used over the Internet, the one for which more content is produced, and also the privileged tongue of the majority of its users. With a handful of languages dominating the web, there is a *linguistic divide* that parallels and reinforces the digital one. Exactly as there are areas of the world deprived of access to the Internet, there are entire languages that cannot get to the Internet. The consequences of such a digital language divide are severe.

Since only the speakers of some dominant languages can hope to access the Internet, its use and usability is dramatically affected. The amount of information and services

1 Source: Language Connect: www.languageconnect.net

2 http://w3techs.com/technologies/overview/content_language/all

that are available in less widely spoken languages is reduced, thus creating inequality at several different levels:

- inequality of linguistic rights and digital opportunities for all languages and all citizens.
- inequality of information and access to services;
- unequal access to technological development and unequal digital dignity;
- unequal opportunities for language survival.

Let's briefly review them in more detail.

Inequality of information and access to services: with 55.9% of all online content estimated in English, it is plain that only those who can read English can access the majority of the information available on the Internet. Machine Translation is a way to get hold of the content available in another language, yet Google Translator is available - with very different degrees of accuracy - for ninety languages only, of which 39 from Europe, 38 from Asia, 10 from Africa, 1 from the Americas and 1 from the Pacific region.

The largest and most linguistically diverse online encyclopedia, Wikipedia, is available in 290 languages, a fairly remarkable number. However, there are striking asymmetries in the amount of information available for the different language editions. The German Wikipedia, which is the fourth largest after the English one, has less than half the number of articles that are available for English[3]. On the other side of the spectrum, there is a near absence of any content in many African and Asian languages[4].

To use the Internet at its fullest means to get access to the whole array of available services such as social media, or reviews sites such as TripAdvisor, or marketplaces like Amazon, eBay, Etsy or Booking.com, to name just a few. But unless you are fluent in one dominant language, you will never be able to use these services: Facebook supports 147 languages[5], Twitter 32[6]. TripAdvisor is available in 29 languages[7], and Booking.com in 43.

Speakers of major languages can access apparently unlimited amounts of Web content, easily perform searches, interact, communicate through social media and voice-based applications. They can enjoy interactive ebooks, have fun with word games for mobiles, engage in multi-player video-games, or take advantage from innovative language learning facilities for other widely spoken languages.

So called "smaller" languages do not enjoy the same range of opportunities as more widely spoken languages. Occitan authors and publishers could not upload and sell ebooks in Occitan over Amazon's Kindle platform, because Occitan is not among the

3 German Wikipedia: 2.022.598 articles; English Wikipedia: 5.332.011.

4 www.zerogeography.net/2012/10/dominant-wikipedia-language-by-country.html

5 https://www.facebook.com/translations/FacebookLocales.xml

6 https://dev.twitter.com/web/overview/languages

7 https://developer-tripadvisor.com/content-api/supported-languages/

languages supported[8]. There is no Wikipedia for Mansi; speakers of Tongva have no localized interface for Facebook, and there is no Google translation for Sardinian, or Quechua, or Inupiaq[9].

In addition to unavailability of Internet services in some countries and to poor digital skills of large parts of the planet, the lack of support for languages other than the major ones implies that speakers of 94% of the languages spoken cannot access Internet services unless they are fluent in one major language as well.

Unequal access to technological development and unequal digital dignity: latest technological development embedded in current, everyday digital devices such as smartphones or tablets are not accessible to speakers of less-widely spoken languages. For instance, Apple's Siri, one of the latest voice-enabled smart personal assistants for smartphones, has been developed for 25 languages only. It covers nine out of 30 EU official languages; an Irish speaker cannot use Siri in her language and has to turn to English instead, and still, problems with an Irish accent can be experienced. This inequality of digital opportunities further discriminates less widely spoken languages, by relegating them once more to the realm of family communication and restricted topics. Less digitally represented languages are under the serious risk of being marginalized, and eventually dialectalized over the years. According to Carlos Leáñez (cited in Prado 2011), "the less valuable a language is [in the eyes of its speakers], the less it is used, and the less it is used, the more it loses value". Shrinking contexts of uses can have a devastating effect, eventually leading to the abandonment of a language in favor of another, better supported one. Should this happen, the consequences for a language profile would be dramatic: any language that cannot be used over digital contexts will engage in a "digital diglossia" relationship with another, better supported language.

Unequal opportunities for language survival: less and less digital contexts of use is what can bring languages to digital extinction (Rehm et al. 2012). It is common to associate the concept of extinction with very exotic languages, or those spoken by a restricted minority. However, the concept of "digital extinction" describes a condition that could prove true for many languages, even those far from being endangered outside the digital world. This condition holds whenever a language is used less and less over the Internet because of lack of Language Technology support: then the range of contexts where it is used dramatically collapses and gradually brings the language to disappear from the digital space. Where there is no favorable environment for a language over digital tools, then its use over the Internet and through digital devices becomes cumbersome, communication is difficult, and usability of the language is dramatically affected. By pushing the naturalistic metaphor further, we can think of a "digitally hostile environment": one where it is not possible to type, make searches, have translations, hold a conversation over digital devices. In such a context, a language easily language goes extinct.

According to the principles of the World Summit on the Information Society endorsed by the UN, the "Information Society should be founded on and stimulate

8 https://kdp.amazon.com/help?topicId=A9FDO0A3V0119

9 http://translate.google.com/about/intl/en_ALL/languages/index.html

respect for cultural identity, cultural and linguistic diversity" (UN 2003). However, as new information and communication technologies are opening new frontiers for innovation, creativity, and development, not everybody is able to participate, contribute and benefit equally.

The digital language divide, thus, holds back entire societies from sustainable development, from the information and the means of communication necessary for health and education, from opportunities to engage in cultural, political and economic development. The imperative to bridge the digital language divide, therefore, is rooted in the basic right of all communities, languages, and cultures to be "first class citizens" in an age driven by information, knowledge and understanding.

How to increase and mantain Digital Language Diversity?

For a more equitable world, we need *digital language diversity*, in the same way we need language diversity to preserve the entire heritage of human culture.

Increasing the level of Digital Language Diversity requires to increase the representation of languages over the Internet, either in terms of available content and in terms of possible uses. Availability of content, although desirable, is a necessary but not sufficient condition in order to guarantee true language digital vitality. A typical case is when there is a Wikipedia in a given language, but not localized interfaces of most popular applications and programs. A user cannot really interact using the language over digital devices. He can only access some web pages: in order to access the Internet and take profit of the services available on it, a user must switch to another language.

The vast majority of regional and minority languages (RMLs) are poorly represented digitally (Rehm, 2014e), (LTInnovate).

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, which is peculiar to Europe and other strongly monolingual States, is the fact that virtually no citizen is monolingual in a regional or minority language: everyone can always make use of an official major language instead of a minority one, thus making regional and minority languages not essential for communication purposes. This makes 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 [10]. According to the META-NET study, the majority of European 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 flourishes. 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, 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 majority of everyday tasks taking place over the Internet, from as simple ones such as writing emails to more complex ones such as listening to automatic speech translation, are supported by some kind of Language Technology (LT). This term broadly encompasses data and software that allow the automatic processing and recreation of natural language, such as spelling and grammar checkers, electronic dictionaries, localized interfaces, as well as search engines, automatic speech recognition and synthesis, language translators or information extraction tools. Language Technology can make content accessible, e.g. though cross-lingual information retrieval and machine translation. It can open up the possibilities for making purchases and perform transactions over the Internet across national boundaries. It can enable e-Participation, and thus contribute to social involvement. It can enable richer interaction among people from different linguistic backgrounds, and thus foster exchange of knowledge and social dialogue and cohesion.

Language Technology, thus, is a cornerstone of digital language diversity. It represents an *enabling technology* by means of which speakers can interact with machines and devices using their natural language (for a review of the crucial role of Language Technologies for fostering multilingualism and enabling the preservation of cultures and languages, see for instance Mariani 2015, and AA.VV. 2015). If we want to save and preserve language diversity, and especially minority and regional languages, we must necessarily let these lesser-used languages have access to the tools and resources of the same technological level as those of "bigger" languages.

However, despite its increasing penetration in daily applications, Language Technology is still under development for major languages. According to a research carried out by the META-NET Network of Excellence[11], culminated in the publications of 30 "Lan-

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

11 www.meta-net.eu

guage White Papers" (Rehm and Uszkoreit 2012), one for each official EU language, 29 European languages are at risk of digital extinction because of lack of sufficient support in terms of language technologies.

The study reports how Language Technology support varies considerably from one language community to another, and about dramatic and alarming differences in technology support between the various languages and areas are dramatic and alarming: in the four areas, English is ahead of the other languages but even support for English is far from being perfect. While there are good quality software and resources available for a few larger languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text analytics and essential resources. Others have basic resources but semantic methods are still far away. A recently update of the study (Rehm et al. 2014), demonstrates, drastically, that the real number of digitally endangered languages is, in fact, significantly larger.

The META-NET study described above clearly shows that, in our long term plans, we should focus even more on fostering technology development for smaller and/or less-resourced languages and also on language preservation through digital means. Research and technology transfer between the languages along with increased collaboration across languages must receive more attention.

However, it must be recognized that this represents a big challenge as well, as fast development of high quality LT is required to keep up with the pace of technological development. If a language does not enjoy good quality Language Technology, it won't be used in the latest voice or language-based applications; it will be replaced by another language and may thus get into the loop eventually leading to digital extinction. On the other hand, if Machine Translation is available for that same language, it will keep being used, even in confrontation with much more widely used languages.

Despite having improved enormously over the last decades, Language Technology is still far from being a perfect solution for multilingualism. As everyone knows, there are striking imbalances in applications and the overall final quality is acceptable for a final user for a handful of languages only. However, it must be recognized that their level of development is good enough to justify for more investment and for enlarging the technology to more languages. Some major companies, mostly from the US, are now starting to recognize the importance of multilingualism for their business but they mostly invest in languages of some economic interest.

In order to increase the presence of languages on the Internet and digital devices, i.e. in order to increase Digital Language Diversity, language technology must be enabled for as many languages as possible. It is by no means simple, for a minority language, to get engaged in the digital world. Small languages need to be given the voice, in technological terms. The challenges - ranging from digital divide and connectivity access, problems in terms of scripts and their digital encoding, lack of terminology, etc. to availability and development of language technologies - can be daunting. However, going digital is not impossible for languages, as long as some minimal conditions are met. Careful consideration and planning are needed in order to develop a roadmap for advancing the sustainability of less widely used languages in the digital world.

The strategy we propose here starts from two assumptions. The first one is that under the current data-driven paradigm of development of Language Technologies, production of digital data represents a major bottleneck: the development of language-based applications crucially depends on the availability of large quantities of open data (Soria et al. 2014). The second assumption is that since lesser used languages are of little economic interest to the major players and developers of language-based digital applications, it cannot be expected these solutions be nicely offered to the public, at least not in the short term. At the same time, further delay in development would only deepen the language digital divide by making the possibility more remote for lesser used languages to keep up the pace of the technological development available for better-resourced ones. Therefore, the moment is now: if we don't act quickly and effectively now, if carefully planned and focused intervention is not immediately carried out, it might be too late.

The Digital Language Diversity Project: putting the fate of languages in the hands of their speakers

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, but little public or private resources are devoted to the development of Language Technologies for smaller languages, as they are not of strong commercial interest for big companies. A way out of this can be offered by unleashing the power of speakers as data producers. We are digital "Tom Thumbs": speakers produce data, at an incredible pace. It has been estimated that every minute, Twitter users tweet 277.000 times, Facebook users share 2.460.000 pieces of content, email users send 204.000.000 messages, and YouTube users upload 72 hours of new video[12]. And this data has economic value since data is what is needed to develop Language Technology.

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[13].

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,

12 Source: https://www.domo.com/learn/data-never-sleeps-2

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

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 and for different types of user categories.

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.

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 '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 linguistic 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). 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 avail-

ability 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).

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, Karelian, Basque and Breton.

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.

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. The results of the survey are to be published in February 2017.

In addition to the assessment tools and self-educational instruments described in the previous sections, the DLDP project will make available to regional and minority language communities a sustainable instrument to help them support the 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) - 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 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 [14], the META-NET Strategic Agenda [15], and the FLaReNet Blueprint for Actions and Infrastructures [16].

Its innovative character lies in its specific focus on the particular needs and challenges of regional and minority languages.

Conclusions

Using the words of John Hobson (quoted by Kevin Scannell, (Scannell 2013), "The internet and digital world cannot save us. They cannot save Indigenous languages. Of course these things have benefits but they are not the Messiah. We don't need another website or DVD or multi-media application, these are short term, quick fix solutions. What we really need is sustainable initiatives, to create opportunities for Indigenous

14 http://www.npld.eu/uploads/publications/313.pdf

15 http://www.meta-net.eu/sra

16 http://www.flarenet.eu/sites/default/files/D8.2b.pdf

language users to communicate with each other in their native tongue. To get people speaking again."

It is only by using the languages through the Internet that they can be successfully revitalized and kept healthy, and this in turn is possible if current technology embeds language technology for a larger number of languages than those for which it is currently possible.

A widening of Digital Language Diversity is desirable and possible, as there is no limitation, in principle, to the number of languages accessing the Internet and content be provided in those languages.

Even if Digital Language Diversity will never be able to mirror the world's linguistic diversity, we can and should aim at least at a partial reflection of it. International and national policy makers should support and foster the digital presence of minority languages in particular - those more at risk of digital extinction. The range of technical and political challenges involved is very vast, and must be addressed at once in order to endow languages with the minimal necessary instruments in order to access the Internet and start producing content. The development of reliable indicators of Digital Language Diversity is also desirable and we argue that such an initiative should be collectively and collaboratively pursued. These indicators could be used to build and Index of Digital Language Diversity, to be used as a monitoring tool to assess digital language diversity in a certain area and highlight areas where intervention is needed (for instance, by singling out where effort should be channelled and funding directed).

Although the destiny of a language is primarily determined by its mother-tongue speakers and its broader cultural context, a Digital Language Planning could help directing the technological development of an under-resourced language, thus affording the language the strategic opportunity to have the same "digital dignity", "digital identity" and "digital longevity" as large, well-developed languages in the Web.

References

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

Brenzinger et al. 2003 - 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. *Language vitality and endangerment. Ad Hoc Expert Group Meeting on Endangered Languages*, March, 2003.

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

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

Fishman 2001 - Fishman, J. A. editor. *Can Threatened Languages Be Saved?* Multilingual Matters, 2001.

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

Harmon and Loh 2010 – Harmon, D., Loh, J. "The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages". *Language Documentation and Conservation, 4*, pp. 97-151, 2010.

Harrison 2010a – Harrison, K. D. *The Last Speakers*. National Geographic, 2010.

Harrison 2010b – Harrison, K. D. "The tragedy of dying languages". *BBC News,* 2010. Available too: http://news.bbc.co.uk/2/hi/8500108.stm.

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

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

Lewis et al. 2013 – Lewis, M. P, Simons, G. F., Fennig, C. D. (Eds.) *Ethnologue: Languages of the World*. Seventeenth edition. SIL International, 2013.

Loh and Harmon 2005 – Loh, J., Harmon, D. "A global index of biocultural diversity". *Ecological Indicators, 5*, pp. 231-241, 2005.

Loh and Harmon 2014 – Loh, J., Harmon, D. *Biocultural diversity: threatened species, endangered languages*. WWF Netherlands, 2014.

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

MAAYA 2012 – VV.AA. *Net. Lang: Réussir le cyberespace multilingue*. C&F Edition, 2012.

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

Nettle and Romaine 2000 – Nettle, D., Romaine, S. *Vanishing voices: the extinction of the world's languages*. Oxford: Oxford University Press, 2000.

Paolillo et al. 2005 – Paolillo, J., Pimienta, D., Prado, D. *Mesurer la diversité linguistique dans l'internet*. UNESCO, 2005. Available too: http://unesdoc.unesco.org/images/0014/001421/142186f.pdf.

Pimienta 2001 – Pimienta, D. "Quel espace reste-t-il dans l'internet, hors la langue anglaise et la culture 'made in usa'?" *Les Cahiers du Numérique, 2(3-4)*, 2001.

Pimienta et al. 2009 – Pimienta, D., Prado, D., Blanco, A. *Douze ans de mesure de la diversité linguistique dans l'internet: bilan et perspectives*. UNESCO, 2009. Available too: http://unesdoc.unesco.org/images/0018/001870/187016f.pdf.

Prado 2011 – Prado, D. Languages and cyberspace: Analysis of the general context and the importance of multilingualism in cyberspace. In E. Kuzmin and E. Plys (Eds.), *Linguistic and Cultural Diversity in Cyberspace. Proceedings of the International Conference*. Moscow: Interregional Library Cooperation Centre, pp. 72-82, 2011.

Rehm and Uszkoreit 2012 – Rehm, G., Uszkoreit, H. (Eds). *META-NET White Paper Series: Europe's Languages in the Digital Age*, 2012. Berlin: Springer. Available too: http://www.meta-net.eu/whitepapers/overview.

Rehm et al. 2014 – Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M. U., Mermer, C., Váradi, T., Kirchmeier-Andersen, S., Stickel, G., Prys Jones, M., Oeter, S., Gramstad, S. "An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era* (CCURL 2014), Reykjavik, Iceland, May 2014, pp. 30-37, 2014. Available too: http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Scannell 2013 – Scannell, K. "Endangered languages and social media". Presentation at the Workshop at INNET Summer School on Technological Approaches to the Documentation of Lesser-Used Languages, September 2013.

Soria et al. 2014 – Soria, C., Calzolari, N., Monachini, M., Quochi, V., Bel, N., Choukri, K., Mariani, J., Odijk, J., Piperidis, S. "The language resource Strategic Agenda: the FLaReNet synthesis of community recommendations". *Language Resources and Evaluation, 48(4)*, pp. 753-775, 2014.

Sutherland 2003 – Sutherland, W. J. "Parallel extinction risk and global distribution of languages and species". *Nature, 423*, pp. 276-279, 2003.

Turin 2013 – Turin, M. *Globalization helps preserve endangered languages*. Yale-Global Online, 2013. Available too: http://yaleglobal.yale.edu/content/globalizationhelps-preserve-endangered-languages.

UN 2003 – *UN Declaration of principles - building the information society: a global challenge in the new millennium*. UNESCO, 2003. Available too: http://www.itu.int/wsis/docs/geneva/official/dop.html.