Technologies for endangered languages: The languages of Sardinia as a case in point

Adrià Martín-Mor[1,2]

Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental, Universitat Autònoma de Barcelona (Catalonia)[3]

Abstract

The world's cultural diversity is at risk because of the current process of language desertification. Few places in the Mediterranean can boast the language diversity of Sardinia —a territory of 24,000 km^2 filled with languages and dialects. According to Moseley (2010) —and a similar scenario is described by Simons and Fennig (2017)—, all the local languages (Algherese Catalan, Gallurese, Sassarese, Sardinian and Tabarchin Ligurian) are «definitely endangered» and being replaced by Italian. Language policies at the official level do not seem to be able to revert the dramatic situation with these endangered languages, and their preservation is mostly left to the commitment of individuals, often with little recognition or help. As a result, the Sardinian languages live in a situation of diglossia, being mainly associated with folkloric matters which, in turn, reinforces a perception of uselessness.

Nonetheless, studies published by the Sardinian government show that society considers that the languages of Sardinia must be protected, and some interesting grass-roots actions related to technologies have been carried out. This article will describe recent examples of digital products which have been translated into or developed in some of the languages of Sardinia, mostly by volunteers and activists, with the aim of exploring how endangered communities can use technologies to contribute to the preservation of their languages.

Keywords: minoritised languages, languages of Sardinia, translation technologies, language planning, endangered languages.

1. The languages of Sardinia

Sardinia and the surrounding smaller islands form an archipelago of around 24,000

[1] This article is signed, as a citizen of the Catalan Republic proclaimed by the legitimate government of Catalonia, in protest against the imprisonment of political activists and members of the Catalan government and in solidarity with all the citizens who suffered reprisals by the Spanish state following the Catalan self-determination referendum held on the 1st October 2017.

[2] ORCID: 0000-0003-0842-3190.

[3] This work has been supported by the Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental.

km^2 in the middle of the Mediterranean. According to the law passed by the Sardinian government on the 15th October 1997[4], the languages of Sardinia are Sardinian, the "Catalan language of Alghero" (Algherese Catalan), "Tabarchin" (Ligurian), "Sassarese" (also called Turritan) and "Gallurese" (Corsican). While Algherese and Tabarchin are part of the Catalan and the Ligurian systems, the latter two are transitional languages between Sardinian and Corsican, with Gallurese being clearly closer to Corsican than Sassarese[5]. These languages are located in the north (Algherese Catalan, Gallurese Corsican and Sassarese) and in the south-west, in the archipelago of Sulcis (Tabarchin Ligurian), while Sardinian occupies the largest area of the island, as depicted in the map below[6].

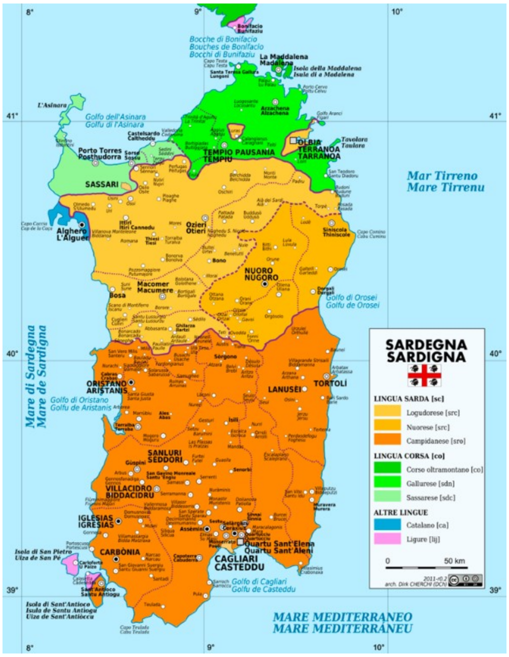

Image 1. The languages of Sardinia (image by Dch, https://commons.wikimedia.org/w/index.php?curid=6344881/)

In the map above, different shades of colour are used to distinguish the varieties

[4] http://www.regione.sardegna.it/j/v/86?v=9&c=72&file=1997026/.

[5] In contrast to the case of Algherese Catalan, the above-mentioned law avoids any reference to the linguistic systems to which Sassarese and Gallurese belong, through the use of an ambiguous formulation: "the Sassarese and Gallurese dialects".

[6] https://commons.wikimedia.org/w/index.php?curid=6344881/.

of the Sardinian language. The traditional confusion at various levels between languages and diatopic varieties or regional dialects has been exacerbated by strong linguistic prejudices subsumed by Giuseppe Corongiu (2013) under the umbrella of *orientalism*. Among these linguistic attitudes, as summarised by Joan Elies Adell (2013: 118) – such as an alleged *archaism* of the Sardinian language, its interference in children's process of learning Italian and the inferiority of the southern varieties –, there is the cliché that the differences between northern and southern varieties are "irreconcilable".

To address the confusion, in 2011 the Sardinian government issued the following language map, where each of the five languages of Sardinia are assigned a colour.

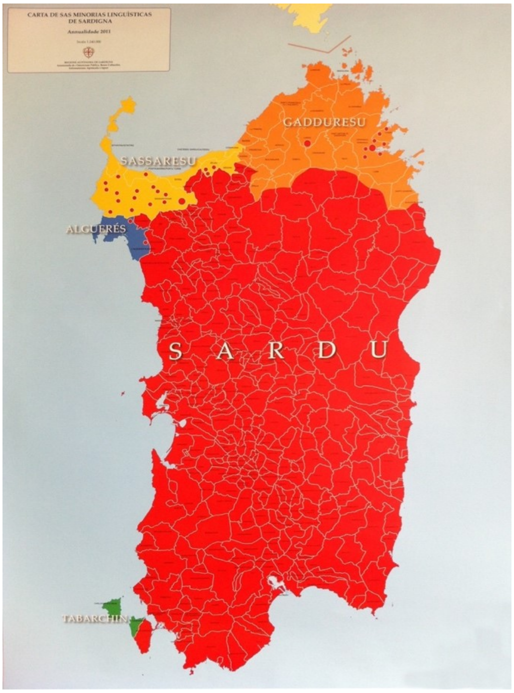

Image 2. Map of the languages of Sardinia according to the Sardinian government

This article will refer to the five languages of Sardinia as Algherese Catalan, Gallurese Corsican, Sassarese, Sardinian and Tabarchin Ligurian. Other languages spoken on the island include Italic languages – such as Venetian, Friulian and Istriot – brought to Sardinia as a result of migration during the fascist period, and Arbaresca (or Romaniska), a language spoken in the town of Isili (Aresu n.d.).

This number of languages and dialects is quite considerable taking into account the size of the territory and its population – 1.6 million. However, all the languages of Sardinia are in a context of diglossia (Russo & Soria 2017: 5), in which Italian is the only language in all social contexts on the island. As for the number of speakers and the status of each language, alarming figures are provided by *Ethnologue* (Simons & Fennig 2017) and Unesco's *Atlas of the World's Languages in Danger* (Moseley 2010).

Ethnologue – which considers Sardinian to be a "macrolanguage" including Campidanese Sardinian, Gallurese Sardinian, Logudorese Sardinian and Sassarese Sardinian[7] – identifies around 1,200,000 speakers of the Sardinian "macrolanguage"[8]. The following table includes the language name given by *Ethnologue*, the number of speakers and the status of each language on a scale which goes from 0 (International language) to 10 (extinct)[9].

Language	***Ethnologue* name**	**Speakers**	**Status**
Algherese Catalan	Catalan[10]	20,000	7 Shifting
Gallurese Corsican	Sardinian, Gallurese[11]	100,000	6b Threatened
Sardinian	Sardinian, Campidanese[12]	500,000	6a Vigorous
	Sardinian, Logudorese[13]	500,000	6b Threatened
Sassarese	Sardinian, Sassarese[14]	100,000	6b Threatened
Tabarchin Ligurian	Ligurian	?	?

Table 1. The languages of Sardinia according to Ethnologue

[7] This would conflict with the view of Gallurese and Sassarese being autonomous languages or dialects of languages other than Sardinian (cf. images 1 and 2).

[8] https://www.ethnologue.com/language/srd/.

[9] The complete scale includes international (0), national (1), provincial (2), wider communication (3), educational (4), developing (5), vigorous (6a), threatened (6b), shifting (7), moribund (8a), nearly extinct (8b), dormant (9) and extinct (10). For a detailed description, see https://www.ethnologue.com/about/language-status/.

[10] https://www.ethnologue.com/language/cat/.

[11] https://www.ethnologue.com/language/sdn/.

[12] https://www.ethnologue.com/language/sro/.

[13] https://www.ethnologue.com/language/src/.

[14] https://www.ethnologue.com/language/sdc/.

Tabarchin Ligurian does not appear in *Ethnologue*, apart from the reference to "possible scattered settlements on and around Sardinia" in the Ligurian information sheet[15].

Unesco's Atlas, on the contrary, lists four languages on the island of Sardinia: Algherese Catalan, Gallurese, Sassarese and Sardinian. The Atlas does contain information about Ligurian as a general language, including an explicit reference to Sardinia – "settlements in the towns of Carloforte on the San Pietro island and Calasetta on the Sant'Antioco island off the southwest coast of Sardinia, Italy"[16], which represents a population of around 10,000 people –, but Tabarchin Ligurian is not listed as a language of Sardinia. The Atlas does not provide any data, probably because of a technical error, on the number of speakers of Algherese Catalan[17].

Language	**Atlas name**	**Speakers**	**Status**
Algherese Catalan	Algherese Catalan	N/a	Definitely endangered
Gallurese	Gallurese[18]	100,000	Definitely endangered
Sardinian	Sardinian[19]	1,300,000	Definitely endangered
Sassarese	Sassarese[20]	120,000	Definitely endangered
Tabarchin Ligurian	Ligurian	?	Definitely endangered

Table 2. The languages of Sardinia according to Unesco

As can be observed in table 2, all the languages of Sardinia (probably including Tabarchin Ligurian, despite the fact that no specific data is provided for the Tabarchin variety of the Ligurian language) are considered to be "definitely endangered", on a scale ranging from "safe" to "extinct" including "definitely", "severely" and "critically" endangered[21].

[15] https://www.ethnologue.com/language/lij/.

[16] http://www.unesco.org/languages-atlas/en/atlasmap/language-id-377.html.

[17] Information on "Logudorese Sardinian", with ID number 381, was available at Unesco's Atlas but was removed at the time of writing of this article.

[18] http://www.unesco.org/culture/languages-atlas/en/atlasmap/language-id-356.html.

[19] http://www.unesco.org/culture/languages-atlas/en/atlasmap/language-id-337.html.

[20] http://www.unesco.org/culture/languages-atlas/en/atlasmap/language-id-408.html.

[21] Since Unesco's Atlas defines "definitely endangered" as "children no longer learn the language as mother tongue in the home", it could be argued, observing Sardinian society, that language statuses

Both tables, therefore, point to a worrying scenario for all the languages of Sardinia. A closer examination of sociolinguistic surveys reveals more interesting data, especially with regard to the intergenerational transmission of the languages of Sardinia.

Anna Oppo (2007: 34) shows that the population of Sardinia has quite a high level of knowledge of the corresponding local languages, as the following figures suggest.

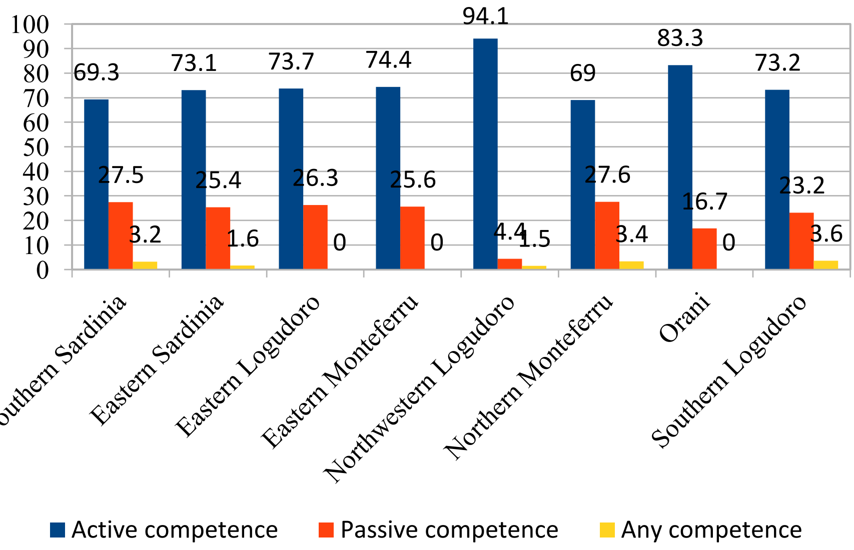

Figure 1. Language competence in Sardinian by areas according to Oppo (2007; our elaboration)

might be shifting towards "severely endangered", defined as "language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves".

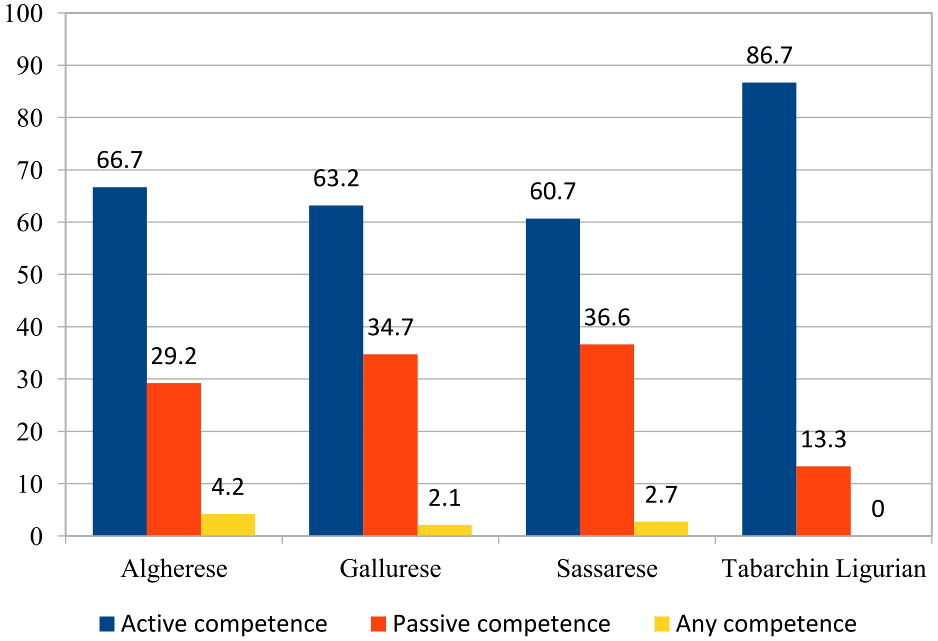

Figure 2. Language competence in the other languages of Sardinia according to Oppo (2007; our elaboration)

However, while Italian is the first language for 16.8% of the population over 65 years old, it is the first language for almost 90% of people between 15 and 24 years old and for 100% of children between 6 and 14 years old (Oppo 2007: 42). These data, collected more than ten years ago, would indicate that nowadays there are hardly any new speakers of the languages of Sardinia as a first language. The issue is exacerbated if one takes into account that the languages of Sardinia are not a compulsory part of the school curriculum.

As for the linguistic competence of the speakers of the languages of Sardinia, the survey in Oppo (2007: 70) shows a significantly lower figure for young people than for older people, except for Tabarchin Ligurian, which is clearly high at all ages.

Area	Local language	Age 15-34	Age 35-59	+60 years	N
Logudorese	Sardinian	59,7	78,3	95,3	321
Campidanese	Sardinian	55,6	69,8	84,7	631
Algherese	Algherese	34,6	56,0	58,5	84
Sassarese	Sassarese	42,8	41,7	38,3	156
	Sardinian	17,9	25,8	43,6	238
Olbia	Gallurese	34,4	34,5	59,5	77
	Sardinian	20,3	55,2	59,5	86
Tabarchin	Tabarchin	84,0	86,1	86,2	76

Table 3. Percentage of people professing to speak local languages according to Oppo (2007)

Nevertheless, Sardinian society overwhelmingly agrees that local languages should be protected to some extent (cf. table 6.13 in Oppo 2007: 58). Some action has been taken at the institutional level – such as the creation of a spell checker for the Sardinian language (see the case of Sardinian, below) –[22], and there are interesting private initiatives (such as publishing houses), research projects (Mura & Virdis 2015) and community efforts such as the localisation of social networks (Beccu & Martín-Mor 2016, Russo, Pisano & Soria 2016), web browsers and messaging systems (Martín-Mor 2016) into Sardinian, or the development of the machine translation system Apertium[23] for the Italian-Sardinian (Tyers et al. 2017) and the Catalan-Sardinian language pairs (see section 3.3 Sardinian, for more details). Some of the Sardinian languages have also reached a minimum consensus on their language model in recent years – such as Algherese Catalan (Scala 2003) and Tabarchin Ligurian (Toso 2005) –, although there are still some differences of opinion on the models to be adopted (Corongiu 2013).

The following section will address actions that can be taken in the technological field to contribute to a reversal of the dramatic situation of the languages of Sardinia. Basing on these actions, we will try to explore a methodology for the preservation of minoritised languages that can be applied to similar contexts.

[22] See Curretore Regionale Ortogràficu Sardu at http://www.sardegnacultura.it/cds/cros-lsc/.

[23] See www.apertium.org.

2. Preserving multilingualism through technologies

Endangered and minoritised languages can be preserved through several actions at different levels. One of these is through technologies. Indeed, the very act of increasing the digital presence of a language could be considered, ultimately, an action for its preservation, in the truest etymological sense of the word. Furthermore, as will be argued below, by making minoritised languages available through technology, not only will the interested users be able to use products in their languages – however few they may be –, but this will also enable software developers to create new language resources.

Three levels of technological actions are identified in this article: translation, localisation and the development of language tools and resources. Both linguistic competences and technical skills are required in each case, albeit at different levels. In the first and second cases, all free and open-source software[24] and many crowd-sourced products allow access and encourage language communities to localise software into as many languages as possible. Apart from the positive effects that translations contribute to any linguistic system (e.g., their contribution to the standardisation process, a key aspect in minoritised languages), the localisation of digital products has another positive effect on minoritised languages, in that it helps to raise their social profile by generating new terms for specialised subjects. As for the development of language resources, although it requires a level of specialisation, it is also true that it depends heavily on previously available language resources (see 2.3 Development of language tools). In other words, the translation of any kind of texts in a digital format facilitates the development of language technologies.

For these reasons, some language communities have identified an opportunity in technological volunteering and activism to preserve their minoritised languages[25].

Taking the experiences of these language communities into account, this section will highlight technological actions that can be taken to preserve endangered languages. Special attention will be given to free and open-source projects, since their licenses allow access and modification of the source code. This implies that anyone can participate in the project by contributing to the development of code or

[24] This article will refer to the *free software* definition provided by the Free Software Foundation, i.e. software that respects the user's four essential freedoms: run, change, redistribute and distribute modified copies of a piece of software (see https://www.gnu.org/philosophy/free-sw.html).

[25] To name a few, Librezale (www.librezale.eus) for the Basque community, Trasno (http://trasno.gal) for the Galician, Softastur (www.softastur.org) for the Asturian community and Softcatalà (www.softcatala.cat) for the Catalan.

by translating it. Furthermore, as noted by Adrià Martín-Mor and Alessandro Beccu (2016: 97), translating proprietary products might mean that translations cannot be reused in other projects, simply because the owner does not allow them to be exported:

> [E]st paradossale chi, a pustis de àere creadu unu de sos corpus parallelos prus mannos pro su sardu (Facebook bortadu dae s'inglesu a su sardu; fortzes su prus mannu in die de oe), sa comunidade sardòfona non potzat esportare nen recuperare sas tradutziones pro las torrare a impreare in àteros progetos.[26]

2.1 Translation

If we think of the internet as the greatest ever repository of information, one can easily understand how crucial it is for the preservation of endangered languages to be *archived* on the network. Among the various technological actions that can be undertaken, translation is the one requiring mainly linguistic competences, rather than technical skills. Many wiki projects, such as those by the Wikimedia foundation, require the user's collaboration in order to crowd-source content into several languages. Among the most successful projects by the Wikimedia foundation, there are dictionaries (www.wiktionary.org), news (www.wikinews.org) and the well-known Wikipedia (www.wikipedia.org). The fact that any user can modify and create new pages in these projects makes them especially interesting for endangered language communities. Editing digital information is free (compared to traditional publishing houses) and immediate (there are no intermediaries between the editor and the reader). The process of creating and modifying pages hardly requires technical skills: Wikimedia projects, such as Wikipedia, feature a graphical interface for any action the user might wish to take. Moreover, the Content Translation feature allows users to translate pages easily (see 2.3 Development of language tools).

2.2 Localisation

The goal of the actions included in this section is to make some products available in minoritised languages. As mentioned above, many digital products rely

[26] Paradoxically, after having created one of the biggest corpora for the Sardinian language (Facebook translated from English into Sardinian; maybe the biggest to date), the Sardinian-speaking community cannot export nor retrieve those translations in order to reuse them in other projects [our translation].

on crowd-sourcing to translate their interfaces into many languages. This is the case with most free and open-source software, and also for some proprietary projects. Localising digital products requires both knowledge of translation and technical skills, which might include the use of translation software or the knowledge of localisation conventions (variables, translation length, hotkeys, etc.). The following table gives examples of free projects we have identified for a range of categories of common software.

Name	Category
TuxPaint	Educational software
Telegram	Messaging service
Mozilla Firefox	Internet browser
LibreOffice	Office suite
GNU/Linux distribution	Operating system

Table 4. Common free digital products (compiled by author)

The software in the previous table covers most of the basic needs of any computer user. It also includes software for different levels of expertise: from educational software for children to complete operating systems. These programmes, furthermore, are all published under a free-software license.

TuxPaint (www.tuxpaint.org) is a free drawing programme for children. The main interface of the programme is available (totally or partially) in 128 local languages, according to its website[27]. The fact that it is educational software makes it especially suitable for translation by minoritised communities, because it allows children to learn basic concepts in the local language.

Telegram (www.telegram.org) is a free messaging service available for a wide array of mobile devices and operating systems. At the time of writing of this article, Telegram had just announced the creation of an online localisation platform[28]. Telegram allows any user to create and modify external language packs and apply them in their devices. Minoritised language communities – such as the Galician, the Basque and the Esperanto communities – get together in Telegram groups to distribute their localised language packs.

[27] http://www.tuxpaint.org/help/po/.

[28] http://translations.telegram.org/.

Mozilla Firefox[29] is a free web browser for mobile devices and computers. In 2017, Mozilla announced[30] that its localisation platform would move to Pontoon[31]. The localisation effort in the case of Mozilla Firefox, which is just one of the projects of the Mozilla Foundation, is much bigger in terms of translatable strings than in the previous examples.

LibreOffice is a free office suite[32], forked in 2010 from OpenOffice[33]. It includes a text processor, a spreadsheet programme and a presentation programme, among others, and is available in more than 100 languages. As in the previous example, LibreOffice is crowd-sourced by users through an online platform[34]. The effort involved in localisation is substantial, especially because the suite includes several programmes, as well as documentation and help files.

In a tentative continuum from smaller to bigger products in terms of the number of words, a final stage would probably be represented by the localisation of an entire operating system. Most associations of volunteers, such as those mentioned in 2. Preserving multilingualism through technologies, also maintain localisations of various distributions of the GNU/Linux operating system, such as Debian (www.debian.org), Linux Mint (www.linuxmint.com) and Ubuntu (www.ubuntu.org).

In our opinion, having the above-mentioned products translated into local languages allows users to meet most of their technological needs in their own languages, thus raising the visibility and the social perception of minoritised languages.

2.3 Development of language tools

Language resources, in paper and in digital form, are often not available for minoritised languages. This drawback hinders the above-mentioned translation and localisation actions, but, on the contrary, the existence of digital texts facilitates the creation of language resources and can contribute to the standardisation of minoritised languages (cf. the case of Sassarese, 3.4 below).

[29] http://www.mozilla.org/firefox/.
[30] https://blog.mozilla.org/l10n/2017/07/24/making-change-with-pootle/.
[31] http://pontoon.mozilla.org/.
[32] http://www.libreoffice.org/.
[33] http://www.openoffice.org/.
[34] https://translations.documentfoundation.org/.

Some of the most common language tools are spelling and grammar checkers. While spell checkers are commonly based on lists of words, grammar checkers tend to include rules (often in programming languages). Among the free resources for spelling and grammar checkers are the Hunspell spell checker[35] and the LanguageTool grammar checker[36].

If for a given language there are a coherent collection of texts (i.e., written with the same orthographic model) available in digital format, corpora can be built to assist translators and linguists in their linguistic decisions, and to build language support tools. Some programmes allow users to easily build corpora both by providing a collection of files or by downloading complete websites. Furthermore, some projects are devoted to building corpora by gathering published texts on Internet. This is the case with the OPUS[37] corpus (Tiedemann 2009), An Crúbadán[38] (Scannell 2007) and, although it does not only include digital texts but any kind of resource, OLAC[39].

Among the most powerful language tools, there is Machine Translation (MT). MT can be used for assimilation purposes– to allow users to understand texts – and for dissemination purposes – for the publication of texts (Forcada 2009). MT can have considerable effects on minoritised languages thanks to its potential contribution to the dissemination of languages. There are free platforms that allow the customisation of Statistical Machine Translation (SMT) engines (Martín-Mor 2017, Martín-Mor & Piqué 2017), which are built based on parallel and monolingual corpora. Most minoritised languages, however, do not have a sufficient number of texts in digital format, because of a lack of digital texts, a lack of consensus on the standardisation models, etc. In those cases, Rule-Based Machine Translation (RBMT) is especially useful, since rules can be manually written even when languages are not fully standardised (Martín-Mor, Piqué & Sánchez-Gijón 2016). Apertium is probably the best-known free MT platform. It has around forty languages, many of them minoritised languages with few or non-existent language resources.

The above-mentioned tools are often integrated. For instance, Wikipedia's

[35] https://hunspell.github.io/.
[36] http://www.languagetool.org/.
[37] http://opus.lingfil.uu.se/.
[38] http://www.crubadan.org/.
[39] http://www.language-archives.org/.

Content Translation[40] uses Apertium for some language pairs, so that editors can machine-translate and postedit articles. Wikipedia articles can be exported with the aim of creating corpora, which, in turn, can be used to create new language resources such as MT engines (Martín-Mor & Peña-Irles 2017).

3. The languages of Sardinia in technology

This section aims to summarise some of the technological products that are already available in the languages of Sardinia. Each sub-section will include references to the above-mentioned levels of technological actions – translation, localisation and language resources.

3.1 Algherese Catalan

Of all the languages of Sardinia, Algherese belongs to a community with the largest number of speakers. Indeed, the fact that Algherese shares its language with a community of almost 10 million people[41] guarantees that some basic linguistic needs are covered. The Algherese variant, however, has very little presence in technology.

The Catalan Wikipedia includes several pages written in Algherese. Some of these are found under the category "L'Alguer"[42] but, to the best of our knowledge, there is no way of exporting all the pages written in Algherese for the purposes explained in 2.3 Development of language tools.

The spelling and grammar checkers available for Catalan, even those – such as LanguageTool – that include several specific rules for the Valencian and the Balearic varieties, do not include Algherese either.

Despite the fact that all programmes listed under 2.2 Localisation are translated into Catalan – and some are even translated into other varieties of Catalan (cf. LibreOffice and Mozilla Firefox in Valencian) –, none of them is translated into Algherese. The same applies to the machine translation engine Apertium, which includes Valencian Catalan as a target language in some language pairs, but not Algherese.

Lastly, either OLAC[43] nor An Crúbadán[44] list specific resources for Algherese.

[40] https://www.mediawiki.org/wiki/Content_translation.
[41] See footnote 9.
[42] https://ca.wikipedia.org/wiki/Categoria:L%27Alguer.
[43] http://www.language-archives.org/language/cat/.
[44] http://crubadan.org/languages/ca/.

3.2 Gallurese Corsican

Gallurese is not present in any of the resources listed above. There are no Wikimedia projects in Gallurese and no software (to the best of our knowledge) is translated into Gallurese. OLAC contains a few reference works – mostly paper records[45] – and no corpora are collected by OLAC nor by An Crúbadán.

However, there are quite a number of resources and programmes available in the Corsican language, including Wikipedia (https://co.wikipedia.org/[46]), the free text editor for Windows Notepad++ (www.notepad-plus-plus.org), Facebook and even an MT engine by Google.

3.3 Sardinian

The Sardinian Wikipedia (sc.wikipedia.org) contains more than 5.500 pages, which makes it, at present, one of the biggest corpora available for a language of Sardinia, despite the fact that not all pages are written using the same language model (see Martín-Mor 2016: 116). There are no other Wikimedia projects for Sardinian. There is, however, a Wiktionary project in the Wikimedia incubator, a wiki devoted to testing the addition of new languages. This *Wikitzionàriu* contains 274 terms at the moment of writing this article[47].

As for localised digital products, as mentioned above, many programmes are translated into Sardinian. To name a few[48], the free text editor for Windows Notepad++ (www.notepad-plus-plus.org), Facebook, the web browser Vivaldi[49], and some components of the mobile operating system Ubuntu Touch[50], such as the uNav GPS navigator[51]. Telegram is translated and maintained by the Sardware team (www.sardware.tradumatica.net) and, as for GNU-Linux distributions, some of

[45] http://www.language-archives.org/language/sdn/.

[46] At the time of writing this article, the Corsican Wikipedia has 5,497 articles.

[47] https://incubator.wikimedia.org/w/index.php?title=Wt/sc/P%C3%A0gina_Base/.

[48] According to the website Aplicatziones in sardu (https://aplicatzionesinsardu.wordpress.com/), which monitors the applications localised in Sardinian, there are at least five other applications available in Sardinian language, of which four are Android apps.

[49] https://translations.vivaldi.com/languages/sc/.

[50] https://translate.ubports.com/languages/sc/.

[51] http://sardware.tradumatica.net/unav.html/.

them are partially translated into Sardinian (such as Debian[52], Linux Mint[53] and Ubuntu[54]).

Several corpora can be found on Internet. The OPUS corpus provides an Ubuntu corpus with around 1,000 translation units using Sardinian as the target language. Many blogs and sites use the Sardinian language, as listed in Russo and Soria (2017). Also An Crúbadán (which distinguishes between Sardinian[55] and Logudorese Sardinian[56]) hosts data extracted from Sardinian websites. Terminological term-bases (in TXT and TBX file formats)[57] and translation memories (in TMX file format)[58] are provided by the Sardware team, which updates and maintains bilingual (English-Sardinian) files of all products localised by them. As mentioned in section 1 (The languages of Sardinia), the Sardinian government sponsored the development of the CROS spell checker and of the Sintesa text-to-speech system[59]. Furthermore, the government's website for the Sardinian language provides some "experimental" language resources – such as a glossary (in PDF format), rules and grammatical guidelines[60] –, dictionaries[61], educational resources[62] and even a multimedia repository[63]. Lastly, Sardinian has been included in a recent release of the proprietary virtual keyboard Swiftkey.

3.4 Sassarese

The Wikimedia incubator hosts a Wikipedia project in Sassarese language with around one hundred articles[64], and the Italian Wiktionary holds around 200 terms in "Sassarese Sardinian" [sic][65]. Both projects, however, have few contributors and are rarely updated. The An Crúbadán project provides some linguistic data based on 44

[52] https://www.debian.org/international/l10n/po/sc/.

[53] https://translations.launchpad.net/linuxmint/latest/+lang/sc/.

[54] https://wiki.ubuntu.com/Ubuntu-Sardu/.

[55] http://crubadan.org/languages/sc.

[56] http://crubadan.org/languages/src .

[57] http://www.tradumatica.net/sardware/glossariu/.

[58] http://www.tradumatica.net/sardware/resursas/.

[59] http://www.sardegnacultura.it/index.php?xsl=267&s=7&v=9&c=28610&nodesc=1/.

[60] http://www.sardegnacultura.it/j/v/258?s=20340&v=2&c=2730&t=7/.

[61] http://www.sardegnacultura.it/j/v/290?s=7&v=9&c=2731&c1=2802&o=1&bt=1&na=1&n=1000/.

[62] http://www.sardegnacultura.it/linguasarda/ilsardo/multimedia.html.

[63] http://www.sardegnadigitallibrary.it/.

[64] https://incubator.wikimedia.org/wiki/Wp/sdc/P%C3%A0gina_prinzipari.

[65] https://it.wiktionary.org/wiki/Categoria:Parole_in_sardo_sassarese.

documents culled from the Sassarese Wikipedia[66]. Togo[67] is the name of an online dictionary for the Sassarese-Italian language pair containing around 3,500 words, which is available also as an Android app[68].

Interestingly, as Sardinian, Sassarese has also been included in a recent release of the proprietary virtual keyboard Swiftkey[69], probably due to the existence of the above-mentioned resources[70].

3.5 Tabarchin Ligurian

Despite the fact that, to the best of our knowledge, there are no digital products translated nor localised into Tabarchin Ligurian, there are some in general Ligurian. Among these, a Ligurian Wikipedia, with – at the time of writing this article – more than 3,000 articles[71], a Ligurian language pack for the Mozilla Firefox web browser[72] and a localised interface for the text editor Notepad++.

An Crúbadán contains a corpus of almost 300,000 words under the name Tabarkin, but it is made up of texts written in general Ligurian language (as the codification lij, on the website seems to indicate)[73].

4. Discussion and concluding remarks

The following table summarises the common free digital products presented in section 2 (Preserving multilingualism through technologies) and their availability in the languages of Sardinia.

Language	Wikimedia projects	TuxPaint	Telegram	LibreOffice	Firefox	Apertium	Linux
Algherese Catalan	Some articles in	Catalan	Catalan	Catalan	Catalan	Catalan	Catalan

[66] http://crubadan.org/languages/sdc.

[67] http://www.togo.sassari.tv/.

[68] https://play.google.com/store/apps/details?id=com.telesassari.togo.

[69] https://blog.swiftkey.com/swiftkey-keyboard-android-update-brings-sleek-redesign-new-themes/.

[70] This example illustrates how it is possible to develop, with few resources, language tools that facilitate the creation of more texts in digital format. As the developer of the software puts it, "[i]t requires at least 5,000 words in a language to be able to build a keyboard for it". This corpus, that can be accessed from online sources, can subsequently be increased, since "[a]s the language model gains users, its vocabulary grows more quickly" (https://blog.swiftkey.com/multilingual-milestone-swiftkey-reaches-support-for-150-languages/).

[71] https://lij.wikipedia.org/wiki/Pagina_prin%C3%A7ip%C3%A2/.

[72] https://www.mozilla.org/lij/firefox/new/.

[73] Cf. http://www-01.sil.org/iso639-3/documentation.asp?id=lij.

	the Catalan Wikipedia						
Gallurese	no	no	no	no	no	no	no
Sardinian	Active Wikipedia; Wiktionary (incubator)	no	yes	no	no	yes	yes
Sassarese	Wikipedia (incubator)	no	no	no	no	no	no
Tabarchin Ligurian	Ligurian	no	no	no	Ligurian	no	no
Language	Wikimedia projects	TuxPaint	Telegram	LibreOffice	Firefox	Apertium	Linux

Table 5. The availability of common free digital products in the languages of Sardinia

Despite the fact that some Sardinian languages are available in many projects, the table above shows how difficult it is for speakers of Sardinian languages to carry out common technological processes in their local languages.

This article has sought to gather examples of digital products and resources for the languages of Sardinia, with the aim of exploring a methodology for the digital preservation of minoritised languages through the selection of common free digital products (wiki websites, educational software, messaging services, web browsers, office suites and operating systems). This proposal distinguishes three levels of actions (translation of digital products, localisation of digital products and development of language tools), in which the output of each level can provide the input for the next task, thus creating feedback loops between stages. As mentioned above, translation and localisation can provide materials for the creation of language tools (such as corpora and machine translation systems), which, in turn, can assist translators and localisers and thus increase the volume of published texts. Furthermore, associating minoritised languages with technology can have a positive effect on their social profile. At the language planning level, it is important to release output resources under free licenses, as these grant access to the source code and encourage users towards its translation, modification and redistribution (see footnote 26) at no cost.

The global process of language desertification is clearly affecting modern and European societies such as the Sardinian, with all its five indigenous languages considered to be "definitely endangered" (see 1. The languages of Sardinia). In this dramatic context and considering the digital presence of the languages of Sardinia that this article has reviewed, one may have the impression that Sardinian, paradoxically, is the wealthiest of the minoritised languages of Sardinia, given the relatively high number of digital products available.

Nonetheless, as stated in the European Charter for Regional or Minority Languages (signed by Italy in the year 2000)[74], regional and local authorities should promote minoritised languages, encourage their use ("in speech and writing, in public and private life") and promote their study and research at universities. The Charter explicitly mentions, in article 12, "the use of new technologies" before stating that the parties should undertake

> to create and/or promote and finance translation and terminological research services, particularly with a view to maintaining and developing appropriate administrative, commercial, economic, social, technical or legal terminology in each regional or minority language.

In line with the Charter, therefore, this article has tried to outline some actions that follow the footsteps of other language communities. It is important to observe that most of the products that have already been translated or localised into the languages of Sardinia are the result of voluntary and collaborative projects, as are the proposals made in this article. If given adequate institutional support, these and other similar projects could be carried out economically and rapidly, giving birth to new job opportunities based on knowledge, and perhaps eventually reducing the island's unstoppable depopulation, while at the same time promoting research and a respect for minorities.

Such a language policy would require, in our view, at least a linguistic and technological training programme which is closely linked to educational institutions, for students of all five languages. A subsequent stage would involve the creation of interlinguistic teams (speakers of all five languages) and intralinguistic teams (in the case of languages spoken also outside Sardinia), with the aim of translating and localising digital products and developing language resources. These teams would need to be strongly coordinated in order to share resources, optimise results and promote actions. For instance, in the case of Sardinian languages also spoken outside Sardinia (such as Algherese Catalan, Tabarchin Ligurian and

[74] https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/148/.

Gallurese Corsican), intralinguistic teams could explore the possibility of automating processes through search-and-replace macros or scripts (cf. with existing automated variant adapters[75], such as those for Valencian Catalan[76] or for British English, as compared to American English[77]).

As Russo and Soria state (2017: 17), "[t]he 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 [Sardinian] to become a fully digital language". There is no reason not to think that if other languages follow the same steps, they too can become fully digital languages.

Adrià Martín-Mor
Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental
Universitat Autònoma de Barcelona (Catalonia)
adria.martin@uab.cat

Bibliography

Adell, Joan Elies (2013). "Diversitat lingüística a Sardenya". *DivÈrsia* 4. http://www.raco.cat/index.php/diversia/article/download/271361/359014 (accessed 16/09/2018).

Aresu, Massimo (n.d.). *S'arromaniska. Il caso dei ramai isilesi*. http://xoomer.virgilio.it/almelis1/albertomeliszingarimassimoaresuc.htm (accessed 16/09/2018).

Corongiu, Giuseppe (2013). *Il sardo: una lingua 'normale'*. Cagliari: Condaghes.

Forcada, Mikel L. (2009). "Apertium: traducció automàtica de codi obert per a les llengües romàniques". *Linguamática* 1 (1), 13–23, http://www.linguamatica.com/index.php/linguamatica/article/view/18 (accessed 16/09/2018).

Martín-Mor, Adrià (2016). "La localització de l'apli de missatgeria Telegram al sard: l'experiència de Sardware i una aplicació docent". *Tradumàtica: tecnologies de la traducció* 14, 112–127, http://revistes.uab.cat/tradumatica/article/view/n14-martin-mor (accessed 16/09/2018).

[75] https://www.altlang.net/.

[76] http://www.softvalencia.org/adaptador/.

[77] https://codewordsolver.com/american-british-english-translator/.

— (2017). "MTradumàtica: Statistical machine translation customisation for translators". *Skase Journal of Translation and Interpretation* 11 (1), 25–40. http://www.skase.sk/Volumes/JTI12/pdf_doc/02.pdf (accessed 16/09/2018).

Martín-Mor, Adrià & Beccu, Alessandro (2016). "Sa localizatzione de Facebook in sardu". *Tradumàtica: tecnologies de la traducció* 14, 85–99. http://revistes.uab.cat/tradumatica/article/view/n14-martin-mor-beccu (accessed 16/09/2018).

Martín-Mor, Adrià, Piqué, Ramon & Sánchez-Gijón, Pilar (2016). *Tradumàtica: Tecnologies de la traducció*. Vic: Eumo.

Martín-Mor, Adrià & Peña-Irles, Víctor (2017). "Creació d'un motor de TAE especialitzat en medicina per a la combinació romanés-castellà". *Linguamática* 9 (2), 45–53. http://www.linguamatica.com/index.php/linguamatica/article/view/v9n2p4 (accessed 16/09/2018).

Martín-Mor, Adrià & Piqué, Ramon (2017). "MTradumàtica i la formació de traductors en Traducció Automàtica Estadística". *Tradumàtica: tecnologies de la traducció* 15, 97–115. http://revistes.uab.cat/tradumatica/article/view/n15-martin-pique (accessed 16/09/2018).

Moseley, Christopher (ed.) (2010). *Atlas of the World's Languages in Danger*, 3rd edn. Paris, UNESCO Publishing. Online version. http://www.unesco.org/culture/en/endangeredlanguages/atlas (accessed 16/09/2018).

Mura, Riccardo & Virdis, Maurizio (eds) (2015). *Caratteri e strutture fonetiche, fonologiche e prosodiche della lingua sarda: il sintetizzatore vocale SINTESA*. Cagliari: Condaghes.

Oppo, Anna (ed.) (2007). *Le lingue dei sardi. Una ricerca sociolinguistica*. Cagliari: Regione Autonoma della Sardegna. http://www.regione.sardegna.it/documenti/1_4_20070510134456.pdf (accessed 16/09/2018).

Russo, Irene, Pisano, Simone & Soria, Claudia (2016). "Sardinian on Facebook: Analysing Diatopic Varieties through Translated Lexical Lists". Basile, Pierpaolo, Corazza, Anna, Cutugno, Francesco, Montemagni, Simonetta, Nissim, Malvina, Patti, Viviana, Semeraro, Giovanni & Sprugnoli, Rachele (eds). *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016, 5–6 December 2016, Napoli*. Torino: Academia University Press, 263–267. http://dblp.uni-trier.de/db/conf/clic-it/clic-it2016.html#RussoPS16 (accessed 16/09/2018).

Russo, Irene & Soria, Claudia (2017). *Sardinian — a digital language?* http://www.dldp.eu/sites/default/files/documents/DLDP_Sardinian-Report.pdf (accessed 16/09/2018).

Scala, Luca (ed.) (2003). *Català de l'Alguer: criteris de llengua escrita: model d'àmbit restringit de l'alguerès, document aprovat per l'Institut d'Estudis Catalans*. Barcelona: Publicacions de l'Abadia de Montserrat.

Scannell, Kevin P. (2007). "The Crúbadán Project: Corpus building for under-resourced languages". *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, vol. 4, 5–15. http://borel.slu.edu/pub/wac3.pdf (accessed 16/09/2018).

Simons, Gary F. & Fennig, Charles D. (eds) (2017). *Ethnologue: Languages of the World, Twentieth edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com (accessed 16/09/2018).

Tiedemann, Jörg (2009). "News from OPUS-A collection of multilingual parallel corpora with tools and interfaces". Nicolov, Nicolas, Bontcheva, Kalina, Angelova, Galia & Mitkov, Ruslan (eds). *Recent advances in natural language processing*, vol. 5. Amsterdam/Philadelphia: John Benjamins, 237–248. http://stp.lingfil.uu.se/~joerg/published/ranlp-V.pdf (accessed 16/09/2018).

Toso, Fiorenzo (2005). *Grammatica del tabarchino*. Recco: Le Mani.

Tyers, Francis M., Alòs i Font, Hèctor, Fronteddu, Gianfranco & Martín-Mor, Adrià (2017). "Rule-Based Machine Translation for the Italian-Sardinian Language Pair". *The Prague Bulletin of Mathematical Linguistics* 108 (1), 221–232. https://doi.org/10.1515/pralin-2017-0022 (accessed 16/09/2018).

Adrià Martín-Mor is a lecturer at the Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental (Universitat Autònoma de Barcelona, Catalonia). He is a member of the Tradumàtica research group at UAB, where he coordinates training programmes in translation technologies such as the Tradumàtica MA and the Tradumàtica Summer School. He holds an MA and a PhD on Translation Studies and his research interests are the automation of the translation process (Computer-Assisted Translation, Machine Translation), minoritised languages and free software for translation.