The Digital Language Diversity Project

Digital Language Survival Kit

The DLDP Recommendations to Improve Digital Vitality

Erasmus+

Imprint

The DLDP Digital Language Survival Kit

Authors:
Klara Ceberio Berger, Antton Gurrutxaga Hernaiz, Paola Baroni, Davyth Hicks, Eleonore Kruse, Valeria Quochi, Irene Russo, Tuomo Salonen, Anneli Sarhimaa, Claudia Soria

This work has been carried out in the framework of The Digital Language Diversity Project (www.dldp.eu), funded by the European Union under the Erasmus+ Programme (Grant Agreement no. 2015-1-IT02-KA204-015090)

Cover design: Eleonore Kruse

Disclaimer
This publication reflects only the authors' view and the Erasmus+ National Agency and the Commission are not responsible for any use that may be made of the information it contains.

www.dldp.eu
www.facebook.com/digitallanguagediversity
dldp@dldp.eu
www.twitter.com/dldproject

Recommendations at a Glance

Digital Capacity		
Indicator	**Level**	**Recommendations**
Digital Literacy	2,3	Increasing digital literacy among your native language-speaking community
	2,3	Promote the upskilling of language mentors, activists or disseminators
	2,3	Establish initiatives to inform and educate speakers about how to acquire and use particular communication and content creation skills
	2	Teaching digital literacy to children in your language community through the medium of your language from the outset
Character Encoding, Input and Output Methods	2,3	Ensuring that you have a dedicated keyboard for your language
Availability of Language Resources	2,3	Develop basic language resources
	2,3	Dictionary making
	2,3	Spell Checker
	2,3	Start up the corpus experience
	2,3	Use tools such as concordancers for corpus querying
	4	Develop intermediate and advanced language resources
	4	Dictionary making: diversity, size, specialization and dissemination
	4	Increase corpus size and diversity
	4	Collect publicly available linguistic data from social media
	4	Develop a part-of-speech tagger
	4	Use tools for corpus analysis and feed your dictionary with data about language in use
	4	First steps toward speech synthesis and recognition

Digital Presence and Use		
Indicator	**Level**	**Recommendations**
Use for E-Communication	2,3,4	Estimating the value of RML use for interpersonal communication
Use on Social Media	2,3	Visualizing the value of RML use on social media
Availability of Internet Media	2,3,4	Increase the amount of content and diversify the types of Internet media
	2,3,4	Increase the amount of text-type content (websites, blogs, forums)
	2,3,4	Create or feed a web-based archive of documents and recordings
	2,3,4	Stream online using free software tools
	2,3,4	Record digital stories in your own language
	2,34	Promote subtitling initiatives
Wikipedia	2	Create a Wikipedia in your language
	3,4	Take your Wikipedia to a higher level
	3	Promote an initiative to increase the number of wikipedia entries in your language
	4	Initiatives to increase the size and quality of Wikipedia

Digital Performance		
Indicator	**Level**	**Recommendations**
Availability of Internet Services	3,4	Expand the range of possibilities to use Internet services in your language
	3,4	Collect information and experiences from your RML users' community, to determine which are the most important and used services
	3,4	Estimate the value of using users' language in business
	3,4	Develop smartphone apps
Localised Social Network	3,4	Initiatives for localising social media user interfaces
Localised Software: Operating Systems and Basic Software	3	Start a community effort to localise free software
	4	Strengthen initiatives to localise the general purpose free or proprietary software most used in the language community
	4	Consider video games as a valuable revitalisation opportunity
Machine Translation Services	3,4	Pave the way to Google translation through community involvement
	3	Develop and promote at least one MT system to and from the majority language
	4	Expand the number of language pairs; if it is not already, try to make English one of the languages included
Dedicated Internet Top-Level Domain	4	Time to get an internet domain for the language

1. Recommendations to Improve Digital Vitality

The Digital Language Survival Kit is an instrument which aims at allowing regional and minority languages' speakers and communities to a) self assess the degree of vitality of their language and b) learn about the kind of concrete actions and initiatives that can improve this level of vitality. This document is devoted to the second objective.

In this set of recommendations, we suggest some actions that may be taken—mostly at the grass-roots level — to make a language progress towards the next steps of digital vitality.

The recommendations are organized in three sections, each one related to a type of indicator of digital vitality.

The recommendations are intended for three levels of digital vitality: Dormant, Emergent and Developing. Some recommendations are specific for a level, and some others are suitable for different levels.

1.1 Three Types of Digital Vitality Indicators

1.1.1 Digital Capacity

By *digital capacity* we mean the extent to which a language is infrastructurally and technologically supported and so that it is able to function in the digital world. A precondition is that the language should have at least one writing system; without it, no function in the digital world is possible. Basic conditions such as availability of Internet connection and digital literacy must be met for a community to use a language digitally. Similarly, the existence and availability of language resources and tools greatly determines the functionality of a language in digital contexts. For instance, functionalities such as spell checkers on smart phones can boost - in principle - its use by making its typing easier and faster. A language's digital capacity only refers to its potential to be used digitally, but it by no means guarantees that a community will use it. This is the case, for instance, of many European regional or minority languages: although most of these languages meet all the requirements of digital capacity, they are often little used in comparison to the official state language of the countries where they are spoken. Other factors (psychological submission, lack of competence in the written language, lack of available digital spaces - forums, blogs - where the language is used) can determine a poor digital use.

1.1.2 Digital Presence and Use

Once the infrastructural level of digital capacity is secured, it becomes possible for a language to be used on a variety of different media and for a wide range of different purposes. The second group of indicators (from 6 to 9) refers to how, and how much, a language is digitally used: whether and the extent to which it is used for communicating, for creative content production, or for educational entertainment purposes, for example. The common denominator for this group of indicators is that they relate to the creation of digital content in the language, whether it is used for communicating or for other purposes. Again, the indicators are ordered so as to suggest a certain progression upwards: texting, messaging, and e-mailing are seen as more basic functions than, for instance, writing Wikipedia articles or developing e-books or video games in the language. However, that does not imply that this order should be taken as a ladder, to go up rung by rung. Therefore, it is not mandatory that we must have Wikipedia to start producing or localising video games. The communicative function is considered as more basic than the other ones, following Gibson [8]. These digital uses of the language also encompass a progression from more private uses of the language to more public (and often official) ones. We have highlighted four indicators for this class: *Use for e-communication, Use on social media, Availability of Internet media,* and *Wikipedia.*

1.1.3 Digital Performance

Digital Performance groups together indicators referring to what can be digitally done with a language. This is a different way of looking at the extent to which a language is used over different media. The perspective here focuses more on the purposes for which it is used rather than on the

range of available media where it is used. We have identified five indicators for this group: *Availability of Internet services, Localised social network, Localised software, Machine translation tools/services,* and *Dedicated Internet top-level domain.*

1.2. Levels of Digital Vitality

In order to portray digital vitality, different levels have been described in the document "How to use the Digital Language Vitality Scale" (Ceberio et al., 2018). The scale has six levels, and the recommendations in this document are intended for three of them:

Level 2 – Dormant

For a language at the Dormant level of vitality, connectivity is ensured and there is some degree of Internet penetration; at least some language speakers are digitally literate. However, there is no technological support for the language (e.g. there is no keyboard support for writing the language). The language could be digitally used but it needs some basic technological development.

Level 3 – Emergent

For a language to be Emergent means that connectivity is well developed to enable pervasive use of the Internet and thus potentially use of the language digitally. Internet penetration is good and speakers are digitally literate. Overall, the language has limited technological support, but some basics are available and it is sometimes used at least for texting and instant messaging (i.e. private communication). A Wikipedia for the language might exist, but it is small and speakers might either not use it or not be aware of it. A few basic language resources might exist already.

Level 4 – Developing

A language at the Developing level of digital vitality shows some usage over communication and social media, although frequency may still be occasional. Some digital media and services may be available, as well as a medium-sized Wikipedia; basic language resources are in place, and there might be evidence of more advanced ones. At least one among the social media and the operating systems used by the speakers' community might be localised. An online machine translation service or tool might be available, for one language pair at least.

The lowest one, Pre-digital, and the highest two levels, Vital and Thriving have not been considered. In the case of Pre-digital, the recommendations should be aimed at increasing the connectivity of the communities, the digital literacy of the users, and solving the problems related to the encoding of characters. The languages examined by the DLDP are not in that situation. As for the two upper levels of the scale, there are few minoritised languages that enjoy this status; mostly, at these levels we find official languages of the EU.

1.3 Structure of the Recommendations

Each recommendation is structured as follows:

- The level(s) for which the recommendation is suitable: some recommendations are specific for a level, and some others are suitable for more than one level.
- Explanatory text to motivate and describe the recommendation.
- Addressees for whom the recommendation is intended: from individuals, users' groups and associations, to organisations and institutions, including research groups, software developers and companies.
- Examples: successful or interesting cases in which the initiatives proposed in the recommendation have been carried out, or that can illustrate how it could be implemented.
- Further reading: articles, blog posts or academic papers providing additional information on the recommendation.
- Related module in the Training Program (TP) that contains information relevant to the recommendation.

2. Digital Capacity

Main Course of Action: Preparing your language for the digital environment

- As a basic skill, promote literacy in the RML.
- Ensure good, up-to-date, connectivity and pervasive internet penetration.
- Promote (medium-high) digital competence of RML speakers (potential digital users).
- Develop language resources and tools, involving different agents (users' communities, research groups, companies, policy makers).

2.1 Digital Literacy

R1 Increasing digital literacy among your native language-speaking community

Suitable for: *2 Dormant, 3 Emergent*

In order to actively participate in the digital world and to increase content in your own native language, language speakers need first to possess a variety of up-to-date digital competences and feel confident about them, irrespective of the language of use.

Taking DigComp 2.1 as a common reference, we can assume that, at this stage of language vitality, an average speaker would be at an 'Intermediate' proficiency level. The goal is thus to consolidate and increase the amount of speakers possessing this level of skills and at the same time promote the acquisition of 'advanced' proficiency in a variety of competences, particularly those identified under the first three competence areas (see DIGCOMP The Digital Competence Framework For Citizens – The Competences):

- Information and data literacy. Specifically: use of advanced search strategies/functions of different search engines; assess the reliability of the information found; use news feeds to keep updated on topics of interest; awareness of the novelties in information search and storage; use of cloud storage.

- Communication and collaboration: use of a variety of communication tools (e-mail, chats, sms, blogs…), create and manage content with collaborative tools, active participation on online spaces, use of various online services (e-banking, ...), use of advanced communication functionalities and platforms such as video-conferencing and data-sharing.

- Digital content creation: production and modification of complex multimedia content, use of a variety of digital platforms, tools and environments, creation of websites, use of advanced formatting functions of various word processing tools, correct application of licenses and copyright, use of programming languages, databases.

R1.1 Promote the upskilling of language mentors, activists or disseminators

As these actors are the drivers and catalysts of the production of digital content, it is fundamental they possess higher skills than the average person in the community. Large associations and institutions should offer free online courses on advanced technology and encourage language activists and teachers especially, and other disseminators, to take these courses in order to upgrade their skills.

Addressees: associations, institutions, policy makers.

Examples:

» MENTEP – Materials to develop digital competence in teaching
» MOOC List | Find MOOC and Free Online Courses from the Best Providers
» Google Training Center Learn the best ways to use Google Tools for reporting and storytelling.
» Digital Tools Catalog – Poynter's News University
» Udemy - Video Production – The Basics! Learn Video Production in an hour. – Free Course

Further reading:

» Catálogo – Andalucía es Digital
» Digital Culture and Writing (Culture et Écriture Numériques) (EMMA)
» Europeana Space: Creative with Digital Heritage (edX)
» Learn How to Build Ecommerce Website From Scratch (Eduonix)
» Developing Digital Skills in your Classroom – 2nd Round – EUN Academy

R1.2 Establish initiatives to inform and educate speakers about how to acquire and use particular communication and content creation skills

Suitable for: *2 Dormant, 3 Emergent*

A simple initiative is to select and disseminate, through your personal contacts, on social media, or on your blog or website, relevant MOOCs, e-learning, and other open learning materials for improving digital skills among your community of speakers.

If you feel creative enough and have some expertise, you might also want to create a small video tutorial, a slideshow, or even to organise a face-to-face education meeting on relevant issues for your audience.

Addressees: individuals, language activists, associations.

Examples:

» https://www.wikihow.com
» https://www.aranzulla.it/

Related module in TP: 5

R1.3 Teaching digital literacy to children in your language community through the medium of your language from the outset

Suitable for: *2 Dormant*

Teaching digital literacy to children is of great importance. In this way, the language will be normalized for them in a digital environment and where children are able to progress far more than their parents in terms of coding skills, for example. Some open source material can be localized to help this process, e.g. Scratch, training materials by Raspberry Pi, or the Playful Coding - Learn2code & Code2learn project.

Addressees: individuals, language activists, associations.

Examples:

» About Scratch
» Getting started with the Raspberry Pi

» PlayFul Coding Videos

Further reading:

» Prys, D., Jones, D., & Ghazzali, S. (2017). Using LT tools in classroom and coding club activities to help LRLs Language Technologies in Support of Less-Resourced Languages. In *8th Language and Technology Conference*. Poznan, Poland.

2.2 Character Encoding, Input and Output Methods

R2 Ensuring that you have a dedicated keyboard for your language

Suitable for: *2 Dormant, 3 Emergent*

A vast majority of people might not be able to type properly using their language because of the lack of a keyboard specific to that language, which includes the missing diacritics or symbols that are not universally supported by existing keyboards. This might easily discourage the most passionate writer. Ensure that you have a dedicated keyboard installed on your devices for your language. There are many free software applications for language input which removes the tedium of typing in a less supported language. For instance, SwiftKey makes available virtual keyboards for more than 190 languages and runs on both iOS and Android platforms. Google's Gboard gives support to 185 language varieties.

If a virtual keyboard is not yet available for your language, you will need to lobby (together with local institutions or language activists) software developers to produce one.

Addressees: individuals

Examples:

» SwiftKey Keyboard for Android – faster, easier mobile typing
» Keyman – Type to the world in your language
» Gboard: now available for Android

Further reading:

» SwiftKey for iOS brings support for 100+ languages, design refresh and animated themes!
» 10 Cool Gboard Features and How to Use Them - Hongkiat

Related module in TP: 5

2.3 Availability of Language Resources

Language resources are an essential condition for the development of more advanced language-based computer applications.

Although it is common to use the term *language resources*, it must be taken into account that tools are also included within that denomination. Among the resources are dictionaries, collections of texts or corpora, grammars, lexical and terminological databases, and knowledge bases, ontologies. Among the tools, spelling and style checkers, morphosyntactic analyzers, part-of-speech-taggers (POS taggers) and syntactic parsers, terminology and multiword expressions (MWE) extractors, automatic translators, and speech synthesis and recognition tools.

In order to evaluate the level of vitality of a language in relation to linguistic resources, in the document "How to use the Digital Language Vitality Scale" (Ceberio et al., 2018) we have classified the resources and tools into three levels: basic, intermediate and advanced resources.

- Basic: monolingual and bilingual e-dictionaries; digital corpus (<100 million words); spell checker.
- Intermediate: corpus driven monolingual dictionary; digital corpus (>100 million words); parallel corpora; web-corpora; term extraction software; Part-of-Speech tagging; basic machine translation (rule-based); speech synthesis.
- Advanced: large corpora (>one thousand million), multilingual corpora; syntactic parsing; WordNet, semantic processing; advanced machine translation (statistical, hybrid, neural); speech recognition.

Of course, this classification is only indicative and should not be considered too strictly, but it can be useful for structuring the values of this indicator, and for organizing the recommendations below.

R3 Develop basic language resources

Suitable for: *2 Dormant, 3 Emergent*

Languages at the Dormant level of vitality suffer from scarcity of linguistic resources. Only the most basic resources are usually available. By far, the most basic resource is the dictionary, and, in the case of minority languages, the first existing dictionary is usually bilingual. There is therefore an evident need to start developing other types of basic resources.

By contrast, at the Emergent level it is very likely that there are already some e-dictionaries, a spell checker, and even a collection of texts or *corpus*. At this level the language community should aim to expand the range of resources, increase the size of them, as well as to advance in the development of new tools.

The following courses of action can be useful and suitable for both levels of vitality, depending on the specific situation of each language.

R3.1 Dictionary making

A dictionary is a vital resource for a language. At the Dormant level, we can expect that at least one bilingual dictionary has been published. At the Emerging level, there is probably one or more dictionaries, be it bilingual or monolingual, and digital versions may also be accessible to be consulted on the internet.

Depending on the specific situation of the language, different actions can be carried out either to create new dictionaries or to enrich existing ones.

Building a dictionary is a task for experts. Until recently, printed and digital dictionaries have been, and most are still, made by lexicographers or, in the case of specialised dictionaries, by terminologists and domain experts. With the advent of the Web 2.0 paradigm, and especially after the success of Wikipedia, the idea of collaborative dictionary making is also progressing. In collaborative dictionaries, users enter data as new entries, definitions, equivalents in other languages, for example. This data is sometimes directly published, and normally reviewed by editors; alternatively, the reviewing work may be done before publishing.

Depending on the type of dictionary, its size and the precision at which it intends to describe the language, the work of experts is essential. But, in certain cases, a collaborative dictionary is able to

contribute to improving the digital capacity of a language, especially for minority languages, since they often lack the necessary infrastructure to make large investments.

As an example, any competent speaker of a language can build word lists, for instance using Poly. Poly is open source, modern software to share and learn every language in the world. Poly streamlines the process of creating and sharing dictionaries between any two languages. Speakers of languages without a written standard, including the world's 200 plus sign languages, are supported by native video functionality. That's why Poly is particularly suited for mainly oral languages.

Another option is Openwords, which defines itself as "a foreign language learning app for the world's open data." It is an open source language learning platform offering authoring lessons for students, and also allowing you to focus on lexicon, and prepare lessons to learn vocabulary.

Another project that offers the possibility of participating in the creation of a dictionary is Wiktionary, a multilingual, web-based project to create a free content dictionary. It is collaboratively edited via a wiki, and it is available in 171 languages.

You may also look at Webonary, SIL International's platform for publishing dictionaries online.

For languages lacking a written standard, or for people not feeling confident enough in writing the language, a viable alternative is contributing to an audio dictionary. For example, Lingua Libre and Forvo.

Please refer to the DLDP Training Programme for a list of available programmes and interfaces.

Addressees: individuals, users' groups, collectives.

Examples:

- Getting started with Poly – Wikitongues – Medium
- Openwords tutorial
- Wiktionary tutorial
- Webonary – Dictionaries of the World
- Lingua Libre - Help
- Forvo: the pronunciation dictionary. All the words in the world
- Stimmen fan Fryslân (Voices of Friesland)

Further reading:

- Learning every language in the world with Poly – Wikitongues – Medium
- Dictionary making in endangered speech communities
- How would you start a User Generated Dictionary for Minority Languages?

Related module in TP: 6

R3.2 Spell Checker

In a digital environment, a spell checker is an essential tool. The use of automatic spell checking is widespread both in classic office applications and web environments, in the Cloud, and in apps for smartphones.

The construction of an automatic corrector is a task that requires skills at a computer and linguistic level. On the other hand, its complexity depends largely on the type of language. At a basic level, a spell checker uses a word list as a reference for acceptable spellings, is able to mark the words in the text that don't match any word in that list, and to propose to users some correct optional words as substitutes for the incorrect one.

Thus, as a starting point, we need a dictionary. The problem is that usually not all the word forms used in texts are dictionary entries. For example, plural and gender variants and verb forms. You may include those forms in the list, but in languages with a highly inflected and productive morphology, this would not solve the problem.

One of the solutions on offer is Hunspell, an spell checker and a morphological analyzer designed for languages with a rich morphology and complex compound word formation, designed in principle for the Hungarian language. Hunspell is free software, distributed under the terms of a GPL, LGPL and MPL tri-license. Hunspell is used in a large variety of software applications, e.g. LibreOffice, OpenOffice, Mozilla Firefox, Thunderbird and Google Chrome, and it is also used by proprietary software packages, like macOS, InDesign, memoQ, Opera and SDL Trados.

In order to analyze and correct a text in your language, Hunspell requires two files: a dictionary containing words and applicable flags (.dic), and a set of morphological rules (.aff), which specifies how these flags will control spell checking, for example the prefixes and suffixes that each type of dictionary entry can admit.

But, before embarking on the construction of this type of resource, you will need to check if there is a version for your language. Maybe there are resources for spell-checking in your language, but these may have not been installed or activated in your software or app. Hunspell dictionaries can be downloaded from here. If you find a package for your language it is not difficult to start using it; we have provided some examples below.

Addressees: research groups, software developers.

Examples:

- How to install downloaded dictionary file on LO - Ask LibreOffice
- Add and remove Hunspell dictionaries in different languages in InDesign
- I have a hunspell dictionary in my language I want to add, how do I do this?
- hunspell - format of Hunspell dictionaries and affix files
- Creating a spell check dictionary add-on - Mozilla | MDN

Further reading:

- Xuxen, a spell checker for Basque
- What is the best way to go about developing a spell checker for a morphological rich and agglutinative language?

Related module in TP: 6

R3.3 Start up the corpus experience

In a language at the Dormant or Emergent level of digital vitality, it is very likely that there is no corpus.

A corpus is basically a collection of texts that has been built with the goal of being representative of language use, either in general (a reference corpus) or in a given domain, genre, register, or time (a specialised corpus). Today, in 2018, it is normal for these texts to be in a digital format. At the basic level a corpus contains raw text. From there metadata or linguistic annotations may be added to enrich the corpus and increase its usefulness.

Corpora contain a large amount of information that can be used or consulted for different purposes. From a corpus we can learn a lot of things about our language. For example, new words, what are the most used words or variants of a word, which are the most used adjectives with a given noun, and so on. That information is useful, for example, to edit a dictionary. Other purposes can

be the development of language resources and tools (e.g. spell checkers, machine translators), and testing linguistic tools.

Depending on the objectives and the scope, building a corpus is a substantial task which requires a collective effort, and will demand a considerable investment of time and funds. An initial design phase is usually desirable and some texts must be digitalized and revised.

Most RMLs may have difficult at this stage. The main problem is usually how to get the texts into a digital format and for this there are some options that can help us overcome this obstacle. Nevertheless, before launching an initiative to build a corpus we recommend visiting the site An Crúbadán – Corpus Building for Minority Languages.

One possibility is to use work already available on the internet where texts have been digitized. But the viability of this web-based approach, also known as *web as corpus* approach, depends on the amount of web content in the RML, and our capacity to find and collect it (for instance, we must identify the pages written in our target language). Thus, it could be more realistic to consider this approach for a higher level of vitality. You can find more detailed information in the recommendation for level 4, Developing, Increase corpus size and diversity.

Another option is to promote the creation of text repositories. This approach is closely related to the recommendation Create or feed a web-based archive of documents and recordings in the section Digital presence and use. The repository of Nenek project is a good example.

Finally, publicly available RML social media messages and posts may be useful. This type of data can be incorporated into a corpus with texts crawled from the web, and, in addition, can be used to analyse how uniform and coherent the different variants of the same language are. One mandatory step before embarking on this course is to check carefully the license of the data. This action has been more developed and detailed in the recommendation for level 4, Developing, Collect publicly available linguistic data from social media.

Addressees: individuals, collectives, research groups, organisations.

Examples:

- An Crúbadán – Corpus Building for Minority Languages
- BootCaT front-end tutorial
- TextSTAT 2.7 User's Guide
- The Corpus of Facebook Welsh Language Texts
- indigenoustweets.com

Further reading:

- History of Corpus Linguistics
- Corpus Linguistics: The Basics
- Nenek – A cloud-based collaboration platform for the management of Amerindian resource languages
- Kilgarriff, A., Reddy, S., Pomikálek, J. & Avinesh, P.V.S. (2010). A Corpus Factory for many languages . In *Proceedings of LREC 2010, Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta.
- Honeycutt, C. & Cunliffe, D. (2010). The use of the Welsh language on Facebook: An initial investigation. In *Information, Communication & Society*, 13(2), pp.226-248.

Related module in TP: 6

R3.4 Use tools such as concordancers for corpus querying

As mentioned, it is very important for a language to have resources such as text corpora.

There are some simple but powerful tools, concordancers, which could be very helpful to obtain information about words and their use, and that can be used with raw text, that is to say, with text that has not been analyzed to add linguistic information about the words. You can get information about the frequency of words, the most frequent combinations or collocations, and use the results for dictionary making and other language processing tasks.

Nevertheless, in order to exploit this kind of resource in the most efficient way, the ideal situation is to have those texts linguistically processed. A word or *lemma* can occur in different forms or *tokens*, as different genre forms, plural forms, irregular verbs, or, in some languages, inflectional forms. Thus, to obtain quality data, it is recommendable to have the text analyzed, at least at a basic level, such as knowing which lemma corresponds to each word in the text, and what is its part-of-speech. Taking a step further, sometimes morphological analysis is almost a must, as it is usually the case that RMLs are morphologically rich and often this compounds the data sparsity issue.

A part-of-speech tagger, or POS tagger, is a tool to perform this task automatically, and it is an important milestone in the technological development of a language. If this kind of tool exists in your language it is recommended to use it, and to browse the corpus once it has been tagged. If this is not the case, it is highly recommended to promote initiatives for the development of this technology. In the recommendation R4.4 Develop a part-of-speech tagger for_level_4 , Developing, you will find detailed information on this topic.

Addressees: individuals, collectives, research groups.

Examples:

» AntConc 3.4.0 Tutorial 1: Getting Started – YouTube
» TextSTAT – Tutorial - YouTube

Further reading:

» AntConc: A freeware corpus analysis toolkit for concordancing and text analysis
» TextSTAT – Simple Text Analysis Tool
» What are the most useful programs for forming text corpus or dictionary?

Related module in TP: 6

R4 Develop intermediate and advanced language resources

Suitable for: *4 Developing*

At this level, it is likely that the basic resources are already available to users, and that some intermediate resources have already been developed. At this point it's time to complete the set of intermediate resources and to look toward the development of advanced resources and tools.

R4.1 Dictionary making: diversity, size, specialization and dissemination

At the Developing level, there is likely to be more than one electronic dictionary available, most of which will be bilingual, and these are often accessible online.

Depending on the specific situation of the language, different actions can be carried out either to create new dictionaries or to enrich existing ones.

The recommendations can be articulated around these four key words: diversity, size, specialization and dissemination. In other words, a language at the Developing level needs bilingual and monolingual dictionaries, of different size and coberture, and intended for different types of users and purposes (as students, learners, translators, or domain specialists). In addition, it is vital that these dictionaries are accessible on the internet and, especially, have apps for consultation. Successful experiences such as Euskalbar and Hiztegiapp, which allow the user to consult practically all the monolingual and bilingual Basque dictionaries available on the Internet, as well as the corpora, can be inspiring.

Building a dictionary is a task for experts, but a competent speaker of a language can participate at least in some tasks on a dictionary making project. Nowadays technology offers the possibility of creating, editing and publishing dictionaries much more easily than, say, twenty years ago.

However, some types of projects require human, technological and financial infrastructure, which may be complicated for minority languages. For this reason projects have traditionally been carried out with the participation of or promotion by institutions, or by private publishers, with a business model based on the sale of printed versions.

All that is changing at great speed as the sale of printed dictionaries decreases from year to year, so much so that it is deemed as unrealistic to think that the market can now finance the publication of a dictionary, let alone a monolingual one, in a minority language.

Thus, public support, crowdfunding and collaborative work are the basis for dictionary making, more so in the case of minority languages.

A first recommendation is, therefore, launching initiatives to involve institutions and groups of speakers in such projects.

Below are some references for tools that can facilitate the work of preparing and publishing dictionaries.

Addresses: collectives, companies, institutions.

Examples:

Lexicographic dictionaries (monolingual or bilingual):
- FieldWorks (FLEx Quick Tour on Vimeo)
- Lexonomy (Gentle introduction to Lexonomy)
- Tshwanalex (TLex Suite: Dictionary Compilation Software)

Specialized, technical or terminological projects:
- TermWiki In My Language | TermWiki.com
- TermKate - An integral web platform for the creation and publishing of terminological dictionaries

Dictionary apps:
- SIL Dictionary app builder

Further reading:
- Kroskrity, P.V. (2015). Designing a dictionary for an endangered language community: Lexicographical deliberations, language ideological clarifications. University of Hawaii Press.
- Garret, A. (2018). Online dictionaries for language revitalization, to appear in The Routledge handbook of language revitalization, edited by Leanne Hinton, Leena Huss, and Gerald Roche (Routledge, 2018).
- Kotorova, E. (2016). Dictionary for a Minority Language: the Case of Ket. In *Proceedings of the XVII EURALEX International Congress*.

Related module in TP: 6

R4.2 Increase corpus size and diversity

In a language at the Developing level of digital vitality, there are certainly one or more corpus resources. The next step is to aim for a corpus of considerable size, and also to begin preparing the construction of bilingual or parallel corpora, and specialised corpora.

Although in a broad sense any collection of texts could, in principle, be considered a corpus, it is commonly accepted that there are certain conditions for this. In the first place the objective of being useful to obtain data about the use of the language in general (reference corpus), or about the use in a certain domain, genre, register or time (specialized corpus). This objective is one of the most discussed issues in corpus design, and is related to the size and representativeness of the sample, or, at least, to the balance between the different types of text included. Secondly, today it is highly unlikely that a corpus would not be in a digital format. Finally, at the basic level, a corpus contains raw text, but, especially in the case of languages with rich morphology, it is usually considered almost essential that the corpus is linguistically annotated, including, for example, information on the lemma, part of speech and case of each token. Recommendations for corpus design and building can be found at Corpus design criteria (Atkins et al., 1992) and Corpus Linguistics: corpus building principles (Markus Dickinson, 2015); for encoding, see 15 Language Corpora - The TEI Guidelines - Text Encoding Initiative.

A corpus with these characteristics is a rich source of information that can be used or consulted for different purposes: linguistic analysis of language use, development of other resources and tools, testing of hypotheses, and training of tools for automatic language processing. For example, dictionary making has benefitted a lot from corpus analysis: we can learn about new words, what are the most used words or variants of a word, or which are the most used adjectives with a given noun.

However, the fulfilment of these conditions, especially the first one, can be a challenge even for majority languages, which usually are not poorly resourced. Building such a corpus is a substantial task which requires a collective effort, and demands an investment of time and funds that may be beyond some RML community's means.

One of the main problems is how to get a considerable amount of texts in digital format. Dealing with publishers and authors can become an arduous task, and works are not always available in a format that can be incorporated into the corpus at low cost. A possibility that can help us overcome this obstacle is to adopt a web-based approach, or web as corpus approach. On the internet, texts are already digitized, but the viability of this strategy depends not only on the amount of web content in the RML, but also on our capacity to find and collect it. For instance, we must identify the pages written in our target language. If you don't have the means to develop a tool for that, there are some tools that facilitate that task, such as TextSTAT and BootCaT. The first one uses a *crawling* system, and you must specify the URL you want to catch. In order to identify which websites has content in the language you are working with it can be helpful to ask the language community to offer URLs of known existing language specific sites. BootCat offers an additional functionality: you can specify a set of seed words, and, using Bing as a search engine the tool combines those words to perform several searches and collect the documents retrieved. You can use words that are exclusive to your language or that belong to a given domain (to build domain-specific corpora).

A complementary way to collect are text repositories. This approach is closely related to the recommendation Create a web-based archive of documents and recordings in the section Digital presence and use.

When talking about the considerable size for a monolingual and general corpus of a RML, we could take 100 million words as a desirable challenge. Although smaller corpora have been used in RML studies and projects, and we could hardly question their usefulness, there is a strong reason to insist on the need for the corpus to be large. The quality of data-driven NLP tools that are trained on linguistic data, often in the form of either raw or annotated text, depends largely on the size of the corpus. Corpora are also useful for evaluating language tools on gold standard versions, which are human revised and extremely valuable.

Parallel corpora are collections of texts which are translations of each other (actually, a source text and its translation). Parallel corpora are a very valuable resource for translators, but also for bilingual dictionary making, as we can know how a given word, expression or term has been translated. The most basic type of parallel corpus is a collection of bilingual documents. If we know which documents are a translation of each other, then we can say that the corpus is aligned at document level. It is far more useful to obtain parallel corpora aligned at sentence level. An ideal procedure to build them is to take translation memories (TM) as a starting point. TMs are the result of the human translation process when a Computer-Assisted Translation (CAT) tool is used. Otherwise, we can aligned bilingual documents automatically using tools such as Wordfast Aligner or hunalign – sentence aligner. Another possibility is to detect bilingual texts in the internet. This is not a trivial task but there are some experiences in RMLs that can be inspiring, for instance, Bitextor, ILSP Focused Crawler and PaCo2. Another possible source of parallel corpora are the sections of articles from different languages' wikipedias which are translations of each other. Wikipedias are not identical. Strictly speaking they are considered as *comparable corpora*. But we know that they share some content, and some techniques have been developed to find and align them.

Finally, specialised corpora are a reliable source of data to make specialized or terminological dictionaries. As in the case of general corpora, the traditional way to build specialised corpora is to collect texts and documents belonging to a given domain, but, we can also choose a *web as corpus* approach which has proven to be quite an adequate source of data (see, for example, Gurrutxaga et al, 2010). To build a domain-specialised corpus from the web the *seed words* approach can be very useful. BootCaT offers this possibility.

Addressees: collectives, research groups, organisations.

Examples:

- BootCaT – Simple Utilities to Bootstrap Corpora and Terms from the Web || BootCaT front-end tutorial
- TextSTAT – Simple Text Analysis Tool
- Building your own corpus – TextSTAT and AntConc
- ILSP Focused Crawler - ILSP NLP
- Bitextor: the automatic bitext generator
- hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene
- W2c- large multilingual corpus | Martin Majlis - Academia.edu
- TenTen Corpus Family – Wikipedia

Further reading:

- Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, *7*(1), 1-16.
- Gurrutxaga, A., Leturia, I., Pociello, E., San Vicente, I. & Saralegi, X., (2010). Exploiting the Internet to build language resources for less resourced languages. In *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages LREC 2010,* Valetta, Malta.
- Lynn, T., Foster, J., Dras, M. & Dhonnchadha, E.U. (2012). Active Learning and the Irish Treebank, In *Proceedings of the Australasian Language Technology Association Workshop 2012* (pp. 23-32), Dunedin, NZ.
- San Vicente, I. & Manterola, I. (2012). PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web. In *LREC* (pp. 1-6).
- Smith, J.R., Quirk, C. & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 403-411). Association for Computational Linguistics.

Related module in TP: 6

R4.3 Collect publicly available linguistic data from social media

The DLDP recommends making full use of publicly available RML social media messages and posts. Especially on Facebook there are dedicated public pages that promote discussions about cultural and language issues related to the minority language, with people naming objects in a picture with the aim to compare if everyone is using the same word or people sharing poems and songs.

There are different routes that can be followed when data from social media are taken into account:

- A corpus with texts crawled from the web could be integrated with data from social media;
- A dataset of textual data from social media can be used to analyse how uniform and coherent are the different variant of the same language;
- The same dataset can be used by developers to build software that handle lemmatization or annotation at the level of part of speeches.

One mandatory step before embarking in this challenge would be to check carefully the license of the data.

This type of text is often considered noisy as it is unedited user-generated content that can contain misspellings or be ungrammatical. Those texts would constitute a specific type of corpus based on this linguistic nature, but it is useful for other purposes (analysing evolution of language use over time, snapshot of language use in real time, analysis of code-switching and other interesting linguistic phenomenon).

Addressees: research groups, software developers.

Examples:

- The Corpus of Facebook Welsh Language Texts
- indigenoustweets.com

Further reading:

- Honeycutt, C. & Cunliffe, D. (2010). The use of the Welsh language on Facebook: An initial investigation. In *Information, Communication & Society*, 13(2), pp.226-248.
- Lynn, T., Scannell, K. & Maguire, E. (2015). Minority language Twitter: Part-of-speech tagging and analysis of Irish tweets. In *ACL-IJCNLP 2015*, p.1.
- Lackaff, D. & Moner, W. J. (2016). Local languages, global networks: Mobile design for minority language users. In *Proceedings of the 34th Annual International Conference on the Design of Communication* (SIGDOC '16)

Related module in TP: 6

R4.4 Develop a part-of-speech tagger

As already mentioned, it is very important for a language to have resources such as text corpora.

In order to exploit this kind of resource in the most efficient way, the ideal situation is to have those texts linguistically processed. A word or *lemma* can occur in different forms or *tokens*, as different genre forms, plural forms, irregular verbs, or, in some languages, inflectional forms. Thus, to obtain quality data, it is recommended to have the text analyzed, at least at a basic level, such as knowing which lemma corresponds to each word in the text, and what is its part-of-speech. A part-of-speech tagger, or POS tagger, is a tool to perform this task automatically, and it is an important milestone in the technological development of a language. There are some downstream applications that benefit (e.g. term and collocation extraction, named entity recognizers, parsers, grammar checkers, information retrieval, etc). Besides, POS taggers and morphological analysers are an important part of some types of machine translation systems. If this kind of tool exists in your lan-

guage, it is recommended to use it and to browse the corpus once it has been tagged. If this is not the case it is highly recommended to promote initiatives for the development of this technology.

Addressees: research groups, software developers.

Examples:

- Tagging Occitan using French and Castillan Tree Tagger
- Uí Dhonnchadha, E. (2009). Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar. Doctoral dissertation, Dublin City University.
- Minority language Twitter: Part-of-speech tagging and analysis of Irish tweets
- Welsh Parts-of-Speech tagger API
- Apertium - Morphological analysis
- EUSTAGGER: Lemmatizer/tagger for Basque

Further reading:

- Learning POS Tagging & Chunking in NLP – GreyAtom – Medium
- Part of Speech Tagging with NLTK
- A Good POS Tagger in About 200 Lines of Python - Matthew Honnibal

Related module in TP: 6

R4.5 Use tools for corpus analysis and feed your dictionary with data about language in use

There are some simple but powerful tools, concordancers, which could be very helpful to obtain information about words and their use, and that can be used with raw text, that is to say, with text that has not been analyzed to add linguistic information about the words. You can get information about the frequency of words and the most frequent combinations or *collocations*, and use the results for dictionary making and other language processing tasks. More detailed information and references about these types of tools have been already provided in the recommendation R4.3 Use tools such as concordancers for corpus querying, for levels Dormant and Emergent.

But if you have a POS tagger in your language the potential for obtaining data about the use of the language is much greater. You can get the frequencies of use of not only word forms in the text (*tokens*) but dictionary entries (*lemmas*), find new words not included in the dictionary, or investigate how words are combined to form expressions, collocations, complex terms. To do this, you need a corpus query system, or tool for corpus analysis (in corpus-analysis.com, you will find several tools of this kind).

Addressees: research groups, software developers.

Examples:

- Corpkit: a tool for investigating text
- The IMS Open Corpus Workbench (CWB)
- NoSketch Engine
- BlackLab - An open source corpus search engine
- Unitex - GramLab

Further reading:

- The IMS Open Corpus Workbench (CWB) Corpus Encoding Tutorial
- 5 Reasons: Anke Luedeling on "Corpus Linguistics"

Related module in TP: 6

R4.6 First steps toward speech synthesis and recognition

Speech synthesis is the computer-generated simulation of human speech. It is one of the two main tasks of speech technology, the other being speech recognition. When a language begins to enter the world of these technologies it is normal to begin with the synthesis because it is a task that is easier than speech recognition.

Developing speech technologies opens a range of possibilities for applications among which there are notably systems for people with disabilities (screen readers for people with limited vision, speech synthesizers for people with speech impairment, , or recognition systems for people with disabilities that preclude using conventional computer input devices). On the other hand, speech synthesisers have proven extremely valuable for online learning when learners do not have access to native speakers. For instance, this has been particularly effective for overseas learners of Irish who may find the Irish orthography difficult to pronounce. In addition, synthesis is necessary, together with recognition, as a component of a system in which the user can communicate using natural language with the computer or mobile devices.

The development of speech technologies is a technically advanced task. Therefore, the recipients of this recommendation are research groups and software developers. However, certain synthesis technologies (such as concatenative synthesis) use real speech samples and at this point it is important to have the participation of speakers who can provide locutions to create a speech database, or even contribute with their own voice.

Addressees: research groups, software developers.

Examples:

- The MARY Text-to-Speech System (MaryTTS)
- Festcat - Catalan Speech Synthesis
- eSpeakNG
- abair.ie – The Irish Language Synthesiser
- Aholab TTS
- The Welsh National Language Technologies Portal - Speech
- Mozilla Common Voice

Further reading:

- Speech Synthesis for Minority Languages: A Case Study on Scottish Gaelic
- Issues in Porting TTS to Minority Languages
- Frisian TTS, an example of bootstrapping TTS for minority languages
- Chasaide, A. N., Chiaráin, N. N., Wendler, C., Berthelsen, H., Murphy, A., & Gobl, C. (2017). The abair initiative: Bringing spoken irish into the digital space. In *Proceedings of Interspeech 2017*.

3. Digital Presence and Use

Main Course of Action: Promote use and content creation and sharing

- » Find and try ways to encourage people to use their RML in private e-communication and social media;
- » Promote the creation of these types of contents: web pages and websites, blogs, forums, but also Internet radio and TV;
- » Initiatives for uploading and sharing media in RMLs;
- » Crowdsourcing subtitling;
- » Wikipedia: create, edit, correct, update.

3.1 Use for E-Communication

R5 Estimating the value of RML use for interpersonal communication

Suitable for: *2 Dormant, 3 Emergent, 4 Developing*

Regarding digital interpersonal communication among people there are different types of e-communication. Instant messaging, for example, shares many features with orality. Thus, in a digital environment, a language that is not used in e-communication is a language that is not spoken. Another important aspect of e-communication is its use in more formal registers such as as e-mailing in a professional context.

Addressees: individuals, collectives, organizations, institutions.

Examples:

- » Tribo, a loita normalizadora e cotiá entre os máis pequenos
- » Campaign to use Welsh in the European Football Cup
- » Mintzanet: an Option for Practicing the Basque Language Anywhere in the World

3.2 Use on Social Media

R6 Visualizing the value of RML use on social media

Suitable for: *2 Dormant, 3 Emergent*

Social media represents an important new domain with the potential to impact positively on the use of a minority language. As such, they are a crucial indicator of the digital vitality of a language. Social media has the power of turning everyone into content creators.

In the case of a language at the Emergent level of digital vitality there may already be an incipient use of language in a social network. You can take advantage of that reality and use it to further boost the use of your language in that network, and to encourage speakers to introduce the use of the language in networks where it is not very present.

The key message here is the ability to connect individuals or groups who previously would have had no connection or who are otherwise dispersed geographically. Increasing the networks online increases the presence and the strength of the language. Normalising its use within these forums encourages others to do so.

Besides, social media provides a more relaxed environment for learners of a language. The informal settings allows learners to feel comfortable with their level and to be more confident in exchanging with other speakers.

This is what happened to Luxembourgish. Luxembourgish is a spoken language on a small territory and the online community became a huge, connected community. All Luxembourgers or those who have relations with Luxembourg all over the globe started to communicate in Luxembourgish on the internet. The community became much, much bigger than its linguistic space was, and because the spelling and the grammar would often vary in personal messages, this led the government to launch a new initiative to reframe the way Luxembourgish is written. Read more at: Reading: 'Endangered languages could be saved by the internet'.

Another dimension of social media is that should be exploited as much as possible to promote language related activities that are happening locally or nationally. Correct use (strategic marketing strategies) can increase the awareness of organised events, meet-up groups and so on.

Addressees: individuals, collectives, organisations (everyone using social media for communication).

Examples:

- Language use on social media project (Language use of Frisian bilingual teenagers on social media)
- Facebook as a potential site for non-standard Breton
- How social media breathes life into the Irish language | Teresa Lynn | TEDxFulbrightDublin
- A simple proposal for increasing social media use by language learners

Further reading:

- BBC News - Micro-blogging in a mother tongue on Twitter
- What motivates language users to write in minority languages in social media, and what kinds of challenges and opportunities do they meet?
- Jones, A. (2015). Social Media for Informal Minority Language Learning: Exploring Welsh Learners' Practices. In *Journal of Interactive Media in Education*. 2015(1), p. Art. 7
- Bhroin, N.N. (2013). Small Pieces in a Social Innovation Puzzle? Exploring the Motivations of Minority Language Users in Social Media In, Storsul T. & Krumsvik, A.H. (2013). Media Innovations: A Multidisciplinary Study of Change. Göteborg: Nordicom pp. 219-238
- Jones, E. H. G. & Uribe-Jongbloed, E. (Eds.). (2012). Social media and minority languages: Convergence and the creative industries (Vol. 152). Multilingual Matters.

Related module in TP: 3

3.3 Availability of Internet Media

R7 Increase the amount of content and diversify the types of Internet media

Suitable for: *2 Dormant, 3 Emergent, 4 Developing*

The presence on the internet of content in a given language is a very noticeable indicator of its vitality. The content on the web is accessible to everyone and a language with little content in the form of web pages, blogs, forums, audio or video, lacks visibility and presence in the digital environment.

Some effective measures could be taken to increase the amount of media types already present in the language, and to incorporate new types.

R7.1 Increase the amount of text-type content (websites, blogs, forums)

The presence on the web of this type of content is basic and essential for any language. For most regional and minority languages which have achieved a considerable presence on the Internet, much of the web content belongs to these categories.

Therefore, everything that promotes and facilitates the creation and publication of this type of content will benefit the language and its users.

Nowadays, there are many applications and services that facilitate this task. Among the well-known names we can mention WordPress, Blogger, and Wix, but you may find many more options in the references of the "Further reading" section.

Addressees: individuals, collectives, organisations.

Examples:

- The Maya Tz'utujil initiative: the Maya Tz'utujil initiative is a space on Facebook, Twitter and WordPress to teach, learn, and broadcast the Tz'utujil language, which primarily covers the geographic national area of Guatemala, and includes multiple collaborative partners who have become fully linked to the digital initiative.
- Basque Language on the Web: Making an Impact

Further reading:

- 6 Differences Between Blogging in a Minority Language versus English
- How to Choose the Best Blogging Platform in 2018 (Compared)
- The 16 best free blogging platforms
- Best 10 Free Website Builders | 2018's Best Website Builders

Related module in TP: 5

R7.2 Create or feed a web-based archive of documents and recordings

For a language having very little digital vitality, a web-based archive of recordings or documents can be created. Non-speakers of the language can contribute as well. The resource may prove useful for re-establishing a language, or for drawing interested but insecure speakers to start using the language.

- Create and feed repositories of texts in your language
- Diversify the types of media in your language, including audio and video content
- Contribute to existing repositories, such as Wikitongues

Addressees: collectives, organisations, institutions, individuals.

Examples:

- Indigitization – Toolkit for the Digitization of First Nations Knowledge
- Endangered Languages Documentation Programme (ELDP)
- Greenstone Digital Library Software
- Wikitongues
- BasaBALIWiki (see also Alissa Sterns' message in the TP)
- Yadiko Ukruri initiative 'jitómagaro uai'
- Cultural Codex
- AIKUMA Project - preserving endangered languages

Further reading:

» Nichols, D.M., Witten, I.H., Keegan, T.T., Bainbridge, D. & Dewsnip, M. (2005). Digital libraries and minority languages. In *New Review of Hypermedia and Multimedia*, 11(2), 139-155
» Nenek – A cloud-based collaboration platform for the management of Amerindian resource languages

Related module in TP: 5

R7.3 Stream online using free software tools

An online radio station can be an effective means of spreading a language and making it regain visibility, in particular for dispersed or scattered communities. The availability of free software makes it relatively easy. You can take inspiration from the stories reported below.

Addressees: individuals, collectives, organisations.

Examples:

» Brazil's First Indigenous Online Radio Station Uses Digital Media to Promote Native Languages and Communities
» Meet the Young Ecuadorians Behind the First Kichwa-Language Radio Show in the US

Further reading:

» Icecast, a streaming media (audio/video) server that can be used to create an Internet radio station.
» Top 5 Free Tools to Live Stream Your Event Online – Capterra Blog

Links to TP module: 5

R7.4 Record digital stories in your own language

Digital storytelling is defined as (Barret, 2005):

> "Digital Storytelling is the modern expression of the ancient art of storytelling. Digital stories derive their power by weaving images, music, narrative and voice together, thereby giving deep dimension and vivid color to characters, situations, experiences, and insights."

Digital storytelling is being used, among other areas, in education (Hoven, 2009):

> "Digital storytelling is increasingly being used as a means of encouraging reflective learning and/or assessment in many language and language teacher education programs around the world. Digital storytelling can provide language teachers and learners with rich scope for helping learners to identify with foreign languages and cultures or first languages (L1) and cultures (C1) that are being lost, and to feel more comfortable about using these languages for real purposes."

From that perspective, it can be useful for language revitalization purposes.

Addressees: individuals.

Examples:

» Digital Storytelling: What it is… And… What it is NOT
» How Storytelling Can Do Wonders in Blogging

Further reading:

» Frequently-Asked Questions about Digital Storytelling
» Hoven, D. (2009). Digital Storytelling in Indigenous Education: Internet Technologies to (Re-) Establish L1 and C1 Literacy and Fluency. In *Internet-Based Language Learning: Pedagogies and Technologies*, p.47.

Related module in TP: 5

R7.5 Promote subtitling initiatives

In the digital environment the subtitling of films and videos has benefitted from the possibilities of working collaboratively. There are many projects in which volunteers translate the subtitles of a film or movie into a growing number of languages. This is an opportunity for RMLs. Something that was previously difficult and expensive may now be a reality thanks to the collective work of RML speakers, either to be able to watch films in their original version with subtitles in their RML, or to give more output to original works in the RML, subtitling them in languages of greater diffusion.

Addressees: individuals, collectives.

Examples:

» Amara: Caption, Subtitle and Translate Video
» PerMondo – Introduction to subtitling
» TED translations
» Contribute translated content - YouTube Help

Further reading:

» Crowdsourcing Subtitles for Endangered Languages
» Review: Amara is a Web-based service that lets anyone transcribe and translate online video
» How Crowdsourced Video Translation Works: Webinar Q&A with Amara
» Is crowdsourcing translation a threat or an opportunity for the audiovisual market?
» Azpitituluak, a Project for Basque Subtitles
» Dowling, M., Lynn, T. & Way, A. (2017). A Crowd-sourcing Approach for Translations of Minority Language User-Generated Content. In *Proceedings of 1st Workshop on Social MT*, Prague, Czech Republic.

Related module in TP: 5

3.4 Wikipedia

It is very probable that there is a wikipedia in your language, but it could be small and speakers might either not use it or not be aware of it. Taking into account that Wikipedia is a very indicative resource of the vitality of a language, it is important to make it grow and improve. In the Detailed list of Wikipedias, you can check whether there is a Wikipedia in your language and, if so, its most relevant figures (as the number of articles or active users).

In the case that there is no Wikipedia in the language, the recommendation is to start a project and/or translate the Wikipedia user interface.

R8 Create a Wikipedia in your language

Suitable for: *2 Dormant*

The Wikimedia Foundation establishes a series of criteria and requirements for the creation of a new wikipedia. Before submitting an application, we must consult the page Language proposal policy, where we are informed of the criteria that the foundation follows to consider a proposal eligible, as well as the process for its approval or rejection.

One the requisites for final approval is the translation of Wikipedia interface into the new language. Localisation statistics describe the current availability of translations for the MediaWiki interface into different languages. The translation of the software behind Wikipedia is done on a website called translatewiki.net (see Translating the software that powers Wikipedia).

On the other hand in the page Requests for new languages we can verify if a proposal for our language has already been presented, and at what stage it is.

Once we have decided to request the creation of a new wikipedia, we will follow the instructions given on instructions for adding to new request.

If the requirements for eligibility are met, the language committee should verify it as "eligible". The next step is to begin writing a test project on the Incubator wiki, where potential Wikimedia project wikis in new-language versions can be arranged, written, tested and proven worthy of being hosted by the Wikimedia Foundation. For a full list of wikis on Wikimedia Incubator, see Incubator:Wikis.

Finally, if all requirements have been met and a detailed investigation finds no unresolved problems, the language committee will notify the whole community of pending approval by adding a notice to Talk:Language committee. If no further problems come up the request will be approved and developers will be asked to create the wiki.

R9 Take your Wikipedia to a higher level

R9.1 Promote an initiative to increase the number of wikipedia entries in your language

Suitable for: *3 Emergent*

A typical case of a language at Emergent level would be to have a wikipedia of 10,000 entries approximately. Any increase should be considered an achievement, but, when considering a challenge, reaching 100,000 entries would be a great success.

How to activate users to work on Wikipedia?

» Orientate user's to include local or regional content on the RML Wikipedia. To attract more users, it is important to differentiate your RML Wikipedia as a resource where we will find information that is not found in other wikipedias, or that is more detailed and elaborated. But

that should not result in abandoning the inclusion of articles of a more general nature (such as science, humanities, history, art).

» Engage secondary level schools to educate pupils on wiki editing (class creates a page on your local town/sports club, for example).

» Set up a workshop to train trainers (through the RML), establishing the correct terms. Providing a workshop to new editors in their RML is more enticing than doing so in the alternative major language.

» Identify the most popular wiki pages (in general) and highlight them as priority articles for editors to translate.

» Try to involve local authorities in the promotion of Wikipedia, through campaigns that encourage its use, give support to wikipedians' communities, and help with editing and correction.

Addressees: individuals, collectives.

Examples:

» Les secrets de Wikipedia en breton – Agence Bretagne Presse (in French)
» La version en breton de Wikipedia a atteint 30 000 articles (in French)
» 10th anniversary of Wikipedia in Asturian
» The founders of Wikipedia highlight the Asturian version (English version by Google Translate from the original in Spanish)
» Welsh Wikipedia reaches 100,000 articles

Further reading:

» Oportunidades y retos para el conocimiento libre en lenguas indígenas en Wikipedia
» The keenest Wikipedians – Languages on the internet – The Economist
» Aboriginal language Wikipedia faces cultural hurdles, say researchers
» Crean Wikipedia en Náhuatl; van por otras en Maya y Mixe

Related module in TP: 4

R9.2 Initiatives to increase the size and quality of Wikipedia

Suitable for: *4 Developing*

A wikipedia already exists in the language, most likely of medium size (between 10,000 and 100,000 articles), but taking into account that Wikipedia is a very indicative resource of the vitality of a language, it is important to make it grow in the number and extension of the articles, and improve their quality.

Wikipedia is a resource that includes a wide variety of content types, and, therefore, gives the opportunity to work in different areas and aspects to improve it. We will consider here the followings characteristics relating to the articles:

» **Number:** Any increase should be considered positively, but, when considering a challenge for a Developing language, reaching or even approaching 100,000 entries would also be a big achievement. Needless to say, going beyond this milestone should be considered a success of the first magnitude.

» **Type:** Orientate user's to include local or regional content on the RML Wikipedia. To attract more users, an option is to differentiate your RML Wikipedia as a resource where we will find information that is not found in other wikipedias, or that is more detailed and elaborated. But, especially at this level, that should not result in abandoning the inclusion of articles of a more

general nature, in the areas of science, humanities, history, art, or any other type of content that can be relevant to the users and their daily life (popular pages on celebrities, brands, sports), particularly if we want to encourage and entice the next generation of speakers to use these digital resources. Without this type of information, the language runs the risk of reproducing in Wikipedia the sociolinguistic situation of diglossia that it suffers in society, and cannot contribute to reverse it. It must be a mixed strategy. A search on the most read and contributed to Wikipedia articles will help inform this approach.

» **Length or extension:** For a language at the Developing level of digital vitality it is worth starting to consider the extension of the articles as a parameter of great importance. If we want users who enter Wikipedia looking for encyclopedic information, beyond the mere definition of the entry, do so in the Wikipedia of the language, it is crucial to give the articles a length that can minimally satisfy that need, even if we are aware that they will always find more data in major language Wikipedias.

» **Language quality:** In some Wikipedias of minority languages, there is a wide concern about the linguistic quality of the articles. It is very important to find a balance: if the quality is insufficient, the prestige of Wikipedia in the language may be damaged, but if too much emphasis is placed on demanding a high level of quality, certain users may be discouraged and may not take part. For example, authors of articles for Wikipedia in Sardinian can choose among three varieties (Limba Sarda Comuna, Logudoresu and Campidanesu) and a link to an external spell checker for these varieties is provided, with the suggestion to check the coherence of the text before posting it on Wikipedia.

Ways to involve users in the Wikipedia edition:

» Organize an editathon during local festivals and promote translation projects in education, also as a way to improve digital literacy.

» Engage secondary level schools to educate pupils on wiki editing (class creates a page on your local town/sports club, for example).

» Set up a workshop to train trainers (through the RML), establishing the correct terms. Providing a workshop to new editors in their RML is more enticing than doing so in the alternative major language.

» Identify the most popular wiki pages (in general) and highlight them as priority articles for editors to translate.

» Try to involve local authorities in the promotion of Wikipedia through campaigns that encourage its use, give support to wikipedians' communities, and help with editing and correction.

Addressees: individuals, collectives.

Examples:

» Celtic Knot - Wikipedia Language Conference
» Basque Wikimedians User Group
» Collaboration with Wikipedia in 2016 and 2017
» Catalan Wikipedia - Creation

Further reading:

» Galipedia, the Wikipedia in Galician, is now 15 years old (English version by Google Translate from the original in Spanish)
» The Basque Wikipedia, Local Knowledge Gone Global (and back)
» Wikipedia in Catalan, leader in the 1,000 most important articles (English version by Google Translate from the original in Catalan)

Related module in TP: 4

4. Digital Performance

Main Course of Action: Create opportunities to do things digitally in your language

- » Promote demand of Internet services in RMLs
- » Localisation of software and users interfaces
- » Machine Translation services
- » Obtain a dedicated domain

4.1 Availability of Internet Services

R10 Expand the range of possibilities to use Internet services in your language

Suitable for: *3 Emergent, 4 Developing*

For a language to be digitally operative, it is essential that you can "do things" on the web. We are talking about services such as e-banking, health, shopping, tourism, culture or news.

R10.1 Collect information and experiences from your RML users' community, to determine which are the most important and used services

In the case of a language at the level 3 of vitality, it is typical, for example, the scarcity of health services, e-banking or e-commerce. However, it may not be so in the case of your language.

A language in level 4 is supposed to have some functionality on the internet. If the language has a certain level of official recognition, the public administration is likely to provide online services in the language. One possible line of work is to expand and strengthen that area, but, at the same time, to work so that the private sector is aware of the convenience of addressing its clients and users in the RML. In any case, our action should be oriented towards the needs of the speakers. It is therefore very important to detect which services the speakers would like to use in their RML.

Therefore, before launching an initiative to encourage administration or companies to offer their services in your language, it is necessary to have precise information about the real situation and the needs of users in order to establish some priorities.

One way to obtain this information is to conduct a survey among users. You can take as a starting point the questionnaire developed in the DLDP project and translate it into your language.

Once you have received the results of the survey you can organize a campaign to have the services in your language most wanted by speakers. Most services are under the responsibility of local or central authorities. Get in touch with the people responsible for those services. Ask whether a language officer is available.

In order to make the services responsible aware you should stress the importance of local language for people's well-being and inclusion.

Addressees: collectives, organisations.

Examples:

» The 'Survey on Digital Fitness' | the Digital Language Diversity Project
» Paricio Martín, S.J. & Martínez Cortés, J.P. (2014). El uso del aragonés en Internet y las nuevas tecnologías: herramientas y repercusión. In *Actas II Jornadas Aragonesas de Sociología.* pp. 105-120. Zaragoza, 2014.

Further reading:

» Why language matters for the Millennium Development Goals - Unesco

R10.2 Estimate the value of using users' language in business

Globalization has further highlighted the importance of language in commerce. It is increasingly accepted that online consumers in most countries indicate a higher level of comfort with websites in their own language. This applies to the case of RMLs, even more so if we are aware of the emotional factors that play an undeniable role in the image that a brand seeks to create with consumers. That's why the language factor deserves to be taken into account when designing a business strategy.

Addressees: companies, organisations.

Examples:

» Let languages shout out your business benefits
» The Benefits of Translating into Minority Languages – Translation

Further reading:

» Language Means Business
» The Importance of Language in Global E-Commerce | TransPerfect
» The Value of Language in e-Commerce white paper
» Cunliffe, D., Pearson, N. & Richards, S. (2010). E-commerce and Minority languages: a Welsh perspective. *Language and the Market, Palgrave Macmillan, Basingstoke*, pp.135-147.

R10.3 Develop smartphone apps

Smartphone apps are a popular way of delivering content for a variety of purposes. Given that smartphones are so widespread an app can be a relatively easy way for a language to be appreciated by a wider public.

You should consider promoting the development of free apps for language learning, for instance. In this area there are many beautiful and replicable examples, such as Wahzhazhe (available for iOS and Android), and Speak Mohawk (iOS and Android versions) provided by First Nations, or the Rawang dictionary app.

Making available words, simple phrases and greetings is a fun way to make someone immerse in a language and culture, and will serve as well as an excellent introduction to the uniqueness of the particular culture conveyed by the language. Additional features may include the recording of words and phrases to allow the user to practice the language. Games and quizzes that allow learners to test their abilities are usually well-received. The more immersive and engaging the experience is, the better.

Unfortunately, developing apps is expensive compared to web sites. But there are options for creating *flashcard apps* using existing platforms like Memrise or Anki that require far less investment by the language community. This approach can result in engaging mobile learning content, al-

though perhaps without the prestige of a stand-alone app. Example of Lakota courses on Memrise: https://www.memrise.com/courses/english/lakota/

Addressees: individuals, collectives, organisations, software developers.

Examples:

» How Technology is Saving Native Tribe Languages
» Wahzhazhe, an Osage language app for phones and tablets
» Six Nations school launches app that teaches people to speak Mohawk

4.2 Localised Social Network

R11 Initiatives for localising social media user interfaces

Suitable for: *3 Emergent, 4 Developing*

We have previously mentioned the important role that social media plays in the vitality of a language. Although the user interface may be in your language, it is not necessarily a condition for you to interact using it in social media, but it is a factor that has some influence on the language use and that can help more users start using it.

Unfortunately, in the last few years the options previously provided by companies such as Facebook, Twitter or Google to localise their interfaces have diminished (see below Scannell, 2012).

It is also necessary to emphasize the importance that the quality of the translation has in the prestige of the language, and in the effective value that this has for the users, as well as possibly increasing the use of the language. In addition, it should also be a goal for the translation to be completed.

Addressees: collectives, organizations, institutions.

Examples:

Sa tradutzione de Facebook in sardu

Further reading:

» Losse, K. (2008). Facebook: Achieving quality in a crowd-sourced translation environment
» Scannell, K. (2012). Translating Facebook into endangered languages. In *Proceedings of the 16th Foundation for Endangered Languages Conference* (pp. 106-110).
» Twitter available in Basque language starting today

Related module in TP: 3

4.3 Localised Software: Operating Systems and Basic Software

Having the possibility to use software in your language is a sign of prestige, it indicates you can live digitally in your language, and all that is of great value, as it can encourage users to choose their RML as the language for communication and work.

R12 Start a community effort to localise free software

Suitable for: *3 Emergent*

Most likely a language in an emerging situation will have, at least, one operating system (either desktop or mobile, either open or commercial) localised in the language. If that is the case, these are the three objectives that may be achievable:

- Operating system: Linux or Android (depending on which one is already localised);
- Web browser: Mozilla Firefox
- LibreOffice

Nevertheless, other alternative strategies could be equally convenient: an initial focus on smaller, friendlier user applications might be a good way to begin developing a new localization community. For instance, in Lakota and Cherokee, localising effort has been made on free education and game apps like eduactiv8.

As we have highlighted in the previous section, the quality of the translation and the fact that it is completed or at least considerably advanced is a factor that affects the decision of the users to choose or not the version in RML.

Addressees: collectives, organisations, institutions.

Examples:

How to Localize Software: 10 Dos and Don'ts for a Watertight Software Localization Process
How to localize LibreOffice in your language
mozilla l10n - Mozilla in your language
How To Localize an Android Application

Further reading:

Localization 101: A Beginner's Guide to Software Localization
How to Localize your Software, App or Game : 7 Best Practices
Localizing LibreOffice: A Community Effort To Expand the Benefits of Free Software
Localization at Mozilla

R13 Strengthen initiatives to localise the general purpose free or proprietary software most used in the language community

Suitable for: *4 Developing*

Most likely, a language in an developing situation will have, at least, one desktop and one mobile operating system (either open or commercial) and some general purpose software (a word processor and a browser) localised in the language.

It is time to go beyond these basic applications, and expand the range of possibilities to use software localised in the language, be it free and open source software, or proprietary software.

In order to achieve the greatest efficiency, and focus the effort on localising the most frequently used software, it is important to know the needs of MRL users, as well as to learn from the path taken by more advanced MRLs.

In the case of open source software, you can start a community effort to localise the general purpose free software most used in the language community. Some examples of that kind of initiatives are Softcatalà (Catalan), Meddal (Welsh), Librezale (Basque), An DROUIZIG (Breton), Softaragonés (Aragonese). For the localisation of commercial software (Windows, MacOS, MS office), local authorities play a key role as interlocutors with companies.

Addressees: collectives, organisations, institutions.

Examples:

» How to Localize Software: 10 Dos and Don'ts for a Watertight
» Software Localization Process

Further reading:

» Localization 101: A Beginner's Guide to Software Localization
» How to Localize your Software, App or Game : 7 Best Practices
» How To Localize an Android Application

R14 Consider video games as a valuable revitalisation opportunity

Suitable for: *4 Developing*

Video games offer to users a high level of interactivity and therefore of engagement with a language. Developing or localising video games in local languages is a great way of giving new boost to them, as well as of showing to the younger generation that the language is alive and fit for the modern world.

As it can be imagined the market for video games for lesser used languages is quite limited. Big companies do not invest in small markets and therefore the responsibility is again on the shoulders of activists and enthusiasts, like Gwenn Meynier who translated into Breton the game *Steredenn*, or Frédéric Antonpietri and Fabien Mariani, who created the game *Winterfall* in Corsican. In the Basque Country Game Erauntsia is a group of Basque gamers who campaign for using the Basque language in the video game world.

Examples:

» Steredenn
» Winterfall

Further reading:

» Get inspired by the story of Kisima Ingitchuna, the first video game in Inupiaq language that, in addition, drawn from indigenous culture for its story and characters: Showcasing Alaska's Inupiat culture through gaming
» PES 2016 Introduces the First Welsh Language Video Game Box Art
» The video game: A challenging universe for minority languages
» Conquering digital worlds in Scottish Gaelic

4.4 Machine Translation Services

The availability of machine translation services is considered as a indicator of strong and active digital use, as it presupposes a wide array of tools and resources. From the point of view of the digital usability of the language, the availability of reliable machine translation for a language is a sign that the language has gained a fairly high level of digital presence and importance.

In environments where a minority language must move and try to survive as a tool for communication, certain uses of machine translation can lead to situations that must be handled with care. A system with a minimum of quality can be useful and effective for assimilation, that is, to produce understandable texts, even if they are not suitable for publication (dissemination). But it is difficult, in every language and especially in the case of an RML, to reach the quality required for dissemination without the text needing to be edited by a translator. If this difference is not taken into account there may be an abuse of machine translation at the expense, unfortunately, of the RML itself. This misuse of machine translation has led to some bad experiences (e.g., the case of the Cherokee Wikipedia), and raised some criticism (see M. Bauer's When things are way, way, WAY worse than you thought they might get; M.B. Měchura's Do minority languages need machine translation?). Therefore, it is vital that potential users of MT are aware of these limitations in order to extend the good use of this technology. See, as an example, the Machine Translation Advice Note emphasising this point for Welsh.

In the case of a language at level 3 of digital vitality, it is not surprising that the technology needed to provide automatic translation services has not been developed, or has not reached the level of maturity required. Even so it is convenient to start activating certain mechanisms that prepare the ground to be able to take that step in the future.

At level 4 of digital vitality, it is highly likely that the language has one online service/tool, with at least one language paired with the RML and available in at least one direction.

It is worth emphasizing here the need for the coordination of public administration of bilingual data. Most governments are not aware of the value of bilingual data to help them develop translation tools to help them meet translation needs (if official status in any way). Usually language technologists are required to advise in this respect, and to highlight the need for coordinated centralisation of translations to ensure the re-use of data (*translation memories*) and the easy access of bilingual data for the purposes of building MT systems. If specifically tuned to public administration, reasonable quality output can be achieved; that would then reduce the overall effort of translators (editing repetitious content instead of from scratch).

R15 Pave the way to Google translation through community involvement

Suitable for: *3 Emergent, 4 Developing*

It is very likely that a popular machine translation service such as Google is not yet available for a language, although the number of languages offered by Google is growing steadily. If this is the case, you do not have to give up but instead try to help the process. Google Translate derives its linguistic models from texts available on the Internet. To develop translation models for less widely used languages Google requires input from users and native speakers. This community input is essential for Google Translate to accommodate lesser-used languages.

You can take inspiration from the initiative carried out for Frisian, a minority language of the Netherlands. A number of Frisian organisations teamed up with Google and organised a "translation week" to make available quality translation from English sentences. This global community effort to translate one million words enabled to add Frisian to its online translation service.

Addressees: collectives, organisations, institutions.

Further reading:

» Frisian Google Translate Day in Groningen
» Learning the lingo: Here's how Google Translate copes with even the rarest languages

Related module in TP: 6

R16 Develop and promote at least one MT system to and from the majority language

Suitable for: *3 Emergent*

For a language at level 3 Emergent of vitality, the choices for the development of a machine translation system can be limited by the difficulties in having good parallel corpora to train statistical MT systems. A free or open source rule-based platform such as Apertium allows the development of MT for these languages, replacing the need of huge parallel corpora by reusable resources such as morphological dictionaries, bilingual dictionaries and transfer rules. If languages are close enough, high quality systems can be developed. A parallel advantage of this approach is that some modules can be used as independent tools (POS tagger, morphological analyzer, dictionaries).

Addressees: research groups, software developers.

Examples:

» Apertium | A free/open-source machine translation platform
» Opentrad translator. Open-source translator of texts and documents

Further reading:

» Tyers, F.M., Alòs i Font, H., Fronteddu, G. & Martín-Mor, A. (2017). Rule-Based Machine Translation for the Italian–Sardinian Language Pair. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), pp.221-232.
» Bowker, L. (2009). Can Machine Translation meet the needs of official language minority communities in Canada? A recipient evaluation. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, (8).
» Martínez-Cortés, J.P, O'Regan J. & Tyers F. (2012). Free/Open Source Shallow-Transfer Based Machine Translation for Spanish and Aragonese. In *LREC-2012*. Istanbul (Turkey).

Related module in TP: 6

R17 Expand the number of language pairs; if it is not already, try to make English one of the languages included

Suitable for: *4 Developing*

At the Developing level of vitality, it highly likely that the language has a rule-based MT system. In that case a first option to diversify language pairs is reusing linguistic data and tools (morphological dictionaries, POS taggers) developed for a pair in the development of a system for new pairs. This evidently limits the range of languages to which we can extend our system since they must be to a certain extent close (not too different in terms of morphology and syntax). Nevertheless, rule-based MT systems have also been developed for very different languages, such as Spanish-Basque (Matxin).

On the other hand it may be the case that the language has a not inconsiderable amount of parallel corpora, something that would enable it to enter the field of statistical machine translation. Some recent experiments with RMLs in the area of neural machine translation have given promis-

ing results. Perhaps this might be the way in which automatic translation for RMLs makes a qualitative leap and bring it closer to being able to be used as a daily tool in translation services.

Addressees: research groups, software developers.

Examples:

- Matxin: an open-source transfer machine translation engine
- Automatic translation - Llengua catalana - Gencat.cat
- Oersetter: Frisian-Dutch Statistical Machine Translation
- You can now translate Wikipedia articles from Spanish into Basque, thanks to an open source machine learning tool

Further reading:

- Popović, M., Arcan, M. & Klubička, F. (2016). Language related issues for machine translation between closely related South Slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)* (pp. 43-52).
- Dowling, M., Cassidy, L., Maguire, E., Lynn, T., Srivastava, A. & Judge, J. (2015). Tapadóir: Developing a statistical machine translation engine and associated resources for Irish. In *Proceedings of the The Fourth LRL Workshop: Language Technologies in support of Less-Resourced Languages*, Poznan, Poland.
- Screen, B. (2017). Machine Translation and Welsh: Analysing free Statistical Machine Translation for the professional translation of an under-researched language pair. *Journal of Specialized Translation*, 28.
- Mayor, A., Alegria, I., De Ilarraza, A.D., Labaka, G., Lersundi, M. & Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. *Machine translation*, *25*(1), p.53.
- Gompel, M.V., van den Bosch, A.P.J. and Dijkstra, A. (2014). Oersetter: Frisian-Dutch statistical machine translation. Ljouwert: Fryske Akademy
- Etchegoyhen, T., Martínez, E., Azpeitia, A., Labaka, G., Alegria, I., Cortes, I., Jauregi, A., Ellakuria, I, Martin, M. & Calonge, E. (2018). Neural Machine Translation of Basque. In *21st Annual Conference of the European Association for Machine Translation* (p. 139).

4.5 Dedicated Internet Top-Level Domain

R18 Time to get an internet domain for the language

Suitable for: *4 Developing*

The importance of having a geographic top-level domain (GeoTLD) dedicated to the language has been pointed out by many experts. But, perhaps more important than that, it has been practically proven by some successful experiences.

It is broadly accepted that a GeoLTD helps strengthen the cultural and social identity of interest groups, such as minoritised language communities, and enhances language diversity. Moreover, a dedicated internet domain provides greater visibility to the language, encourages more agents and users to incorporate the use of the language in their usual communications, and it becomes an evident element of revitalization. The cases of Wales, Catalonia, Brittany and the Basque Country are clear evidence of the wide set of advantages that a minority language can obtain thanks to a dedicated internet domain.

Reaching this goal is by no means an effortless task. The communities that have obtained their own GeoTLD have had to go a long way to meet the requirements set by the Internet Corporation for Assigned Names and Numbers (ICANN) for its concession.

Before applying for a new GeoTLD within the New gTLD Program, we must check whether there is a domain in the language, or an application to obtain it has been submitted. ICANN provides this information at New gTLD Geographic Applications.

Addressees: collectives, organisations, institutions.

Examples:

- GeoTLD Group AISBL - Promoting local digital Identities for Cities, Regions, Languages and Cultures on the Internet
- Fundació punt CAT
- Point BZH - Pik BZH - L'extension internet de la Bretagne
- PuntuEUS

Further reading:

- The Case for a TLD for Wales
- Interview : David Lesvenan, Président de l'association www.bzh
- EU urged to help expand internet domain names in different languages
- The Basque .eus Big Bang
- Cultural diversity in cyberspace: the Catalan campaign to win the new .cat top level domain
- Why a local domain might be the best choice for your business
- The internet domain name system and the right to culture