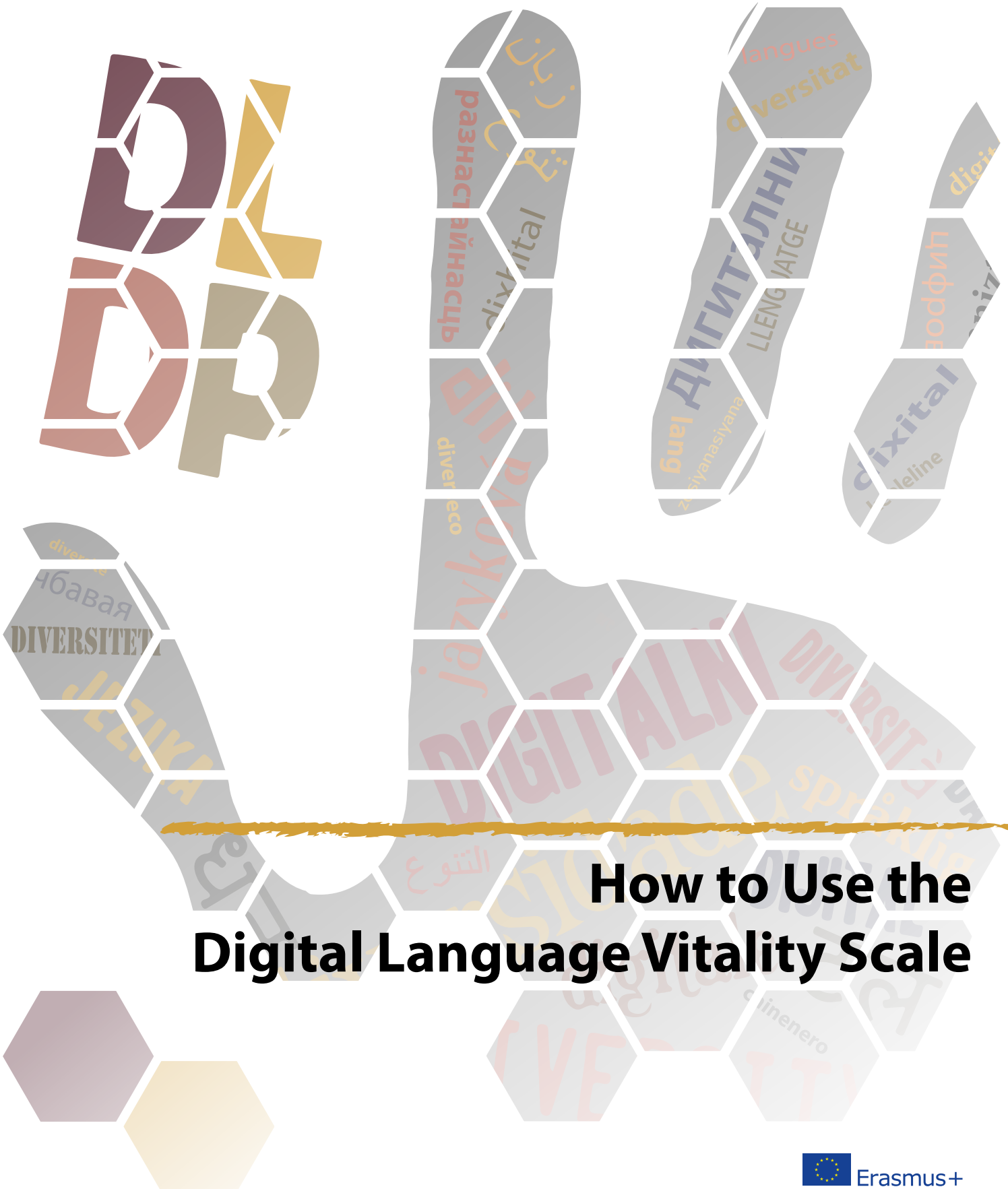
The Digital Language Diversity Project



How to Use the Digital Language Vitality Scale

Erasmus+

Table of Contents

Introduction **3**
Who should use the scale 3
What information is preliminarily required 3
What the scale can be used for 3
The DLDP Scale for Digital Language Vitality **4**
How to apply the scale **5**
Digital Capacity 5
Evidence of connectivity 5
Digital literacy 6
Internet penetration or digital population size 7
Character encoding and input/output methods 7
Availability of language resources 8
Measuring the availability of digital language resources 9
Digital presence and use 9
Use for e-communication 10
How to assess use for e-communication 10
Use on social media 10
How to assess use on social media 11
Availability of Internet media 11
How to assess availability of Internet media 12
Wikipedia 12
Digital Performance 13
Availability of Internet services 13
How to assess availability of Internet services 14
Localised social networks 14
How to measure localised social networks 15
Localised software: operating systems and basic software 15
How to check for localised software 15
Machine translation tools/services 16
How to check for availability of MT services 17
Dedicated Internet top-level domain 17
How to check for the existence of a dedicated Internet top-level domain 18

Imprint

How to Use the Digital Language Vitality Scale
Authors: Klara Ceberio, Antton Gurrutxaga, Claudia Soria, Irene Russo, Valeria Quochi

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.

Introduction

The Digital Language Vitality Scale is an instrument developed within the framework of the Digital Language Diversity Project (www.dldp.eu) for estimating the degree of digital vitality of any given language. It aims to be an instrument for self-assessment of the digital vitality of any language, although it is aimed in particular at identifying current gaps, needs and requirements regarding the extent to which a language community is active/vital on digital media and devices so that adequate digital language planning can be done.

Ideally, the scale contains reliable indicators that should be measured objectively. In practice, this is not always possible. Hence, the purpose of this document is to provide some guidelines on how to apply and measure or estimate the indicators included on the scale in practice, in particular indicating what kinds of sources of information are to be taken into account depending on the indicator and on the particular situation under scrutiny.

Who should use the scale

The scale is a tool for community assessment of the digital vitality of any given language. It can be used either by individuals or by groups, provided that the information required is available.

What information is preliminarily required

Most of the information needed for applying the scale should be available to any person having a deep knowledge of the sociolinguistic situation of the language investigated. For this reason, we recommend that the scale is applied as a result of teamwork, and on the basis of shared and agreed upon evidence.

Some basic knowledge of the Internet and related issues is required. However, we have tried our best to indicate reliable sources of information for every aspect that is taken into account by the scale.

What the scale can be used for

It is expected that the scale will be used mostly for assessing languages - and regional/minority languages in particular - by institutions, associations, governmental offices, and /or scholars working on or interested in language preservation or revitalization.

The current digital vitality status of a language will serve as a baseline for making informed decisions regarding the digital development of that language. The particular types of actions and measures needed will be chosen by the language community, and in this respect the DLDP project wishes to act as an external consultant providing guidance and expertise about the range of possible actions to be taken. These actions are detailed in the DLDP Digital Language Survival Kits, a collection of recommendations applicable for the various digital vitality levels identified by the DLVS.

Therefore, the Digital Language Vitality Scale is the first but necessary step in digital language planning, a process - we stress it once again - that must be community-based and rooted in the community's vision of what is desirable and achievable.

The DLDP Scale for Digital Language Vitality

The DLDP Digital Language Vitality Scale has six levels.

Pre-digital (PD): The language is not present on (online) digital media and it lacks most basic preconditions for digital use. It has no technological support; the infrastructure for connectivity is limited or too expensive for the average person, therefore the language cannot expand on the Internet; people's digital competence is non-existent or very low.

Dormant (D): The language is not present on the Internet although some of the main pre-conditions for digital usage are in place: e.g. connectivity is ensured, there is some degree of Internet penetration and most language speakers are at least basically digitally literate. However, there is no technological support for the use of the language (e.g. there is no keyboard support for writing the language, no t9, auto-correction, etc.), especially online. Therefore, although speakers are in principle capable of using the language digitally, in practice they do not do so.

Emergent (E): The language starts to be used digitally. Internet penetration is good, speakers are digitally literate. Overall, however the language enjoys limited technological support (e.g. such as fonts and keyboards), perhaps a few (digital) language resources (such as e-dictionaries and text collections) might be available. Texting and messaging as well as social media start to be used in the language, albeit not yet extensively. Wikipedia, if present, is small.

Developing (De): The language is visible on the Internet and is used over communication and social media, although frequency may still be occasional. Some digital media and services may be available, as well as a Wikipedia; basic (electronic) language resources exist, 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.

Vital (V): The language is highly present on the Internet, and is used regularly for e-communication and on social media, some of which may have a localised interface. There is a considerable variety of digital media available. Language resources are widely available. Wikipedia projects are big and actively used/participated. The language can be used in all digital domains. Most used operating systems and general purpose software are localised in the language. There is evidence of machine translation tools/services.

Thriving (T): The language is pervasive on the Internet and is used extensively and without any technological barriers in all current digital domains, from communicative to transactional ones. The latest technology is available.

How to apply the scale

Placing a language at one of the levels above requires to be able to assess several factors, called indicators. We distinguish among three groups of indicators: a group pertaining to a language digital capacity, a group related to a language digital presence and use, and a group related to a language digital performance. From the interplay of these axes we can derive a picture of the extent to which a language is able to perform in the digital domain.

In the following, we provide some useful indication on how to practically assign the relevant score to the indicators: where to find the relevant information, how to interpret/use it, and if there is no reliable source of information available we provide a suggestion on how to approximate the estimate. A general recommendation here is that before applying the scale you should do some ad-hoc research to discover whether, for each indicator, there is some new reliable initiative/information source or any specific one dedicated to your language (i.e. the language you want to assess). This is necessary because the situation may change fast in this domain and some speaking communities may have set up specific digital observatories that could provide you with relevant data.

Digital Capacity

By "digital capacity" we mean the extent to which a language is infrastructurally and technologically supported and may function in the digital world. Basic conditions such as availability of Internet connection and digital literacy must be met for a community to use a language digitally. Similarly, the availability of functionalities such as spell checkers over cellphones can boost - in principle - its use by making its typing easier and faster. A language digital capacity only refers to its potential to be used digitally, but it is by no means a guarantee 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 having digital capacity, they are often little used in comparison to the national 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.

Evidence of connectivity

Internet connectivity is the ability of the members of a community to connect to the Internet, be it from a fixed or mobile broadband. If there is no (evidence of Internet) connectivity for the speakers of a given language, then the language is assigned the score ‚1' and automatically placed in category no. 1 (Pre-digital). Any positive evidence of digital connectivity would assign a value of ‚2'. This is based on the assumption that connectivity is considered a basic prerequisite for digital use, i.e. if there is no evidence of connectivity for the target language community, the language will certainly have no significant digital use, and would thus be classified at the pre-digital vitality level. On the contrary, evidence of connectivity does not necessarily imply active digital use of the language, which therefore might be classified minimally at the dormant level.

Label	Score	Micro Indicators
no or scarce	1	scarce BB or 3G internet access
good	2	good BB or 3G internet access

Ideally, one would need official statistics on Internet connectivity among the community of speakers of the language under assessment. At the time of writing, this type of information is hardly ever available. To date, one might use data about Internet connection in the region or country where the language is spoken, keeping in mind that these are rough and potentially incorrect approximations.

If such information is not available rely on your (expert) knowledge about the situation for the target language community, or use national/regional statistics to approximate the estimate, or both.

At present, information about internet connection of the country where the language is spoken can be found from Eurostat statistics:

http://ec.europa.eu/eurostat/statistics-explained/index.php/Internet_access_and_use_statistics_-_households_and_individuals

http://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_digital_society_statistics_at_regional_level

These still do not take into account the linguistic context, but may be more accurate than statistics at national level. In fact, Internet connection might be influenced by social-economic factors that affect that part of the country where the regional or minority language is spoken.

Digital literacy

Digital literacy refers to the skills required to achieve digital competence, i.e. the confident and critical use of information and communication technology (ICT) for work, leisure, learning and communication[1].

Digital literacy is another essential pre-condition for digital language vitality. If speakers of the language are digitally not (or very minimally) proficient, then it becomes evident that their native language has little chance of being digitally active, notwithstanding its spoken vitality or technological readiness. In the present study, 'digital competence' is one of the dimensions that shall be considered for assessing the potential for digital presence/vitality of languages. As such it should address the digital skills that members of a given language-speaking community possess, irrespective of the language(s) they use on digital media and devices. The rationale behind this being that, if a person has high digital skills, she/he is (potentially) able to use her/his own mother tongue digitally. Whether a given language is or is not actually used digitally is another matter, reasons for which require a separate study.

Label	Score	Micro Indicators
low	1	individuals have no or low e-skills
basic	2	individuals have basic or above basic e-skills

We make reference to the European Digital Competence Framework for citizens (DigComp)[2], a framework designed and adopted by the European Commission for the (self-)evaluation and improvement of citizens' digital skills/competences, which provides a common reference for digital competence in Europe.

The EC has made the improving of digital literacy and competence of EU citizens one of its key strategic elements in the Digital Agenda (see for instance the Digital Competence Framework[3] initiative) and runs yearly surveys to assess individuals' level of computer skills'.[4] Such surveys would be an ideal source of information for measuring the digital literacy indicator. However, they show a major limitation: they do not take into account the linguistic (minority) dimension, so that statistics are available only at national levels, whereas for assessing the digital vitality of a RML one would ideally need data on the digital literacy of the speakers of the language. Another potentially useful source of information are the ITU statistics[5], which give some indication of possession and

1 http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Digital_literacy
2 https://ec.europa.eu/jrc/en/digcomp
3 https://ec.europa.eu/jrc/en/digcomp/digital-competence-framework
4 http://ec.europa.eu/eurostat/tgm/refreshTableAction.do?tab=table&plugin=1&pcode=tsdsc460&language=en
5 http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

use of computers, still at national level. However, like connectivity, digital literacy could indeed be influenced by social-economic factors that affect only the part of the country where the minority language is spoken.

To the best of our knowledge, official, public domain information is available for countries and possibly for officially recognized regions, but they do not take into account specifically language communities, age groups, or smaller geographic areas. When applying the scale and using the mentioned sources to measure digital literacy, we must therefore be aware that taking the country statistics as an estimate of literacy in the RML communities is likely to overestimate the situation.

Internet penetration or digital population size

Internet penetration is the percentage of Internet users over total speakers' population and is here again considered as a basic pre-condition for language digital vitality. Internet users are defined as persons who accessed the Internet in the last 12 months from any device, including mobile phones.[6]

For the present purpose of assessing the overall "digital capacity" of a language, we are interested in estimating, within a given language speaking community, the number or proportion of people using the Internet in any language, i.e. the potential for the language being used digitally. The rationale behind this indicator is that if, irrespective of the actual language of use, there is a high number of speakers of the language under assessment that use the Internet regularly, then the potential for that language to be used is higher than if the overall use of the Internet was less diffuse of less frequent. For instance, if most speakers of, say, Sardinian used the Internet regularly, albeit mostly in Italian, then the potential for Sardinian to be used on the Internet would be higher than if they used the Internet rarely overall.

It must be borne in mind, also, that a digital population will always comprise passive digital users of the language and, therefore, an ideal indicator would therefore be one showing the population belonging to a given language-speaking community actively engaged in digitally mediated interaction (in any language). Such an indicator however has not (yet) been developed, for it would be almost impossible to measure it at the time of writing.

Label	Score	Micro Indicators
scarce	1	no regular Internet use (less than once a week)
good	2	regular Internet use (more than once a week)

At present, it is not possible to measure this indicator accurately as the available data does not allow us to distinguish language groups among Internet users, and therefore any calculation will rely on extremely approximate data. Possible data sources are Internet World Stats[7], ITU[8], Eurostat[9]. However, these global statistics or breakdowns per countries are not suited to capture this piece of information for non-national languages. This data could be integrated with an ad hoc survey or more specific data about the digital use in any language among the speakers of a given (minority) language.

Character encoding and input/output methods

Unavailability of character/script encoding (e.g. Unicode) severely limits the digital usability of a language, although it does not completely disallow it. In theory, a language can be digitally present in its oral form, e.g. through video or audio. However, since the Internet is still predominantly a written medium, the possibility for a written digital use represents a clear indicator of digital usabil-

6 https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users

7 http://www.internetworldstats.com

8 https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/Individuals_Internet_2000-2015.xls

9 http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tin00028&plugin=1

ity. Having its script included in the universal character set is crucial for a language to participate in the technological advancements of computing, as well as to promote native language education, literacy, and cultural preservation. Moreover, character encoding based on internationally accepted standards enables worldwide interchange of text in electronic form.

There are many different types of character encodings at present, but the ones we deal most frequently with are ASCII, 8-bit encodings, and Unicode-based encodings. Most language scripts and alphabets are represented by Unicode standard, but nevertheless there is still evidence of spoken languages for which this is not available .

This indicator also takes into account the availability of input methods such as a specific keyboard, when needed. It does not consider whether the language has a standardised writing system, which of course plays a role on overall digital usability and use, as this is considered to have an indirect influence on the indicators pertaining to the digital usability and performance measured elsewhere in the scale. Here we are interested in the technological pre-conditions for using a language on digital devices, irrespective of standardisation.

Label	Score	Micro Indicators
unsupported	1	language with no standardised script encoding and no alternative script is used
informal	2	language with no standardised script encoding; alternative supported script is used
developing	3	language for which a script proposal is available
proposed	4	language with a consistent and agreed upon encoding that may not have entered already the standardisation process
standardised	5	language with a standardised character/script encoding; fonts, keyboards and software may not be fully available
supported	6	language with a standardised character/script encoding; fonts, keyboards and software is updated and available

For a list of Unicode supported scripts see the Unicode site[10]; other sources are Unicode consortium[11]; Script Encoding Initiative[12], and the Wikipedia page on Unicode font[13].

Availability of language resources

Language resources are, for example, corpora, grammars, electronic monolingual or bilingual dictionaries.

The availability of language resources is an essential condition for allowing the development of natural language processing applications, from basic ones, such as spell-checkers, to advanced ones, such as machine translation applications.

Texting and messaging, for instance, is crucially enabled not only by the availability of keyboard characters, but also by the integration of auto-correction devices.

By using the term availability, we refer to the fact that a particular language resource or tool can be used by a technology that enables the speakers to carry out a task or receive a service. The mere existence of a resource does not imply the idea of availability, if it is not exploited to provide

10 http://www.unicode.org/standard/supported.html
11 http://www.unicode.org/versions/Unicode9.0.0/
12 http://linguistics.berkeley.edu/sei/index.html
13 https://en.wikipedia.org/wiki/Unicode_font

services or is in an experimental or prototype phase. For instance, the fact that a prototype spell-checker exists for a language, but is not exploited by any application, makes it not-available to the community.

For the purpose of assessing the digital fitness of a language, we differentiate the following types of language resources (LRs):

» Basic: monolingual and bilingual e-dictionaries; digital corpus (\textless 100 million words); POS-tagging, spell-checker

» Intermediate: corpus driven monolingual dictionary; digital corpus (\textgreater 100 million words); parallel corpora; web-corpora; term extraction; shallow syntactic parsing; basic MT (rule-based); speech synthesis (TTS)

» Advanced: big corpora (Gw), multilingual corpora; deep syntactic parsing; WordNet, semantic processing; advanced MT (SMT, hybridation, neural); speech recognition.

Using those levels of LRs as a reference, we propose the following scale:

Label	Score	Micro Indicators
none	1	no language resources available in digital format
minimal	2	e-dictionary (bilingual or monolingual)
limited	3	at least 2 basic LRs
medium	4	basic LRs and, at least, 3 intermediate LRs
strong	5	most of the intermediate LRs
high	6	most of the advanced LRs

Measuring the availability of digital language resources

We are aware that knowledge about language resources, in the sense used here, very much belongs to a specialist domain and is the kind of information not readily available to non-experts. Moreover, instruments for assessing the availability of language resources for the various languages are still under development.

Possible sources of information are the CLARIN Virtual Language Observatory[14], Linguistic Data Consortium[15], the Universal Catalogue maintained by the European Language Resource Association[16] and the META-SHARE catalogue[17]. By filtering the search with the name or ISO code of a specific language, these sites will return the language resources that are registered on these sites.

Other useful reference works, although developed for the eleven EU official languages only, is the META-NET Language White Papers series[18], where the technological support for natural language processing applications is assessed.

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 varied range of different purposes.

14 https://vlo.clarin.eu
15 https://www.ldc.upenn.edu/
16 http://catalog.elra.info/
17 http://www.meta-share.eu/
18 http://www.meta-net.eu/whitepapers/overview

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 edutainment purposes, among the many uses possible. Again, these 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 ebooks or video-games in the language. These digital uses of the language also encompass a progression from more private uses of the language to those more public ones, including official usage. What they have in common is the fact of referring to the creation of digital content in the language, whether it is used for communicating or for other purposes.

There are four indicators for this class: Use for e-communication, Use on social media, Availability of Internet media, and Wikipedia.

Use for e-communication

This indicator assesses whether the language is being used digitally for interpersonal communication among people. e-communication, especially under the form of instant messaging, shares many features with orality, and as such is one of the privileged contexts where spontaneous forms of expressions emerge. The one-to-one directionality of messages adds a privacy dimension, which fosters incipient uses of a language, often by users not fully confident with its written standard. Typical e-communication media are e-mail, texting and instant messaging, but also audio-only channels such as Skype.

From the perspective of using e-communication as an indicator of digital language vitality, what is important to capture is not only the variety and quantity of communication media available, but also the frequency with which they are used. Therefore, the different values for this indicator are expressed under the form of "rules" for assigning a particular value on the basis of the occurrence of particular frequency conditions.

Label	Score	Micro Indicators
none	2	no use for e-communication
minimal	3	at least one communication medium that is used at least rarely
medium	4	at least two communication mediums that are used at least occasionally
strong	5	more than two communication mediums that are used at least regularly
high	6	more than two communication media that are used everyday

How to assess use for e-communication

Because this indicator mainly refers to private communication exchanges, at present there is no objective or reliable source of information that can be used to estimate its value, and it's quite unlikely that there will be one in the future, for privacy matters.

Therefore, the main source of information for this indicator is a survey of the type carried out in the DLDP project[19]. If information is to be collected through a survey, then we suggest to apply a threshold of at least 60% of respondents to ensure reliability of the information provided.

Use on social media

Use on social media, such as Twitter, Facebook, Instagram, etc., is a strong indicator of active digital use of a language. Social media does not merely function as means for computer-mediated

19 http://www.dldp.eu/content/reports-digital-language-diversity-europe

communication. On top of that, they act as virtual spaces of expression: they transform language speakers from consumers to creators of content, and especially in local contexts, they represent important new domains with the potential to impact upon the use of a minority language.

Similarly to what suggested for the previous indicator, we consider the variety of social media where the language is used and the frequency of use as the main criteria for estimating the use of a language on social media.

Label	Score	Micro Indicators
none	2	no use on social media
minimal	3	one or two social media that are used at least rarely
medium	4	at least three social media that are used at least regularly
strong	5	at least three social media, one of which is used at least everyday
high	6	more than three social media used everyday or at least regularly

How to assess use on social media

At the time of writing we are not aware of tools or applications being able to crawl social media on the basis of a specific language requirement. A notable exception is, for Twitter, the Indigenous Tweets service[20] that allows a reader to see the amount of Twitter accounts and tweets in a number of languages. Therefore, either because not all networks allow for the trawling of their data (as is the case for Facebook), or because language identification is a technology still under development for the vast majority of languages, it is still very difficult to capture the extent to which a language is used on social media.

Until this ideal condition is met, we suggest to resort once again to the information that can be gathered through an informative questionnaire of the type developed by the DLDP project. If based on a questionnaire, we suggest to apply a threshold of at least 60\% to ensure reliability.

Availability of Internet media

With this indicator we aim to capture the variety of Internet media that are available in a given language, with a sufficient amount of content.

The term "Internet media" broadly refers to a range of communicative instruments such as web pages and websites, blogs, forums, but also Internet radio and TV, allowing production and consumption of content in digital form. Internet media are in most cases traditional media delivered for consumption on a new platform (e.g. television or radio programmes delivered online, e-books); others are radically new forms of content production and consumption, and include a dimension of interactivity that was not possible up until relatively recently, where content was published in print and reactions to it would take the form of edited letters.

Because of the relative ease and affordability of their creation and use, Internet media provides a viable alternative for content creation in minority languages that are often deprived of the institutional support that would be essential for traditional media. And because of the centrality of digital media in modern life, their availability in a language is as well a sign of the (digital) vitality of that language, and its use conquering new and important domains.

20 http://indigenoustweets.com

Label	Score	Micro Indicators
none	2	no Internet media available in the language (less than one)
limited	3	some Internet media available in the language (= 1 to 2)
medium	4	some Internet media available in the language (= 3 to 4)
strong	5	a considerable variety of Internet media is available (= 5 to 6)
advanced	6	a wide variety of Internet media is available (more than 6)

Here is a list of relevant types of Internet media (this is the same list that was used in the DLDP survey[21]:

- websites (includes publicly available websites, and private such as intranets and extranets)
- smartphone apps
- Internet television (both television channels delivered by the Internet, such as Netflix service, or individual TV shows offered via the websites of channels)
- blogs and forums/message boards
- streaming audio (including podcasts and Internet radio)
- streaming video (including webcasts, podcasts, YouTube and Vimeo videos/ channels)
- eBooks
- digital libraries and archives

How to assess availability of Internet media

The ideal source of information would be a dedicated survey or official surveys/censuses, if available. If a survey is exploited we suggest to apply a threshold in order to enhance reliability, for instance greater or equal to 30% of respondents.

Wikipedia

Other authors in the past have considered Wikipedia as an indicator of digital vitality (Kornai, 2013[22]). This particular role of Wikipedia as a sign not only of digital use, but also of active interest in the creation of digital content stands behind the decision of considering Wikipedia as a separate indicator of digital vitality.

The sheer existence of a Wikipedia should not be overestimated as a healthy sign of digital vitality, though: not only are there wikipedias for classical languages (this is the case of Vicipaedia, the Wikipedia in Latin language[23]), but it can also happen that a wikipedia project is set up by a small number of activists with no correspondence with frequent and active digital uses of the language outside the Wikipedia community.

The most straightforward sign of the importance of a Wikipedia as an indicator of language vitality is its number of articles. This data is provided by the Wikipedia Foundation at the page "List of Wikipedias"[24]. It is easy to obtain, and fully reliable.

The table below illustrates a grading according to number of Wikipedia articles. The intervals in the number of articles are those used by Wikipedia to classify the different language editions.[25]

21 http://www.dldp.eu/content/reports-digital-language-diversity-europe
22 http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077056
23 https://la.wikipedia.org/wiki/Vicipaedia:Pagina_prima
24 https://meta.wikimedia.org/wiki/List_of_Wikipedias
25 https://www.wikipedia.org/

Label	Score	Micro Indicators
none	1	no Wikipedia
incubator	2	a small Wikipedia is available (less than 100 articles)
small	3	between 100 and 10,000 articles
medium	4	between 10,000 and 100,000 articles
high	5	between 100,000 and 1,000,000 articles
big	6	over 1,000,000 articles

Digital Performance

Digital Performance groups together indicators referring to what can be digitally done with a language, i.e. the available digital services. For instance, is it possible to read online newspapers? Is the software mostly used localised in the language? Is machine translation available? Most regional and minority languages, especially if located in Europe, do not have such advanced facilities. Part of the reason for this lies in the fact that the availability of these services is linked to a) whether a language is officially recognised and b) its potential for market. If a language is not officially recognised, then it often fails to display services that need an institutional support. Similarly, if little commercial interest is associated with it, then little investment is provided by large companies. This latter phenomenon is often compensated by local, small commercial initiatives or by the open software community and digital activists.

Availability of Internet services

This indicator refers to the existence of digital services localised in the language (e.g. Amazon, Booking, Google, etc.)

Here is the list of types of Internet services that was used in the DLDP survey and which are up-to-date, relevant services at the time of writing:

- Online newspapers
- Online news
- Search engines (e.g. Google)
- Edu-tainment products and services, (e.g. videogames, kid-friendly websites, digital apps)
- Entertainment (music channel, movies..)
- Video subtitling
- Health services
- e-Commerce services (e.g. airline reservation sites, local e-shopping services)
- Public Administration/eGovernment (e.g. local county)
- eBanking (online banking services offered by bank)
- Cultural services (e.g. virtual museums and heritage websites)
- Tourist information and services (e.g. tourist agencies)
- Advertising, promotion and marketing
- Customer care

The table below gives the scoring indications.

Label	Score	Micro Indicators
none	2	no digital services available in the language (= less than 1)
limited	3	some digital services available in the language (= 1 to 2)
medium	4	Some digital services available in the language (= 3 to 5)
strong	5	a considerable variety of digital services is available (= 6 to 8)
advanced	6	a wide variety of digital services is available (= more than 8)

How to assess availability of Internet services

Ideally, objective data (e.g. registries or catalogues) would be needed of the number of (different) services existing for the language under assessment. However, such data is currently often not available. Alternatively, one can rely on own independent research of the available services, also relying on one's own knowledge of the local language community and situation, and/or data from a survey like the one carried out for the DLDP project. Census, objective or expert information is always to be preferred to information elicited from users. These often tend to overestimate the availability of services, or can misinterpret the question. Therefore, if you are going to use a survey of the kind we used in the DLDP project, then you should consider a reply as valid if shared by at least 40% of respondents.

Localised social networks

This indicator refers to the presence of social media with a localised interface. The most popular social networks offer their interface localised in a number of different languages: at the time of writing, Facebook offers 113 languages, Twitter 48. It is still debated whether the availability of interfaces translated or localised has any positive influence on the use of a language instead of another. In a study of the use of Puerto Rican Spanish on MySpace[26], Carroll (2008) argued that an English-language interface lead to the adoption of English terminology, and this would encourage to think that the availability of a translated interface would promote the use of the same language. Unfortunately, there are no studies available inquiring this correlation. Cunliffe's research (Cunliffe 2013)[27] revealed little evidence that the Facebook Welsh interface was a positive influence to use Welsh.

In any case, we share Cunliffe's view that that the sheer existence of a localised interface has a positive influence on the perception of the language as modern and suited to being use in ICT contexts.

Label	Score	Micro Indicators
none	3	No social media localised in the language (= less than 1)
limited	4	At least one social media interface localised in the language (= 1)
medium	5	Some social media interface localised in the language (= 2 to 3)
advanced	6	Many social media interface localised in the language (= more than 3)

26 https://www.researchgate.net/publication/279709617_Puerto_Rican_language_use_on_MySpacecom

27 https://onlinelibrary.wiley.com/doi/full/10.1111/jcc4.12010

How to measure localised social networks

Checking whether a social network is available in a language is fairly simple. For Facebook, you'll have to select "Language" from the menu "Settings", and then search for a particular language on the list that appears when clicking on the question "Which language do you want to use Facebook in?". On Twitter, the available languages appear immediately from the "Account" menu. For social networks that are less widely available, a survey might serve as a reliable source of information.

Localised software: operating systems and basic software

Computer software is usually classified into three categories:[28] 1) system software or operating systems; 2) application software, that can be general purpose (word-processing, web browsers, etc.) or special purpose software (accounting, database systems, audio and video production, etc.); and, 3) computer programming tools. The last category has not been considered as an indicator of vitality, since hardly any of those types of tools are localised.

Localisation is "the process of adapting internationalized software for a specific region or language by adding locale-specific components and translating text." [29]

Localisation of software is carried out in different ways. Sometimes, localisation can be done by the system developers themselves, but in the case of regional and minority languages, it is usually promoted or even undertaken by users' communities, local stakeholders, or institutions who want to use the aforementioned software in their own language. At this point, it is important to mention that in some cases localised software does exist in the RML language, but for many reasons, the user's community does not know about it, and therefore, does not use it.

Label	Score	Micro Indicators
none	2	Neither operating system nor general purpose software localised in the language
limited	3	At least one operating system (either desktop or mobile, either open or commercial) localised in the language
medium	4	At least one desktop and one mobile operating system (either open or commercial) + some general purpose software (a word processor and a browser) localised in the language
strong	5	Most used operating systems and general purpose software localised in the language; some specific purpose application software localised
advanced	6	Main operating systems and application software localised in the language

How to check for localised software

In order to give a value to this indicator, one can rely on personal experience with the software and operating systems used. If an assessment of a language different than the one spoken is made (which is not recommended), then the following sources of information can be useful.

- Windows: https://msdn.microsoft.com/en-us/windows/hardware/commercialize/manufacture/desktop/available-language-packs-for-windows
- MacOS: https://support.apple.com/kb/PH18433?locale=it_IT
- iOS: https://support.oneskyapp.com/hc/en-us/articles/206217438-Languages-supported-by-iOS-
- Mozilla: https//addons.mozilla.orj/en/en-US/firefox/language-tools

28 https://en.wikipedia.org/wiki/Software_categories

29 https://en.wikipedia.org/wiki/Internationalization_and_localization

- Chrome: https://www.lifewire.com/change-default-languages-in-google-chrome-4103615
- Android: https://gist.github.com/lopspower/ce384ef98f3d79e61ec78dc9e512644f
- Internet Explorer: https://www.lifewire.com/change-default-languages-in-ie11-4103671
- Thunderbird: https://www.thunderbird.net/en-US/thunderbird/all/
- Firefox: https://www.mozilla.org/it/firefox/all/

The notion of what counts as "main" software and operating systems is clearly variable and cultural-dependent. In the context of the DLDP survey we used a list of software and operating systems and asked people to reply whether they knew if a localised version was available. We report the list here to help the interested reader: Windows, Mac OS X, Linux, Android, iOS, Microsoft Office, LibreOffice, Firefox, Chrome, Internet Explorer, Thunderbird, Adobe Creative Suite, Gimp. This list may need to be modified as needed. If using a questionnaire to elicit this information, we recommend to check afterwards, because we have observed a tendency from respondents to overestimate the availability of localised interfaces.

Machine translation tools/services

This is an indicator of the "proficiency" level of language use on digital media and the availability of machine translation services and tools is considered from the user's perspective, rather than from the point of view of the availability of the technology itself. So this indicator about MT counts differently as in the "Capacity" section.

Since machine translation technology presupposes a wide array of tools and resources, last but not least the availability of big multilingual corpora, this indicator can be taken as a benchmark of strong and active digital performance. Moreover, from the point of view of the digital usability of the language, the availability of reliable MT for a language is a sign that the language has gained a fairly high level of digital presence and importance. The availability of MT "de-minorizes" minority languages (Forcada 2006[30]), and contributes heavily to improving their status, by increasing normality, literacy, and visibility.[31] In the context of the Digital Language Vitality Scale, the availability of MT is an indicator of increasing vitality, use, and prestige.

Here we are ignoring the technical complexity of the realization of MT tools, and focus instead on whether there are (useful) tools for automatic translations (either online or as standalone tools). Therefore, we consider as good indicators the number of different services/tools, the number of language pairs and possibly the directions (e.g. whether from language 1 to language 2 or the reverse as well).

Label	Score	Micro Indicators
none	3	No MT for the language
basic	4	at least one (online?) service/ tool, at least one language pair or one direction
medium	5	at least one (online?) service / tool, at least two language pairs in both directions
advanced	6	more than one (online?) service /tool, more than 5 language pairs

30 https://pdfs.semanticscholar.org/8e9d/b064a62261ba10152c11cd6d05bb306e7cc3.pdf

31 Forcada (2006) also argues that the use of MT systems may contribute to the standardisation of a language by promoting a writing or spelling system

How to check for availability of MT services

It is fairly easy to check if Google provides MT services for a language: the Google Translate page offers a list of available languages.[32]

Another important source of reference for open-source machine translation is the Apertium page[33] (\url{}), as well as the list of tools available here.[34]

Probably, it is quite safe to rely, even if not exclusively, on the knowledge and/or perception of the respondents to the questionnaire, since whether a service or tool is known and used by the digital users of the language should have some weight. It would be a way to account for the spread of the service within the community, i.e. how much it is used.

Dedicated Internet top-level domain

This indicator is related to the existence of a top-level domain name (a geographic top-level domain or GeoTLD) dedicated to a certain linguistic and cultural community (e.g. .cat; .eus; .bhz). A GeoTLD is a generic top-level domain using the name of or invoking an association with a geographical, geopolitical, ethnic, linguistic or cultural community.[35]

The success of the .cat top-level domain name has inspired the creation of several independent Internet identities for linguistic and cultural communities.

The fact that a language has its own top-level domain name cannot be unequivocally interpreted as a sign of high vitality. The precise vitality level depends also on the penetration of the domain in the geographical area where the RML is used, and on the density proportion of websites within the domain that are written in the RML[36]. Moreover, as pointed out by Cunliffe (2007)[37], it is unlikely that small minority language communities will meet ICANN's (the Internet Corporation for Assigned Names and Numbers) criteria the other hand, neither is the existence of a dedicated domain name a necessary condition for vitality, because, in principle, a language can be digitally thriving without a dedicated domain name.

Nevertheless, languages having their own top-level domain name present a non-negligible digital activity, indicating that there is some level of correlation between the existence of the domain and digital vitality. Moreover, a dedicated domain name can be taken also as an indicator of a strong digital identity.

The values proposed for this indicator are yes / no and if no, the minimum corresponding level could be `emergent' (thus scoring 3) while in case of presence of an Internet domain the minimum level would be 'vital' (thus scoring 5).

Label	Score	Micro Indicators
no	3	No dedicated Internet top-level domain
yes	5	There is a dedicated Internet top-level domain

32 https://translate.google.com/intl/en/about/languages/

33 https://www.apertium.org

34 http://fosmt.org

35 https://en.wikipedia.org/wiki/Generic_top-level_domain\#Geographic_gTLD

36 https://www.domeinuak.eus/wp-content/uploads/2015/12/PuntuEUS-Observatory-2015.pdf

37 Minority languages and the internet: new threats, new opportunities. Cunliffe, D. In: Minority Language Media: Concepts, Critiques and Case Studies, M. Cormack and N. Hourigan (Eds.), Multilingual Matters, Clevedon, 2007: 133-150

How to check for the existence of a dedicated Internet top-level domain

Information about current top-level domains (TLDs) is provided by the ICANN website[38]. A database of TLDs is maintained at IANA[39]; however this Wikipedia list[40] can be more useful for identifying the linking of a domain to a language and culture.

How to calculate the Digital Vitality Level for a language

Put all the scores into a table and then calculate the average value. The resulting number will indicate to you the corresponding digital vitality level.

38 https://www.icann.org/resources/pages/registries/registries-en

39 http://www.iana.org/domains/root/db

40 https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains#Geographic_top-level_domains