BLaRKing at minority language speakers

The Digital Language Survival Kit as a speaker-centered approach to digital development of minority languages

Claudia Soria

Institute for Computational Linguistics - National Research Council, Italy
claudia.soria@ilc.cnr.it

Table of contents

1. Introduction
2. The Digital Language Diversity Project
3. The DLDP approach and methodology
    - Assessing vitality and evaluating needs
    - Learning: Training Programme and Digital Language Survival Kit
    - Planning: the Digital Language Survival Kit
4. A closer look to the Digital Language Survival Kit
    - Recommendations concerning Language Resources
5. Conclusions

Introduction

Introduction

I think there is a common ground we share here: the value we place on linguistic diversity and on the importance that all languages, no matter how small, be represented digitally. So it’s not a matter of having the conversation about why we should increase digital language diversity, but rather about how we can achieve it.

The Digital Language Diversity Project

The Digital Language Diversity Project

- A three-year project (2015-2018) funded under Erasmus+ Programme, Adult education
- Aims to address the problem of low digital representation of EU regional and minoritised languages
- By giving their speakers the skills to create, share, and re-use online digital content.

The Digital Language Diversity Project (http://wp.dldp.eu)

"The mission of DLDP was to advance the sustainability of Europe's regional and minority languages in the digital world by empowering their speakers with the awareness, knowledge and abilities about the actions that can be concretely put in place to make their languages survive and possibly advance in the digital context."

Short-term objectives and results

- fostering the notion of digital language diversity and vitality and creating awareness about the risk faced by minority languages of not being technologically adequately supported;
- defining strong, clear and actionable recommendations about what needs and can be done for a language “to go digital”: which are the challenges and difficulties, which areas need to be addressed first, which tools are available;
- providing a widely applicable training programme, targeted to ML speakers to guide them towards effective production of digital content in their languages;
- laying out an indication for the immediate future, especially in relation with other projects and initiatives, with a view to national governments and EU institutions.

Three main assumptions

- digital language diversity is limited
- digital presence is important for revitalisation and preservation
- digital presence can be increased bottom-up through *community empowerment*

Digital language diversity is limited

The world's language diversity is not mirrored in the digital world:

- 87,5% of websites are in one among 10 languages (1)
- 25 languages account for 98% of web site content (2)
- speakers of 94% of the languages spoken on the planet cannot access Internet services unless they are fluent in one dominant language as well (3)
- only 20% of the world (primarily white male editors from North America and Europe) edits 80% of Wikipedia currently (4)
- 84% of Wikipedia articles focus on Europe and North America (5)

Sources: (1, 2): W3Techs statistics (4, 5): https://whoseknowledge.org

Digital presence is important for equality, revitalisation and preser vation

Human language will be the predominant means of communication between human and machines and for accessing collective knowledge and information.

A language that is not digital is considered as not providing any competitive advantage.

Monolingual speakers of minority languages are disadvantaged and discriminated. Multilingual speakers will tend to abandon the digitally minoritised language not to miss the digital train.

Digital presence can be increased bottom-up

Minority languages are usually of little economic interest or enjoy limited institutional support.

There is a wide range opportunities for language speakers to give an impulse to the digital presence and usability of their languages.

Speakers can and must be educated to take action for their languages.

Precursors and inspiration

- Andras Kornai's *Digital Language Death* (Kornai 2013)
- META-NET *Language White Papers* (Rehm & Uszkoreit 2012)
- Basic Language Resource Kit (BLaRK, Krauwer 1998)

Andras Kornai's *Digital Language Death*

Brings the traditional methods of language vitality assessment to the digital realm

Explicitly connects to Fishman's Graded Intergenerational Disruption Scale (GIDS, Fishman 1991; later revised as the Expanded GIDS (EGIDS, Lewis & Simons 2010)

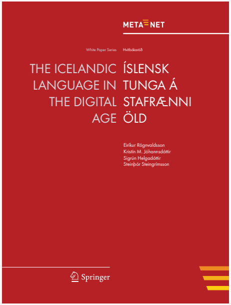
Proposes a four-level classification of digital vitality (Still, Heritage, Vital, Thriving) and a set of proxies.

In our subjective estimate, no more than a third of the incubator languages will make the transition to the digital age.

META-NET Language White Papers

In 2012, the META-NET ’Language White Papers’ introduced the concept of *digital extinction* to refer to the risk faced by the majority of EU official languages and consisting in a dramatic contraction of uses as a result of the lack of technological support.

The concept had enormous success in the media and significantly helped drawing attention to the issue.



META NET

White Paper Series Hvítbókaröð

THE ICELANDIC LANGUAGE IN THE DIGITAL AGE

ÍSLENSK TUNGA Á STAFRÆNNI ÖLD

Eiríkur Rögnvaldsson
Kristín M. Jóhannsdóttir
Sigrún Helgadóttir
Steinþór Steingrímsson

Springer

BLaRK

- Basic Language Resource Kit
- 1998, Steven Krauwer
- The minimal set needed to do any precompetitive research and development
- Similar focus on assessment and priority instrument
- Dynamic nature
- BLaRKette, MyBLaRK
- Successful in minority language contexts.

The DLDP approach and methodology

A complete methodology for digital language planning

The main activities of the project revolved around the development of a complete methodology for ML speakers' communities for digital language planning:

1. **evaluating** the digital needs of a given speakers' community
2. **assessing** the degree of digital vitality of its language
3. **learning** the range of possible actions and activities that can be put in place according to the identified level of vitality.

Phases of the methodology and related instruments

1. Assessing vitality and evaluating needs:
   the Digital Language Vitality Scale & Survey
2. Learning:
   the DLDP Training Programme
3. Planning:
   the Digital Language Survival Kit

Assessing: The Digital Language Vitality Scale & Survey

- the Digital Language Vitality Scale (DLSV), a scale and associated indicators for assessing the degree of digital fitness of a language.
- a model for a survey for eliciting the information needed to apply the Digital Language Vitality Scale.

What is the scale?

An instrument for estimating the degree of digital vitality of any given language.

Aimed 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, we provided guidelines on how to measure or estimate the indicators included in 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 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.

Six levels

1. Pre-digital
2. Dormant
3. Emergent, e.g. Sardinian, Karelian
4. Developing, e.g. Basque, Breton
5. Vital
6. Thriving, e.g. English

How to apply the scale: dimensions

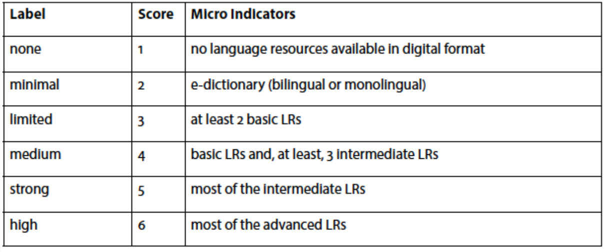
The DLSV uses *dimensions* and *indicators* in order to assess the degree of digital vitality of a language.

- digital *capacity*: the extent to which a language is infrastructurally and technologically supported and may function in the digital world
- digital *presence* and *use*: the amount and type of digital content that is available in a given language, be it for communicative, informational, or recreational purposes, among the many
- digital *performance*: what can be digitally done with a language, i.e. the available digital services

How to apply the scale: Indicators

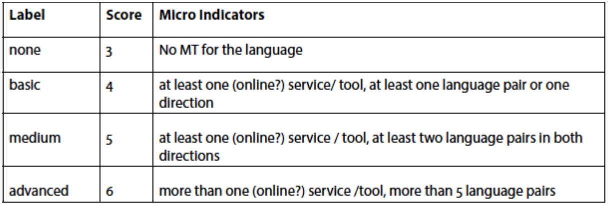
- digital capacity:
    - Evidence of connectivity; Digital literacy; Internet penetration or digital population size; Character encoding and input/output methods; Availability of language resources
- digital presence and use:
    - Use for e-communication; Use on social media; Availability of Internet media; Wikipedia;
- digital performance:
    - Availability of Internet services; Localised social networks; Localised software; Machine translation tools/services; Dedicated Internet top-level domain

How to apply the scale: Examples



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

How to apply the scale: Examples



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

What the scale can be used for

The Digital Language Vitality Scale is the first and 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.

A baseline for making informed decisions regarding the digital development of a language

The particular types of actions and measures needed will be chosen by the language community

Experts can and should provide guidance and expertise about the range of possible actions to be taken.

Assessing: The Survey

The Digital Language Diversity Project

Basque — a digital language?

Erasmus+

The Digital Language Diversity Project

Euskara — hizkuntza digitala?

Erasmus+

Goal of the Survey

Main goal:

to answer the question *”is it possible for regional or minority language speakers to have a digital life in those languages?”*, i.e.

to inquiry about the digital behaviour, desires, and expectations of speakers of regional and minority languages

and secondarily, to gather evidence and information to feed the Digital Language Vitality Scale for Basque, Breton, Karelian and Sardinian.

Structure of the Survey

Designed around three main conceptual blocks:

1. the digital capacity of the language, i.e. if the technological conditions for its digital use are in place, such as the availability of internet connection, or the possibility to type the language
2. the opportunity to make a digital use of the language, under the form of available contexts and purposes for its digital use such as digital media and services
3. the speakers' attitudes towards digital use of the language: if it is felt as desirable, what are the underlying motivations for it, what are the blocking factors, if any.
    - Particular attention was devoted to highlighting the possible problems encountered in using the language digitally.

Questions

- demographic questions (age, sex, place of birth, …)
- competence assessment and attitudes re. language
- digital activism
- digital use of the language:
- use for e-communication: type, frequency, reasons for no use
- language use over the internet: type, frequency, reasons for no use
- character encoding and keyboard availability
- (knowledge) of the existence of (digital) media in the language
- existence and use of Wikipedia
- presence and use of language on social media
- localisation of social media interfaces
- existence of online services in the language
- existence of localization for main operating systems and software
- existence of digital language resources
- free comments

Methodology

Developed by the DLDP consortium with members of Advisory Board

Closed questions where the informants had to tick either only one box or more than one

Questionnaire template in English to ensure max. comparability and reusability

Translated and localised into Breton, Basque, Sardinian, and Karelian

Made available through Google forms between July and September 2016

Participants mostly recruited among partners of the European Language Equality Network (ELEN)

Advertised on social media and through personal contacts

A total of 1.301 replies was received

Some survey results

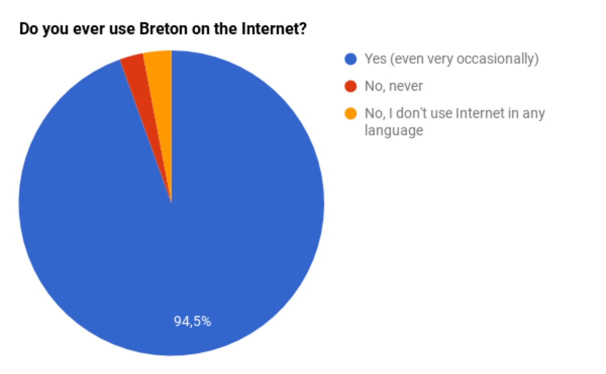


Figure 1: Use of Breton on the Internet

Some survey results

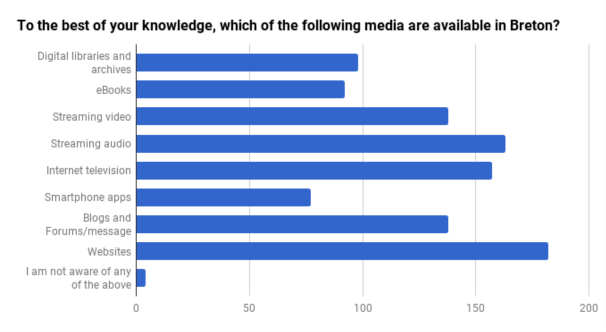


Figure 2: Available digital media in Breton

Some survey results

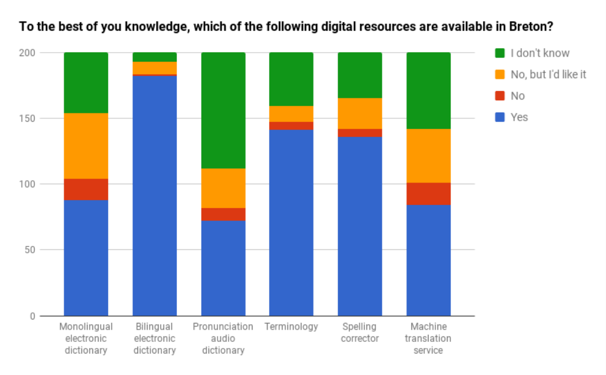


Figure 3: Available digital resources in Breton

Basque (Developing)

- A digitally fit and actively used language.
- Respondents have high linguistic competence and good knowledge of the existing digital tools and resources.
- Widely used on social media
- Despite knowing about the existence of localised digital services, some respondents are not using Basque in their devices, applications or software.
- There is a demand for more entertainment products in Basque and especially addressing young people.
- Most people are consuming computer or mobile games in other languages because finding them in Basque is hard.

Breton (Developing)

- Digitally fairly well developed
- Many respondents to the survey call for more apps, more software, and for Windows in Breton
- there is some provision in social media in Breton, for example, a Breton interface version of Facebook
- demand for machine translation and inclusion on apps such as Google Translate
- people want to be able to live their (digital) lives in Breton

Karelian (Emergent)

- Karelian digital presence is developing, but still needs much work done
- Speakers have a high linguistic self-esteem and this translates to a will to use the language online.
- Many necessary online and digital resources are missing, and speakers often have no or false information about those that do exist.
- In terms of social networks, use of Karelian is very heavily restricted to Facebook.
- Wide unawareness of availability of resources, services, and opportunities to use the language (e.g. keyboard settings)

Sardinian (Emergent)

- Extensively used online by the large majority of its speakers
- Particularly vital on social media, among which Facebook is by far the most used network.
- For Facebook there is even a localised interface available.
- Vitality on social media does not correspond to strong and diffused availability of Internet media.
- Existence and availability of digital services in Sardinian is vastly unknown.
- Online newspapers and news are widely available, as is entertainment and, thanks to a previous investment by Regione Sardegna, some Public Administration services.
- More advanced media such as smartphone apps, Internet TV, audio and video streaming are instead lacking. People express a strong desire to be able to use Sardinian on the Internet as part of their everyday life.

Lessons learned

Complexity of analysing data about minority languages: how can data be gathered? Is a survey the best way to gather such data? How reliable can this data be?

Absence of official data can be first motivation behind choice of a survey

Issues of sample representativeness and reliability of replies should not be underestimated: we found that some digital provision that was reported as available is actually not.

Certain degree of distortion, due either to a misunderstanding of the questions or carelessness in providing replies should be taken into account.

Do people tend to overestimate the digital development of their languages in the attempt of making them look better suited, to promote them in the eyes of external evaluators?

Training Programme

Goal: providing speakers of minority languages with some skills to create and share digital content.

- A collection of material and instructions and examples of what has been done in other languages in order to inspire the TP participants
- Modular structure – seven modules on different topics All modules: intro
- Implemented as a Moodle course

Training Programme Modules

1. Why Do We Need Digital Language Diversity?
    - Introduction and Guide to the Module
    - What Is Digital Language Diversity and Why Is It Important?
    - Success Story: Rising Voices - Promoting Indigenous Languages Online (Eduardo Avila's guest talk)
    - Collection of articles (Webography)
2. Assessing Digital Vitality
    - Introduction and Guide to the Module
    - Assessing the Digital Vitality of a Language
    - How to Use the Digital Language Vitality Scale
    - How to Use the DLDP Questionnaire and Conduct a Survey
    - The DLDP Questionnaire as a Template
    - How to adapt the Digital Language Diversity master questionnaire on Google Forms
    - Overview of Online Survey Platforms
    - The DLDP Survey Reports
    - If You Want to Know More

3. Social Media
    - Introductory Talk: Minority languages and Social Media (Teresa Lynn)
    - Success Story: Localisation of Facebook in Sardinian
    - Tutorial: How to Set Social Media Platforms in Your Language
    - If you want to know more...
4. Wikipedia
    - Introduction and Guide to the Module
    - Introductory Talk: Wikipedia for Regional, Minority Languages (Shubhashish Panigrahi)
    - Success Story: Wikipedia in Basque
    - Success Story: Wikipedia in Karelian
    - If you want to know more ...
5. Multimedia Content Creation
    - Introduction and Guide to the Module
    - Multimedia Content Creation
    - Success story: The BASAbali Wiki (Alissa Stern)
    - How to Set up Your Own Website
    - How to Make Your Own Videos
    - How to Add Subtitles to an Existing Video

6. Language Technologies & Digital Activism
    - Introduction and Guide to the Module
    - Guest talk (Delyth Prys)
    - Online dictionary making
    - Working with corpora

The DLSK

the Digital Language Survival Kit is a set of recommendations addressed at individual speakers and speakers' communities regarding the actions that can be taken – mostly at the grassroots level – to make a language progress towards the next steps of digital vitality.

The Digital Language
Diversity Project

Digital Language Survival Kit

The DLDP Recommendations
to Improve Digital Vitality

Erasmus+

Structure of the DLSK

The Kit is linked to the scale:

- it addresses three levels (Dormant, Emergent, and Developing)
- it is organised into sections corresponding to the scale’s dimensions and according to the various indicators of the scale

A closer look to the Digital Language Survival Kit

Structure I

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

- Digital capacity
- Digital presence and use
- Digital performance

Structure II

The recommendations are intended for three levels of digital vitality: *Dormant*, *Emergent* and *Developing*.

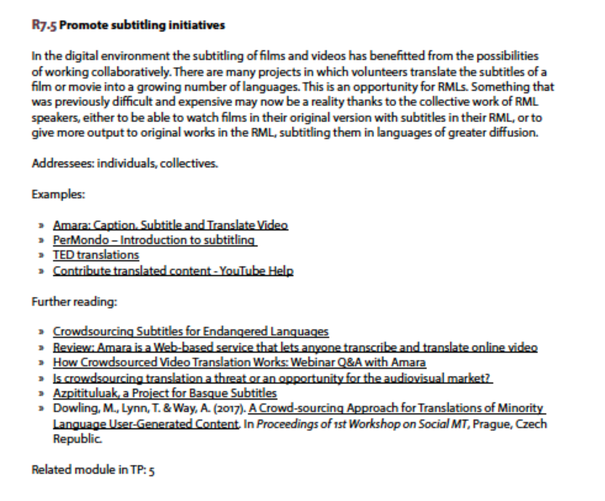
Some recommendations are specific for a level, and some others are suitable for different levels.

Structure III

In addition to being organised into sections, 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
- Description and motivation of the recommendation
- Addressees for whom the recommendation is intended
- 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 readings (articles, blog posts or academic papers providing additional information on the recommendation)
- Link to the Training Program

Example



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.

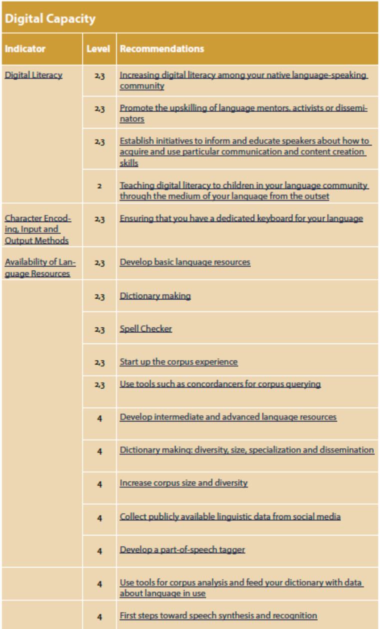
Related module in TP: 5

Recommendations for Digital Capacity

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).

Recommendations for Digital Capacity



Digital Capacity

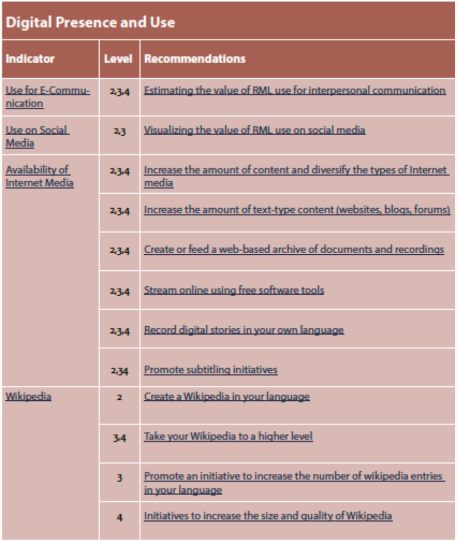
Indicator	Level	Recommendations
Digital Literacy	2,3	Increasing digital literacy among your native language-speaking community
	2,3	Promote the upskilling of language mentors, activists or disseminators
	2,3	Establish initiatives to inform and educate speakers about how to acquire and use particular 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

Recommendations for Digital Presence and Use

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

Recommendations for Digital Presence and Use



Digital Presence and Use

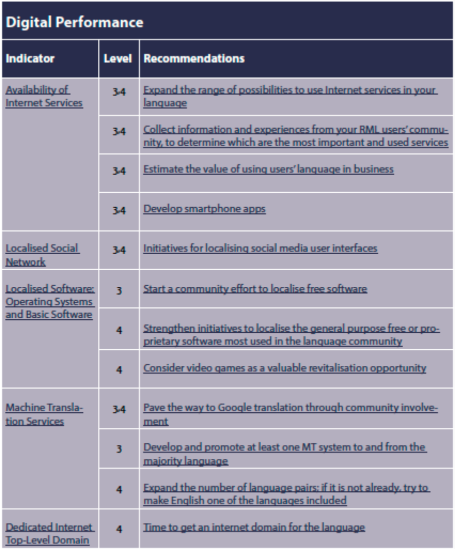
Indicator	Level	Recommendations
Use for E-Communication	2,3,4	Estimating the value of RML use for interpersonal communication
Use on Social Media	2,3	Visualizing the value of RML use on social media
Availability of Internet Media	2,3,4	Increase the amount of content and diversify the types of Internet 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

Recommendations for Digital Performance

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

Recommendations for Digital Performance



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

Addressees

The potential addressees of the DLSK are:

- Individuals
- Users’ groups, collectives, associations
- Research groups, software developers
- Companies
- Organisations, institutions, policy makers

Recommendations concerning Language Resources

- Develop basic language resources
    - Dictionary making
    - Spell Checker
    - Start up the corpus experience
    - Use tools such as concordancers for corpus querying
- Develop intermediate and advanced language resources
    - Dictionary making: diversity, size, specialization and dissemination
    - Increase corpus size and diversity
    - Collect publicly available linguistic data from social media
    - Develop a part-of-speech tagger
    - Use tools for corpus analysis and feed your dictionary with data about language in use
    - First steps toward speech synthesis and recognition

General features

An aggregator of information and available resources, presented in a reasoned way

No previous knowledge is required

Recommendations are more about creating awareness and introducing the concept than instructing about the technicalities

Repeatedly advised that this is work for specialists

Accent on re-use of publicly available data and on the need of informed copyright choices.

Dictionary making

Examples and best practices are provided regarding how to build a basic dictionary or enrich an existing one.

Explicit reference is made to Openwords, Poly, Webonary, LinguaLibre and Forvo

Plain, simple and no technical language is used.

Conclusions

Frame Title

I am not saying that everything was easy and perfect. Most of the things could have been done better. We learned a lot, mostly in terms of what should have been done!

A collective work

The Kit was developed in close cooperation with a wide board of experts (the DLDP Advisors)

Upon publication of the The Digital Language Survival Kit, we received extended support by the community, industry (e.g. Google) as well as requests for sharing, dissemination, and translation.

It was featured in the Irish press, on the Welsh television, and extensively on social media.

This must not end here

A word of caution

The Kit is a drop in the ocean: a first exploration in this direction and a proof of concept.

It has been applied to four European minority languages (Basque, Breton, Karelian, Sardinian) by adapting and localising the recommendations to the specific case of each language and community.

A lot more needs to be done, and this must continue.

Digital language planning as a bottom-up activity

All languages, minority ones in particular, must take digital language planning into consideration

Planning requires informed consideration of the technology available, of the actors to be involved, and above all of the needs, desires and expectations of the speakers' community

A methodology for digital language planning can help coordinate the efforts avoiding fragmentation and duplication and prioritise the resources to be invested

Digital language planning can and should involve minority language speakers, who can be drivers of the digital health of their languages.

Questions?

Backup slides

Sometimes, it is useful to add slides at the end of your presentation to refer to during audience questions.

The best way to do this is to include the `appendixnumberbeamer` package in your preamble and call `\appendix` before your backup slides.

metropolis will automatically turn off slide numbering and progress bars for slides in the appendix.

References i