

# How can we measure the Digital Language Vitality?

Irene Russo (ILC CNR) & Tuomo  
Salonen (KKS)



- Digital Language Vitality as the extent to which a language is usable, present, and actually used over the Internet and digital devices
- How do we measure Digital Language Vitality?

A scale + a set of indicators (+ a survey)

# Digital Language Vitality Scale

- **Pre-digital:** the language is not present on (online) digital media and it lacks the basic infrastructure for digital use.
- **Dormant:** some of the main preconditions for digital usage of a language are in place but there is no technological support for the use of the language.
- **Emergent:** internet penetration is good, speakers are digitally literate. Limited technological support but some basic language resources might be available.

# Digital Language Vitality Scale

- **Developing:** the language shows some usage over communication and social media and some digital media and services may be available. An online machine translation service or tool might be available.
- **Vital:** the language is used regularly for e-communication and on social media, some of which may have a localised interface. Language resources and services are widely available.
- **Thriving:** The language is used extensively and without any technological barriers in all current digital domains, from communicative to transactional ones. The latest technology is available.

# Digital Language Vitality Indicators

Indicator	
1. Evidence of connectivity	
2. Digital literacy	
3. Internet penetration or digital population size <sup>2</sup>	<i>digital capacity</i>
4. Character/script encoding	
5. Availability of language resources	
6. Use for e-communication	
7. Use on social media	<i>digital presence and use</i>
8. Availability of Internet media	
9. Wikipedia size	
10. Available Internet services	
11. Localised social networks	
12. Localised software	<i>digital performance</i>
13. Machine translation tools/services	
15. Dedicated Internet top-level domain	

Table 1: Digital Language Vitality Indicators

## 4 Digital capacity: character/script encoding

- The digital use of a language on the web is mainly written
- Unavailability of character/script encoding limits the digital usability of a language
- Check in the list of Unicode supported scripts

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

Table 5: Input and output methods

## 7 Digital presence & use: use on social media

- Social networking are linked the development of literature outside formal and social contexts
- No official ways to assess the quantity of content (a tentative is done for Twitter with the project [Indigenous Tweets](#))
- Rely on the answers from a survey



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

Table 8: Use on social media

## 9 Digital presence & use: Wikipedia size

- Wikipedia contributes to the overall vitality of a language, building a community
- Number of articles as the best indicator (but also number of active contributors and page views can be considered)
- Ideally a Wikipedia in a minority language should have distinctive content

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
large	5	between 100,000 and 1,000,000 articles
huge	6	over 1,000,000 articles

Table 10: Wikipedia

## 12 Digital performance: localised softwares

- Facebook is localised in 113 languages, Twitter 48
- No direct effect on the use, it has a positive influence on the perception of the language as modern and suited to being used in ICT contexts
- Check in the setting of the social networks. You can organise campaign to translate the interface

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.

Table 13: Localised software

# 13 Digital performance: machine translations tools

- Machine translation presupposes a wide array of tools and resources (i.e. the availability of big multilingual corpora)
- MT improves the status of a language, by increasing normality, literacy, and visibility
- As good indicators the number of different services/tools (e.g. Google and Bing services) and the number of language pairs

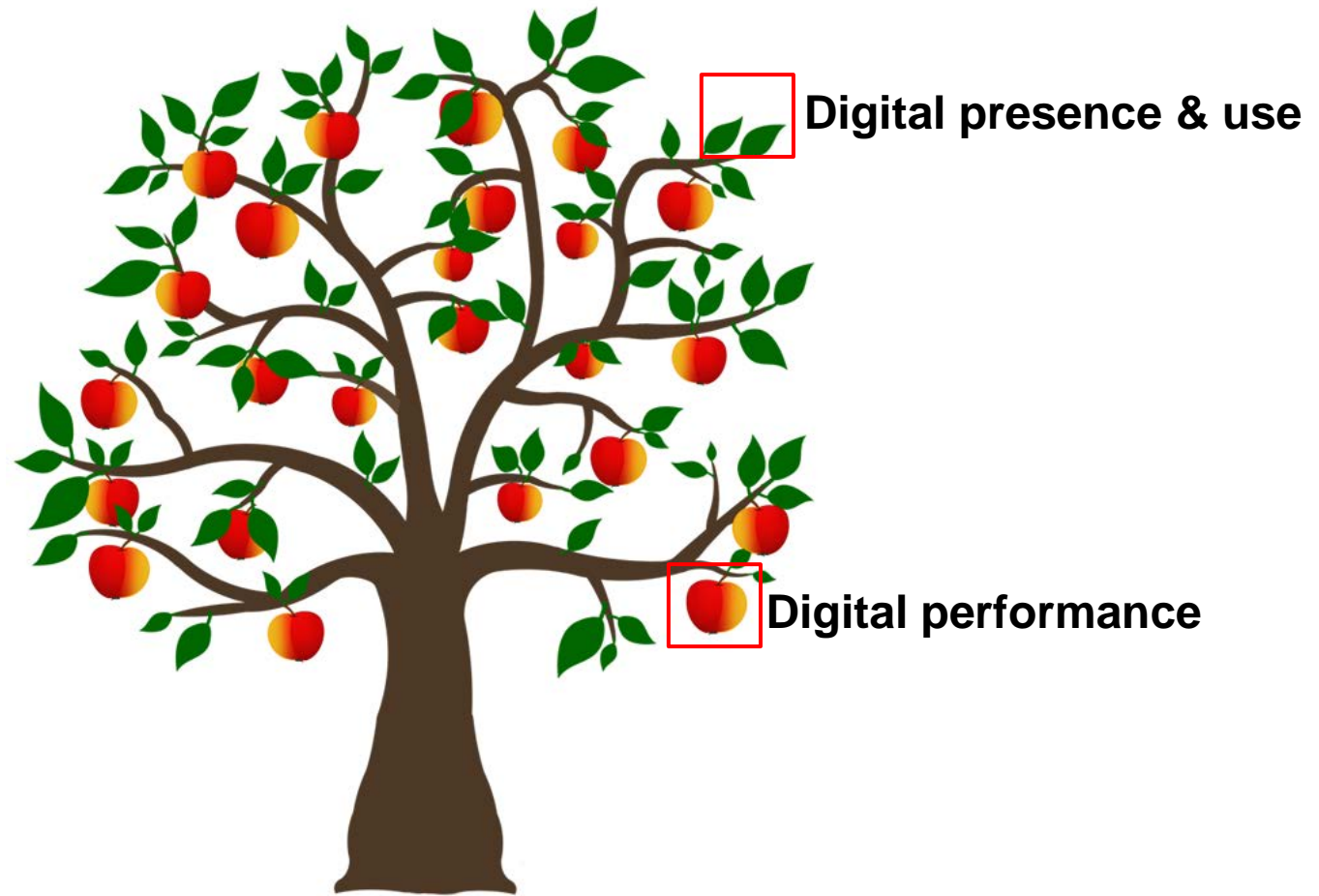
Label	Score	Micro Indicators
none	3	No MT for the language
basic	4	at least one (online) service/ tool, at least one language pair or one direction
medium	5	at least one (online) service / tool, at least two language pairs in both directions
advanced	6	more than one (online) service /tool, more than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

Table 14: Machine translation tools/services

# How to assess Digital Language Vitality with a survey

- For several indicators we don't have official statistics
- Ask activists' opinion or make a survey
- DLDP has a good template for the [survey](#): 31 questions to understand the current digital usage and needs of a given (RM) language by its language speakers community





**Digital capacity**

# Applying the scale to Karelian

- A short introduction
- Karelian Survey
- Digital capacity
- Digital use
- Digital performance
- Karelian digital vitality

# Karelian: a short introduction

- Tens of thousands of speakers in Finland and Russia
- A Finnic language with 3 main dialects
- Non-regional in Finland since World War II
- No official status
- Subjected to strong assimilationist policies
- Severely endangered

# Karelian survey

- Distributed through social media, Karelian online news and Karelian organizations
- 156 respondents representing all 3 dialects
- 46,8% of respondents over the age of 60

# Digital capacity

- High digital literacy and internet penetration
- Informal script encoding (Finnish with Sámi)
- Minimal language resources (Karelian online dictionary)

# Digital use

- Occasional use in e-communication, e-Mail preferred
- Use on social media concentrates on Facebook
- Advanced internet media
- Two Wikipedias: Livvi Karelian with 2633 articles and Karelian Proper still in incubator

# Digital performance

- Medium-level internet services: mainly online news, newspapers and edu-tainment
- No localised software or social networks
- No machine translation tools
- No dedicated internet domain

# Karelian digital vitality

- Overall score of 3.29
- Emergent vitality level
- Good connectivity and penetration, emergent activity on a range of domains
- Issues with localisation, encoding language resources