Assessing Digital Language Support on a Global Scale

Gary F. Simons
SIL International
gary_simons@sil.org

Abbey L. Thomas
The University of Texas at Dallas
abbey.thomas@utdallas.edu

Chad K. White
SIL International
chad_white@sil.org

Abstract

The users of endangered languages struggle to thrive in a digitally-mediated world. We have developed an automated method for assessing how well every language recognized by ISO 639 is faring in terms of digital language support. The assessment is based on scraping the names of supported languages from the websites of 143 digital tools selected to represent a full range of ways that digital technology can support languages. The method uses Mokken scale analysis to produce an explainable model for quantifying digital language support and monitoring it on a global scale.

1 Introduction

The users of endangered languages struggle to thrive in a digitally-mediated world. The opportunities afforded by digital technology differ drastically depending on the language being used. This has been dubbed the "digital language divide" (Mikami, 2008; Young, 2015; Soria, 2016; Matsakis, 2019). As digital modes of communicating and accessing information become increasingly necessary in daily life, lack of digital language support (DLS) for a language means that its speakers must use other languages to participate in the global information society or be left out.

Linguists have been writing for decades about the role digital technology could play in language revitalization (Warschauer, 1998; Buszard-Welcher, 2001; Eisenlohr, 2004; Galla, 2009; Holton, 2011; Cru, 2016). Language technologists are recognizing the inequities facing the vast majority of the world's languages (Bird, 2020; Blasi et al., 2022) and are embracing the challenges of bringing greater equity in DLS (Joshi et al., 2019; Bapna et al., 2022; Edunov et al., 2022).

However, in a world where most people are multilingual and each language fits into its functional niche within an ecology of languages (Lewis and Simons, 2016), full digital support for every language is not a realistic goal nor what those multilingual individuals are necessarily looking for (Bird, 2022). The goal of our research is to develop a method for measuring DLS in every language, so that it will be possible to provide an empirical view of the digital state of the world's languages and to observe the progress as so-called low-resource languages move toward crossing the digital language divide.

2 Related Work

Our primary inspiration has been the seminal work by Kornai (2013) on developing a method for assessing the digital vitality of any language. He proposes a four-way classification of languages as digitally Thriving, Vital, Heritage, or Still, "roughly corresponding to the amount of digital communication that takes place in the language." His method harvests data from the Web, then uses supervised classification to automatically label all known languages. In practice, he adds a fifth level, Borderline, to represent languages that show signs of crossing the gap from Still to Vital. He and his colleagues have applied this method to the languages of India (Kornai and Bhattacharyya, 2014), the former Soviet Union (Kornai, 2015), and the Uralic family (Acs et al., 2017).

In reviewing Kornai's method, Gibson (2015, 2016) focused on the huge gap between Still and Vital. He argues that two additional levels are needed to fill this gap: one for when the needed elements (like a keyboarding solution) are in place for potential digital language use, and another for when digital language use is indeed taking off. We follow Gibson's lead in adding two levels, but use names that achieve better congruence with the geometry of the S-curve model that emerges from our method (see Figure 3).

3 Requirements

Following Kornai's (2013) lead, we seek to develop an automated method for assessing digital

Proceedings of the 29th International Conference on Computational Linguistics, pages 4299–4305
October 12–17, 2022.

language vitality that is based on feature data harvested from the Web. In this way, it can be run periodically to monitor changes in digital vitality for every language. We were motivated to develop an alternative to Kornai's method of analysis in order to meet three requirements:

Digital vitality should be orthogonal to non-digital vitality. We exclude features like population and language vitality from the feature data. Kornai notes that the EGIDS level as reported in Ethnologue (Lewis and Simons, 2010; Eberhard et al., 2022) is "the best predictor of digital status." But digital vitality is distinct from non-digital vitality. For instance, our method reports the "dead" language Latin to be the 80th most digitally vital language in the world. By contrast, Aimaq with nearly two million speakers is found to be digitally Still.

The assessments should be explainable. A standard critique of machine learning models based on black-box methods is that the models cannot explain why they produce the answers they do (Arrieta et al., 2020; Miller, 2019). Kornai (2013) bases his results on the majority outcome from 100 runs of a black-box model that yields a slightly different result each time. Users will be more likely to trust results if they are deterministic and explainable.

The assessment scale should measure a single underlying trait. The data features used by Kornai (2013) covered a variety of digital uses. Some had to do with quantifying the extent to which the language has been documented in digital archives by researchers. Others, like the sizes of Wikipedias, had to do with quantifying the extent of digital language use by the language community itself. Still others looked at specific software products and recorded which languages they support. These strike us as three distinct traits, each of which should be assessed in its own right: digital language preservation, digital language use (DLU), and digital language support (DLS). Of these, the latter two are what speak to monitoring the digital vitality of a language as it moves toward crossing the digital language divide. DLU and DLS are distinct traits that should be assessed separately—speakers of unsupported languages may nevertheless use it digitally (for instance, making do in texting; see Eberhard and Mangulamas (2022)), while speakers of supported languages may choose to use digital resources in another language they know.

We have chosen to focus on DLS since the data for monitoring that phenomenon are openly accessible—the developers of digital tools are usually keen to advertise all of the languages they support. By contrast, data on actual digital use is typically not shared on a language-by-language basis by the vendors concerned. A comparable effort to assess DLU on a global scale is much needed, though we anticipate that it will be significantly harder to acquire the needed data.

4 Methodology

The method we have adopted for building an explainable model of DLS is Mokken scale analysis (Mokken, 1971; Schuur, 2003). Mokken's method is a generalization of the more widely known Guttman scaling (Guttman, 1950). In the latter, the items in a scale form a strict hierarchy. If a subject has an item on the scale, then all lower items also apply. A subject's score on the scale is thus the highest item that is true for the subject.

Intuitively, DLS has these properties. If a language has a good virtual assistant (like Siri), then we can infer that it also has good machine translation—but having good machine translation does not imply having a good virtual assistant. Similarly, if a language has good machine translation, we can guess that it must also have good spell checking, though we cannot assume that the reverse would hold. In a Guttman scale, an exception to the hierarchical ordering is considered an error, but in an arena like DLS we can expect there to be exceptions. Mokken scaling is a method for placing the items of a supposed hierarchical scale into their optimal order, while providing metrics that allow one to evaluate how well the hierarchical model fits.

4.1 Categories of Digital Language Support

The method uses the following seven categories of DLS. They are listed below from easiest (most commonly supported) to hardest (least commonly supported) as determined by the results of our analysis:[1]

- Content — A service offering content in many languages (like Wikipedia, news sites, or Bible sites)[2]

[1] This aspect of the analysis is explained in subsection 5.2 and illustrated in Figure 2.

[2] Having digital content in a language could also be viewed as an evidence of digital language use. We treat the fact that a service offers content in a language as a Boolean indicator of

- Encoding — A system component for representing languages (like keyboards and fonts)
- Surface — A tool with surface-level processing (like spell checking or stemming)
- Localized — A tool with a localized user interface (like operating system, browser, or messaging)
- Meaning — A tool with meaning-level processing (like machine translation)
- Speech — A tool for speech processing (like speech-to-text or text-to-speech)
- Assistant — An intelligent virtual assistant (like Siri or Alexa)

For each category, we sought to identify the top ten tools of its kind globally. In order to ensure that we included the major tools in use outside the English-speaking world, we also included the top five tools in each of the ten most populous countries of the world.[3] The reference authority for these rankings was the *similarweb* service.[4] Then we added any tools found from other sources that supported more than 10% of the median number of languages supported by the top tools in the category. In order for a tool to be used in our analysis, we required there to be a URL from which the names or ISO 639 codes of supported languages could be scraped.

The full sample consists of URLs for 143 digital tools across the seven categories of DLS.[5] The number of tools in each category is shown in Table 1 as the maximum number in the range for level 4.

4.2 Harvesting the feature data

The method works by scraping each URL in the sample to discover what languages each tool supports. The harvested language names are mapped to their corresponding ISO 639-3 code[6] by means of a manually maintained table of name-to-code mappings. After the mapping of the harvested language names, the resulting feature data is a logical matrix with rows for 7,829 ISO 639-3 codes, columns for the 143 digital tools, and a Boolean value at the intersection indicating whether the given language is supported by the given tool.

4.3 Scoring the DLS categories as subscales

When a language is not supported by any tools in a given DLS category it is scored as 0; otherwise, the number of tools supporting that language is converted to a level score on a four-level subscale. The correspondence between the number of tools supporting the language and the level on the subscale is shown in Table 1. The score corresponds to the quartile in the distribution of the number of tools supporting each language; only the languages that are supported by at least one tool in the category are included in that distribution.[7]

Category	Levels			
	1	**2**	**3**	**4**
Assistant	1	2	3–4	5–11
Speech	1	2–3	4–8	9–23
Meaning	1	2	3–6	7–14
Localized	1	2	3–12	13–47
Surface	1	2	3	4–15
Encoding	1	2	3	4–10
Content	1	2	3	4–23

Table 1: Number of tools supporting a language in each level of the subscales for the DLS categories

4.4 From category levels to scale items

In constructing the Mokken scale, the levels of the categories become items in the scale. These items are named Content1, Content2, and so on. Within each subscale, the items form a strict hierarchy, in which being scored at a higher level on the subscale implies also having at least as much support as the lower levels of the same subscale. Thus the count of languages for item Content3 also includes the languages for Content4, and so on going down. The bar graph in Figure 1 shows the items listed from top to bottom in ascending order of the number of languages with at least that level of support in the named category.

5 Results

5.1 Evaluating fit of the model

Mokken scale analysis allows us to evaluate the degree to which the scale depicted in Figure 1 forms

support for the language. To measure digital language use, we would quantify the amount of digital content in each language.

[3]This sampling method allows us to discover widely-used tools that support just one large language, but it admittedly misses tools that have been custom-built for a single smaller language.

[4]https://similarweb.com

[5]A complete list of the 143 digital tools is provided at https://github.com/sil-ai/dls-results.

[6]https://iso639-3.sil.org/code_tables

[7]The quartile boundaries are extended upward to accommodate ties; thus in every case, Level 1 contains more than 25% of the languages with that kind of support.

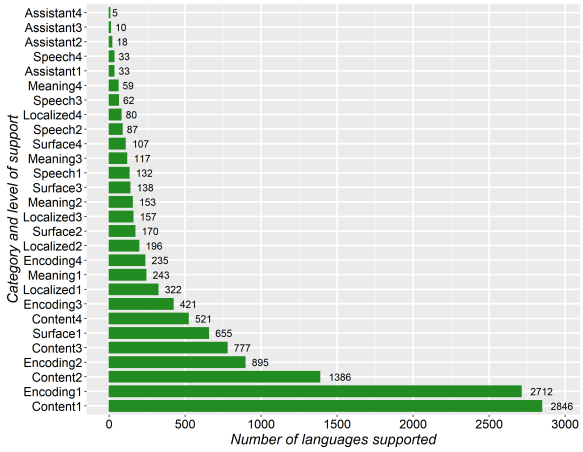

Figure 1: Number of languages supported at each category and level of digital language support

a hierarchical scale. This is done using Loevinger's (1948) coefficient of homogeneity, H.[8] H compares the actual Guttman errors to the expected number of errors if the items were not related in a scale. A value of 1.0 indicates no errors; any value above 0.5 is indicative of a strong scale (Sijtsma and Molenaar, 2002).

Item	H
Assistant	0.987
Speech	0.942
Meaning	0.920
Localized	0.924
Surface	0.885
Encoding	0.707
Content	0.685
Full scale	0.825

Table 2: Coefficient of homogeneity, H, for DLS scale

The results in Table 2 show that the proposed DLS scale is a very strong scale, especially among the categories of support that are hardest to achieve. Thus the total score on all 7 categories (i.e., 0 to 28) serves to quantify the DLS for a given language.

5.2 Relative difficulty of DLS items

Mokken analysis is based on Item Response Theory (IRT)—a methodology developed for educational and psychological testing (Lord, 1980). In IRT, logistic regression is used to derive an Item Response Function (IRF) for each test item; it returns the probability that a subject would produce a positive (or correct) response on that item, given their total score on the rest of the test items. The difficulty of an item is defined as the score (on the rest of the test) at which the subject has a 50% chance of giving a positive response for the item. Figure 2 plots the difficulty for each of the scale items listed in Figure 1. For instance, a language has a 50% chance of getting its first spell-checker (Surface1) if it has 3.6 other DLS items, but the first virtual assistant (Assistant1) cannot be expected until it has 23.4 other DLS items.

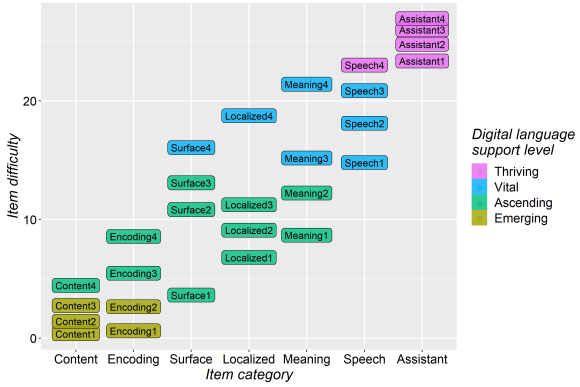

Figure 2: Difficulty of the DLS categories and levels

5.3 DLS as a growth curve

Figure 3 plots the DLS score for 7,829 ISO 639 languages. The vertical axis is the measure of DLS as a proportion: the DLS score achieved divided by the maximum possible score.[9] The horizontal axis is the rank of the language by DLS score, but converted to a log scale and flipped so that lowest DLS is on the left and highest is on the right.

The pattern that emerges is an S-curve as is typical in studies of growth in innovation. We follow the geometry of the fitted curve to assign each language to one of the five summary levels:

- Still — a score of 0
- Emerging — at the bottom where the slope is more horizontal than vertical
- Ascending — below the midpoint where the slope is more vertical than horizontal
- Vital — above the midpoint where the slope is more vertical than horizontal

[8] We have performed these calculations using the "mokken" package (van der Ark, 2007, 2012) in R (R Core Team, 2022).

[9] The DLS scores are also adjusted by scoring each item as the probability returned by its IRF. In educational testing, scoring each positive response as a probability is a way of controlling for random guessing on questions that are too hard for the subject. In the application to DLS it can control for "random" developments that do not have the underpinnings of the expected lower categories of support, such as when there is a one-time philanthropic gesture by a large company or the potentially unsustainable efforts of a solitary developer.

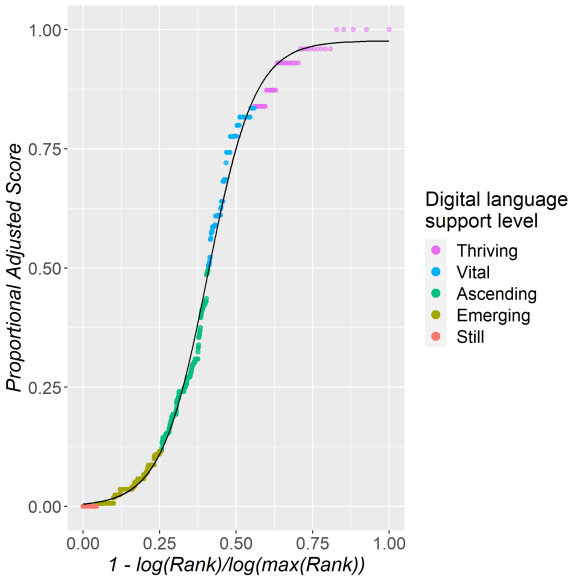

Figure 3: The growth of DLS as a logistic function.

- Thriving — at the top where the slope is more horizontal than vertical

By comparing Figures 2 and 3 one sees what components of DLS correspond to the summary levels.

Table 3 reports the number of languages at each summary level along with the names of example languages, the first being from the upper end of the range and the second from the lower.[10]

Level	Languages	Examples
Thriving	33	English, Hungarian
Vital	95	Nepali, Tongan
Ascending	401	Greenlandic, Hunsrik
Emerging	3304	Dogri, Michif
Still	3996	Aimaq, Yurok

Table 3: Number of languages per DLS level

6 Conclusion

We have presented a method that produces an explainable model for quantifying DLS. We are currently working with Ethnologue to add reporting on DLS in its description of languages, beginning with the next edition. Regularly updating the assessments should serve to document the digital trajectory of every known language.

[10] A sampling of the detailed results produced by the system is provided at https://github.com/sil-ai/dls-results.

Acknowledgements

This research has been funded by the Ethnologue program of SIL International for the purpose of developing a way to add digital language vitality to its reporting on the state of the world's languages. At the outset of the project, SIL International licensed the software developed by Kornai (2013) from the Hungarian Academy of Sciences. We are deeply indebted to Andras Kornai and his PhD student at the time, Katalin Pajkossy, for their help and encouragement as we first replicated his results before developing the method described in this paper. Others who made significant contributions include Steve Moitozo, Erica Oldaker (née Swindle), Steve Woolston, and Daniel Whitenack.

References

J. Acs, K. Pajkossy, and A. Kornai. 2017. Digital vitality of Uralic languages. *Acta Linguistica Academica*, 64(3):327–345.

A. B. Arrieta, N. D. Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

A. Bapna, I. Caswell, J. Kreutzer, O. Firat, D. van Esch, A. Siddhant, M. Niu, P. Baljekar, X. Garcia, W. Macherey, T. Breiner, V. Axelrod, J. Riesa, Y. Cao, M. X. Chen, K. Macherey, M. Krikun, P. Wang, A. Gutkin, A. Shah, Y. Huang, Z. Chen, Y. Wu, and M. Hughes. 2022. Building machine translation systems for the next thousand languages. Technical Report arXiv:2205.03983.

S. Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

S. Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

L. Buszard-Welcher. 2001. Can the Web help save my language? In L. Hinton and K. Hale, editors, *The Green Book of Language Revitalization in Practice*, pages 331–345. Academic Press, San Diego, CA.

J. Cru, editor. 2016. *Digital Media and Language Revitalisation*. Linguapax Review 2016. Linguapax International.

D. M. Eberhard and M. Mangulamas. 2022. Who texts what to whom and when? Patterning of texting in four multilingual minoritized language communities and a preliminary proposal for the language repertoire matrix. *International Journal of the Sociology of Language*, 2022(276):169–205.

D. M. Eberhard, G. F. Simons, and C. D. Fennig, editors. 2022. *Ethnologue: Languages of the World*, twenty-fifth edition. SIL International, Dallas.

S. Edunov, P. Guzman, J. Pino, and A. Fan. 2022. Teaching AI to translate 100s of spoken and written languages in real time. *Meta AI*.

P. Eisenlohr. 2004. Language revitalization and new technologies: Cultures and electronic mediation and the refiguring of communities. *Annual Review of Anthropology*, 18(3):339–361.

C. K. Galla. 2009. Indigenous language revitalization and technology: From traditional to contemporary domains. In J. Reyhner and L. Lockard, editors, *Indigenous Language Revitalization: Encouragement, Guidance, and Lessons Learned*, pages 167–182. Northern Arizona University, Flagstaff, AZ.

M. L. Gibson. 2015. A framework for measuring the presence of minority languages in cyberspace. In *Linguistic and Cultural Diversity in Cyberspace. Proceedings of the 3rd International Conference*, pages 61–70, Moscow. Interregional Library Cooperation Centre.

M. L. Gibson. 2016. Assessing digital vitality: Analytical and activist approaches. In *Proceedings of the LREC 2016 Workshop "CCURL 2016—Towards an Alliance for Digital Language Diversity"*, pages 46–51.

L. Guttman. 1950. The principal components of scale analysis. In S.A. Stouffer, editor, *Measurement and Prediction*, pages 312–361. Wiley, New York.

G. Holton. 2011. The role of information technology in supporting minority and endangered languages. In P.K. Austin and J. Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, pages 371–399. Cambridge University Press, Cambridge, UK.

P. Joshi, C. Barnes, S. Santy, S. Khanuja, S. Shah, A. Srinivasan, S. Bhattamishra, S. Sitaram, M. Choudhury, and K. Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Proc. of the 16th International Conference on Natural Language Processing,*, pages 211–219, Hyderabad, India.

A. Kornai. 2013. Digital language death. *PLoS ONE*, 8(10):e77056.

A. Kornai. 2015. A new method of language vitality assessment. In *Linguistic and Cultural Diversity in Cyberspace. Proceedings of the 3rd International Conference*, pages 132–138, Moscow. Interregional Library Cooperation Centre.

A. Kornai and P. Bhattacharyya. 2014. Indian subcontinent language vitalization. In *Proc. 2014 LREC Workshop on Indian Language Data: Resources and Evaluation (WILDRE2)*, pages 24–27.

M. P. Lewis and G. F. Simons. 2010. Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique*, 55(2):103–120.

M. P. Lewis and G. F. Simons. 2016. *Sustaining language use: Perspectives on community-based language development*. SIL International, Dallas.

J. Loevinger. 1948. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, 45(6):507.

F. M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Mahwah, NJ.

L. Matsakis. 2019. Bridging the Internet's Digital Language Divide. *Wired, 13 June 2019*.

Y. Mikami. 2008. Digital language divide: Measuring linguistic diversity on the Internet. In *UNESCO/UNU Conference on Globalization and Languages: Building on our Rich Heritage*, pages 27–28, Tokyo, Japan.

T. Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

R. J. Mokken. 1971. *A Theory and Procedure of Scale Analysis: With Applications in Political Research*, volume 1. Walter de Gruyter.

R Core Team. 2022. *R: A language and environment for statistical computing*. Vienna, Austria.

W. H. Schuur. 2003. Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, 11(2):139–163.

K. Sijtsma and I. W. Molenaar. 2002. *Introduction to nonparametric item response theory*. Sage, Thousand Oaks, CA.

C. Soria. 2016. What is digital language diversity and why should we care? In J. Cru, editor, *Digital Media and Language Revitalisation. Linguapax Review 2016*, pages 13–28. Linguapax International.

L. A. van der Ark. 2007. Mokken scale analysis in R. *Journal of Statistical Software*, 20(11):1–19.

L. A. van der Ark. 2012. New developments in Mokken scale analysis in R. *Journal of Statistical Software*, 48(5):1–27.

M. Warschauer. 1998. Technology and indigenous language revitalization: Analyzing the experience of Hawai‘i. *Canadian Modern Language Review*, 55:140–161.

H. Young. 2015. The digital language divide: How does the language you speak shape your experience of the internet? *The Guardian, 28 May 2015*.