WikiLingDiv: a dataset for quantifying digital linguistic diversity using Wikipedia page views

Hannes Essfors
TU Wien Informatics,
Favoritenstraße 9-11,
1040 Vienna, Austria
hannes.essfors@tuwien.ac.at

Andreas Baumann
Faculty of Philological and Cultural Studies,
University of Vienna, Universitätsring 1,
1010 Vienna, Austria
andreas.baumann@univie.ac.at

Abstract

With the conflation of digital and non-digital spaces, and NLP technologies being integrated into an increasing number of aspects of daily life, linguistic diversity cannot be fully understood without considering language use in the digital space. To facilitate such research, we introduce WikiLingDiv, an openly accessible dataset for quantifying linguistic diversity in online knowledge retrieval using Wikipedia page views, corresponding to one dimension of digital linguistic diversity. Our dataset is based on yearly page views of 340 language editions of Wikipedia, aggregated across 239 countries and territories over 10 years (2015-2024). Using the dataset, we illustrate spatial and temporal patterns of digital linguistic diversity, suggesting that diversity has both increased and decreased across countries and regions, while highlighting country-specific dynamics in language usage. We release the dataset as an openly available and easily integrable data resource for researchers in computational linguistics, digital humanities, and the broader social sciences, enabling further work on linguistic variation and digital inequality.

1 Introduction

In the postdigital world, the lines between digital and non-digital have become blurred, with communication through digital media and tools moving from representing non-standard use of language to a non-optional part of daily interaction in large parts of the world. As a matter of fact, in December 2025, 73.2% of the global population were internet users and 68.7% were social media users (Statista, 2025a), while during Covid-19, digital working became the de facto standard for much of the knowledge-based part of the global economy (Marsh et al., 2022).

This raises the question whether traditional models of linguistic diversity as a function of the relative abundance of L1 speakers (Greenberg, 1956; Ginsburgh and Weber, 2020; Grin and Fürst, 2022; Civico, 2025) remain sufficient—notably, the bias towards English in the digital space would suggest a large discrepancy between native language and the actual linguistic landscape (Benson et al., 2025). Instead, an operationalization of linguistic diversity based on language usage might be considered a more direct model of how language facilitates the tasks of, e.g., retrieving and producing information in the process of between-human interaction. While modelling language usage in the non-digital space intuitively appears difficult, it presents itself as fairly straightforward in the digital space due to the extensive datafication it has been subjected to (Mejias and Couldry, 2019; Flyverbom, 2019). Notably, such approaches towards linguistic diversity have been taken by, e.g., Magdy et al. (2014); Alshaabi et al. (2021); Hiippala et al. (2019); Väisänen et al. (2022) who assess digital linguistic diversity based on language production on social media, specifically Twitter/X and Instagram. These approaches come with three main caveats: Firstly, social media data is highly sensitive and thus restricted in its usage. For instance, since 2023, the Twitter API is no longer freely available for research (Blakey, 2024). Secondly, using social media data and similar artifacts to model digital linguistic diversity is crucially reliant on high accuracy in the NLP-task of language identification, which is notably challenging for multilingual documents and short texts (Jauhiainen et al., 2019), both of which are common features of social media posts (Alshaabi et al., 2021). Thirdly, it has been shown that only a small fraction of tweets are geotagged (Pfeffer et al., 2023), severely limiting the possibility of analyzing linguistic diversity as a spatial phenomenon.

To address some of these shortcomings and facilitate research on digital linguistic diversity as an approach towards modelling human culture and society, we here introduce *WikiLingDiv*—a dataset

Proceedings of SIGHUM 2026, pages 202–211

for quantifying digital linguistic diversity using Wikipedia page views. Our approach utilizes the fact that Wikipedia is a multilingual, widely used reference work, freely available in most of the world, and as of August 2025, the 7th most visited website globally (Statista, 2025b). Furthermore, the Wikimedia Foundation, which hosts Wikipedia, is committed to open data principles, making records and web activity associated with Wikipedia openly available through the Wikimedia Analytics API. This allows us to collect the number of page views associated with any given language edition of Wikipedia for 239 countries and territories over 10 years, which are made openly accessible through a CC BY-SA 4.0 license. Using this as a basis, we then normalize the page views to yield the proportion of attention given to each language across countries and years, based on which formal measures of diversity are derived.

Explicitly, our approach aims to shed light on digital linguistic diversity as a dynamic phenomenon arising when humans interact with—and through—digital technology. It follows that our work builds on two key assumptions: 1) Wikipedia page views approximate interest in retrieving knowledge in a given language, and 2), Patterns of Wikipedia access are impacted by socioeconomic factors such as internet access and level of education. Therefore, WikiLingDiv does not model digital linguistic diversity as a product of textual output, but rather as a proxy for language consumption. Thus, it complements work focusing on diversity in existing language technology infrastructure, such as the *Digital Language Equality Metric* (DLE) (Gaspari et al., 2022; Grützner-Zahn and Rehm, 2022) of the European Language Equality (ELE) project, which assesses the availability of technology and support for different languages. Inherently, these two dimensions of digital linguistic diversity are linked: the language used depends on the available technological support for that language, and the available support partially reflects the linguistic demand. To understand the extent of this interaction, data on digital linguistic usage and the relative demand of digital technology is necessary, a gap which WikiLingDiv seeks to help to address.

The paper is structured as follows. First, we describe the process of compiling and structuring the dataset, as well as the measures taken in harmonizing the data to allow for integration with other datasets. Then, we provide a descriptive analysis of the dataset. Finally, we showcase how WikiLingDiv can be utilized in a diversity analysis to address the research question of how digital linguistic diversity—from the perspective of language consumption in online knowledge retrieval—has changed over the last decade.

2 Creating the dataset

For the creation of the dataset and all subsequent analysis, the statistical programming language R (4.5.2) (R Core Team, 2025) was used.[1] As alluded to in the introduction, our empirical strategy utilizes the Wikimedia Analytics REST API to understand how users interact with the language-specific Wikipedia editions.[2] More specifically, we use the httr package (Wickham, 2023) to acquire a list of all available Wikimedia projects through the following API-request: `https://commons.wikimedia.org/w/api.php?action=\protect\@normalcr\relaxsitematrix&smtype=language&format=json`. We loop through the list and collect the 2 or 3-letter language codes used by Wikimedia to denote each project, e.g., /en/ for English or /ceb/ for Cebuano, together with its name in Latin script. For each language code, we then query the REST API to acquire the page views per country. The endpoint takes requests according to the following structure: /pageviews/top-by-country/project/access/year/month. We specify *project* using the respective language codes; we specify *access* to 'all-access' to capture both desktop and mobile users; and we specify the time frame as each year and month combination beginning with 2015, since the endpoint serves data from this year and onwards.

From each request to the API endpoint, we receive a value for the number of page views from each country for which at least one pageview exists in a given year and month. However, to protect user privacy, the Wikimedia Foundation does not publish the exact number of page views but instead ceils to the nearest 1000 (Wikimedia Foundation, 2024), meaning that the values are estimates. For example, if two language editions receive 1012 and 1756 views, respectively, they are both assigned

[1] All code associated with the paper is made available at `https://github.com/Eszettfors/WikiLingDiv`. The dataset is published at `https://zenodo.org/records/18526766`

[2] For the exact API documentation provided by the Wikimedia Foundation, see `https://doc.wikimedia.org/generated-data-platform/aqs/analytics-api/`

the value 2000. It naturally follows that the more views an edition has, the smaller the rounding error proportionally is. We then aggregate the values for each year from 2015 to 2024 by summing the page views for each language edition and country across all 12 months. Finally, we restructure the dataset as a time series in long format such that each record consists of a uniquely defined year-country-language triplet specified by the number of page views.

Unfortunately, the page views by country endpoint does not allow specification of the nature of the agent, and accordingly, we cannot purposefully exclude self-identified bots from the statistics. While this naturally introduces some bias to the data, systematic scraping constitutes a fairly stable bias within countries—especially given the low granularity scope the data is aggregated at—meaning that the relative abundances we are seeking to compute is expected to remain robust. Furthermore, we would argue that a substantial fraction of automated information retrieval is indeed driven by human information needs. With this in mind, we suggest the aggregated page views per language and country be interpreted as a proxy for the relative interest in a particular language in that country.

2.1 Data harmonization

To harmonize and make the dataset interoperable, we address the language codes used by Wikipedia, which inconsistently employ both three- and two-letter language codes as well as multiple non-standard identifiers (Wikimedia Foundation, Inc., 2025). There are two major language catalogues that we intend to make the dataset easily mergeable with: Glottolog (Hammarström et al., 2025) for linguistic typology and Ethnologue (Eberhard et al., 2025) for speaker numbers. For this purpose we add both ISO 639-3 codes (ISO, 2023) and glottocodes. Since ISO-codes can denote both macrolanguages and individual languages, such as /ara/ being the macrolanguage identifier encompassing all Arabic varieties—multiple of which have their own Wikipedia language edition—we avoid using macrolanguage identifiers, unless they have a corresponding Glottocode. This is, for example, the case with Serbo-Croatian (hbs), which is recognized as an individual language in Glottolog with the respective standardized varieties Bosnian (bos), Serbian (srp) and Croatian (hrv), which are classified as dialects in Glottolog and thus have a corresponding Glottocode. As a result, both Glottocodes and ISO639-3 function as unique language identifiers in the dataset, allowing for easy integration with, e.g., knowledge-bases used in multilingual NLP such as URIEL+ (Khan et al., 2025).

3 Descriptive analysis

All in all, the dataset comprises 295,510 year-language-country triplets covering a total of 340 languages and 239 countries/territories across 10 years (2015 - 2024). As seen in Figure 1a), the distribution of page views across countries remains fairly stable throughout the time period. While a noticeable increase in median and mean page views can be observed from 2015 (mean = $10^{7.38}$; median = $10^{7.37}$ to 2016 (mean = $10^{7.66}$; median = $10^{7.57}$), only small fluctuations are distinguished until 2022, after which it decreases. Furthermore, the geographical spread of the page views is by no means homogeneous: the boxplots are consistently layered according to continents across the time span, with countries in Oceania and Africa generally making up the first and second quartiles; North America the second and third; and Europe, Asia and South America the third and fourth. The number of countries also remains stable around 235, with a sharp decline to only 200 countries in 2024. This can reasonably be assumed to be a result of the Wikimedia Foundation's data publication guidelines, as the missing countries for 2024 align with the list of countries and territories designated as protected due to the higher risk posed to individuals from their Wikipedia activity (Wikimedia Foundation, 2025). Because of this, the dataset only contains complete time series for 196 countries and territories, which, however, increases to 232 if the year 2024 is excluded.

Country	Page Views (%)	Cumulative
U.S.	22.50	22.50
Japan	7.06	29.56
Germany	6.35	35.91
U.K.	5.58	41.49
India	4.82	46.31
France	4.03	50.34
Italy	3.62	53.95
Russia	3.44	57.39
Canada	2.76	60.16
Spain	1.99	62.15

Table 1: Top 10 countries as a source of Wikipedia page views across the years 2015 - 2024.

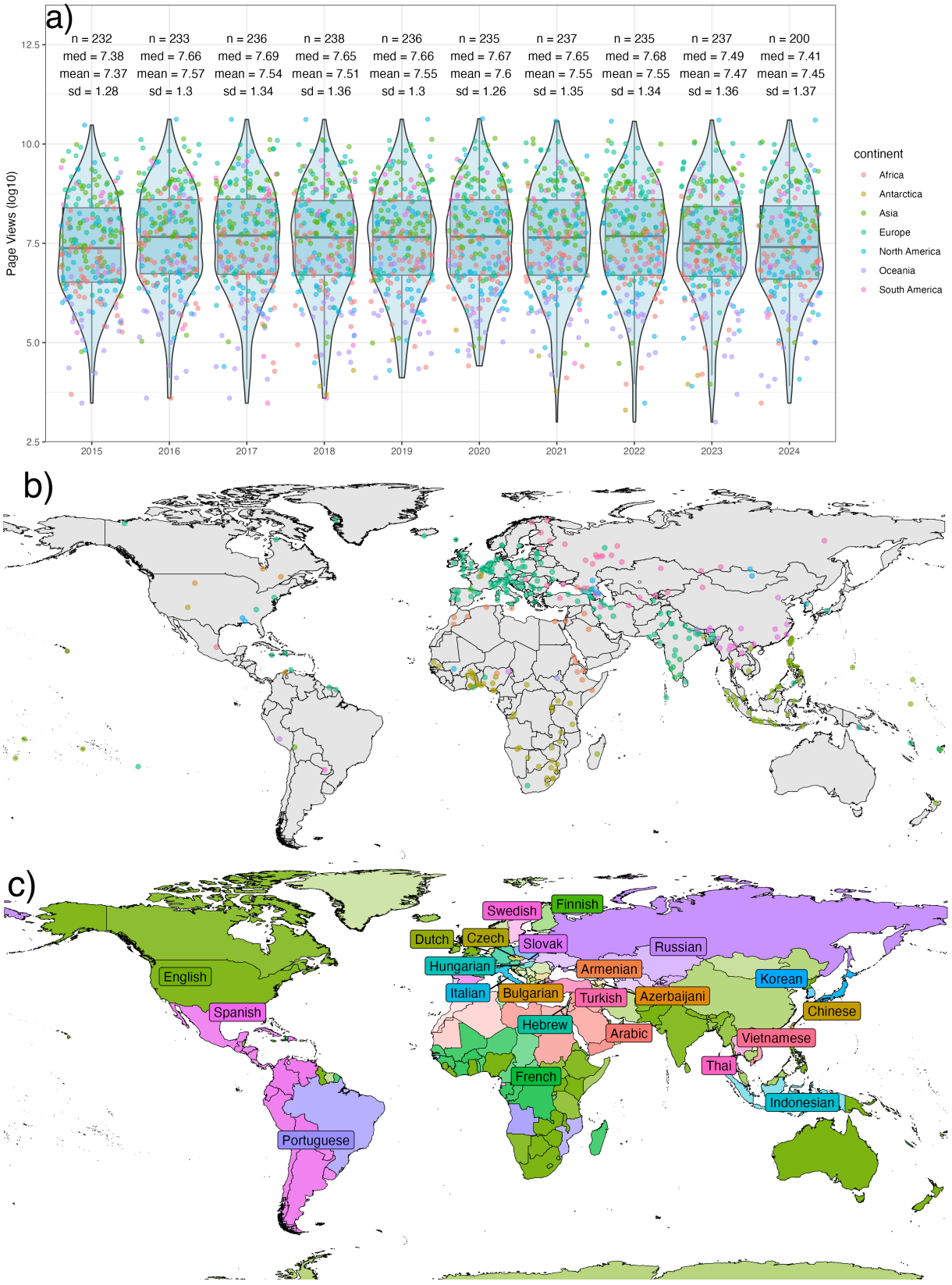

Figure 1: a) Boxplots overlayed by violinplots depicting the distribution of yearly total page views per country from 2015 to 2024, scaled using a base-10 logarithm. Each point corresponds to a country or territory, which is coloured according to its geographical continent of belonging. b) Geographical distribution of languages present in the dataset, coloured according to language family. The coordinates of each language were taken from Glottolog. c) World map with each country coloured according to the most visited language edition of Wikipedia between 2015 and 2024. Not all languages are labeled due to overlap. Colour intensity corresponds to the dominance of the language.

Naturally, Wikipedia is not used to the same extent across countries due to differences in, e.g., population size, digital access, information needs, and government restrictiveness on information access. This is evident from the fact that 22.5% of total page views across all years stems from the United States, with 62.15 % of all page views being sourced from 10—predominantly highly developed—countries as seen in Table 1. This difference is then partially reflected in which language editions receive attention, with 88.61 % of attention directed towards the 10 language editions with the most page views, as seen in Table 2, largely corresponding to the national language of the countries in Table 1. Although the dataset comprises 340 languages in total, almost 50 % of page views are concentrated into a single language, English, and looking at a family level, of the 32 language families present, 87.74 % of attention is directed towards the 137 Indo-European languages in the dataset, of which English is one. Indeed, as can be seen in Figure 1b), most languages are found across Eurasia, with a particularly high concentration of languages in Europe. However, quite a few languages from Central Asia, especially the Caucasus, are present. Also in Africa, notably the Atlantic-Congo (42 languages) and Afro-Asiatic (16 languages) language families are present. Languages from the Americas and Oceania, however, have a noticeably low presence in the data.

Language	Page Views (%)	Cumulative
English	49.77	49.77
Japanese	6.92	56.69
Spanish	6.55	63.24
German	5.96	69.21
Russian	5.07	74.28
French	4.66	78.94
Italian	3.38	82.32
Chinese	2.67	84.99
Portuguese	2.02	87.00
Polish	1.60	88.61

Table 2: Top 10 visited language editions of Wikipedia over 2015 and 2024.

On the other hand, looking at the country level and the dominant language of each country as seen in Figure 1c), only 24 languages dominate the 239 countries and territories in the data. While many instances of the national language being the most dominant language are found in Europe, such as for Bulgaria, Sweden and Finland, a pattern of imperialistic legacy can be observed. For example, West Africa is dominated by French, while East Africa is dominated by English. Portuguese is dominant in the former colonies of Angola, Mozambique and Brazil, while the Russian Wikipedia is most popular in former Soviet states such as Ukraine and Kazakhstan. However, Russian is clearly more dominant in Russia (88% of page views) compared to Ukraine (57%) and Kazakhstan (66%), signaling that there is competition between Russian on the one hand, and the local national languages of Ukrainian (30%) and Kazakh (25%) and to a certain extent English (9.8% & 6.5%) on the other hand. At the same time, in other former Soviet states, such as Armenia (40% Armenian, 36 % Russian, 20% English) and Azerbaijan (42% Azerbaijani, 25% Russian, and 18% English), the local national languages are dominant, with the Russian Wikipedia being the second most used language. In the Baltic states, English dominates instead, followed by the local national language and Russian (Estonia: English 43%, Estonian 35%, Russian 16%; Latvia: English 37%, Russian 31%, Latvian 24%; Lithuania: English 43%, Lithuanian 41%, Russian 11%). Reasonably, these patterns occur as a result of different pressures acting on internet users, such as knowledge and familiarity of the languages in the population, the extent and usefulness of the Wikipedias (Kornai, 2013), and the political status and association with the languages, to name a few; all factors that suggest themselves to be accounted for in follow-up research. We are thus convinced that our dataset has great potential as a resource to model language competition and further our understanding of how technology interacts with—and shapes—both human society and cultural heritage.

4 Analyzing change in linguistic diversity

In this final part of the paper, we showcase how our dataset can be employed to answer research questions pertaining to linguistic diversity. Specifically, we aim to understand if digital retrieval of information has become more or less linguistically diverse over the last decade. Since the country coverage is substantially lower for 2024, we restrict the analysis to the years from 2015 to 2023, giving a time series of nine consecutive years. To increase the reliability of our estimates, we filter away countries that in any given year had fewer than 1,000,000

page views in total, which we consider a reasonable threshold. We then subset to countries and territories present across all nine years, yielding a complete time series with 199 countries. As a strategy to protect against sudden spikes in language richness due to scraping activities, we fixate the languages in each country to those present in 2015. Therefore, any changes in diversity will reflect a redistribution of attention among already established languages, rather than the introduction of new languages in the respective countries. While this strategy runs the risk of missing changes related to large-scale migration that explicitly took place after 2015, the actual impact on the measured diversity is likely limited within the relatively short timespan under study.

For each year, we calculate a relative abundance vector for each country by dividing the number of page views of each language by the total number of page views in that country. As such, we end up with 232×9 vectors, $p = [p_1, p_2, ..., p_n]$ where p_n is the proportion of page view visits to the n:th Wikipedia, in a given country and year, such that

$$\sum_{i=1}^{R} p_i = 1, \tag{1}$$

where R is the length of the relative abundance vector.

We then define linguistic diversity according to the Leinster Cobbold Framework (Leinster and Cobbold, 2012)—suggested by Essfors (2025)—as the effective number of completely dissimilar and equally abundant languages. While we acknowledge that diversity is multivariate in its nature, we only employ one measure of diversity to maintain the example scope of the analysis. Specifically, we use a naive model of diversity, treating all languages as completely dissimilar and distinct from one another. To reduce the influence of rare languages that might have been introduced through automatic scraping and not necessarily purposeful information retrieval, we specify the sensitivity parameter q to 2, such that rare languages are given a lower weighting. In this set-up, our working definition of diversity D_{tc} in the country c at time t formally becomes

$$D_{tc} = \frac{1}{\sum_{i=1}^{R} p_i^q}, q = 2. \tag{2}$$

First, we consider the absolute change in diversity from 2015 to 2023 by calculating the log diversity ratio between the years 2015 and 2023 as an effect size. Formally, we define the log diversity ratio ldr for a given country c as

$$ldr(c) = log_2(\frac{D_{2023c}}{D_{2015c}}) \tag{3}$$

where D_{2023c} and D_{2015c} are the measured diversity of c in 2023 and 2015 respectively. It follows that a positive log-ratio indicates an increase in diversity, and a negative log-ratio a decrease. Furthermore, the measure is symmetric around 0 (i.e., no change), such that a doubling of diversity = 1, while a halving of diversity = -1.

By plotting the ldr of each country as seen in Figure 2, it is evident that more countries have seen an increase in diversity than a decrease. As a matter of fact, we note an increase in 137 countries, more than twice the number of countries observing a decrease in diversity (62), yielding a mean ldr of 0.095, 95% CI = [0.063, 0.13], suggesting that on average, linguistic diversity has increased by approximately 4.5%–9.4% from 2015 to 2023 due to a redistribution of attention away from the most dominant languages within each country. One of the most important insights from this result is that while indeed diversity is suggested to generally be increasing, the fact that the change is heterogeneous—with diversity decreasing for a substantial portion of countries—indicates that the underlying mechanisms are country-specific, and potentially culturally dependent. Considering the spatial distribution of ldr as seen in Figure 4, there

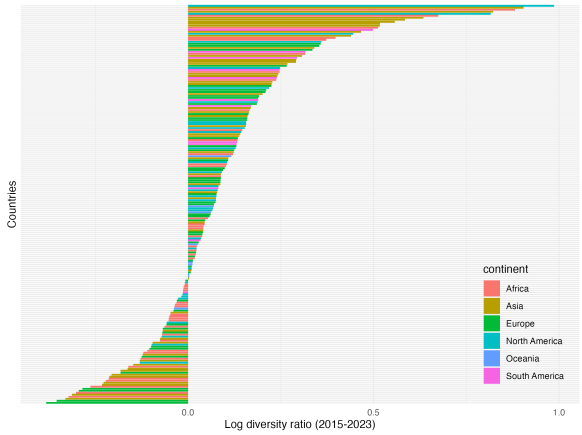

Figure 2: A barchart displaying the change in diversity across countries between 2015 and 2023. Each bar represents a country, with the length of the bar being proportional to the change in diversity. Bars directed to the left indicate a decrease in diversity, while bars directed to the right indicate an increase.

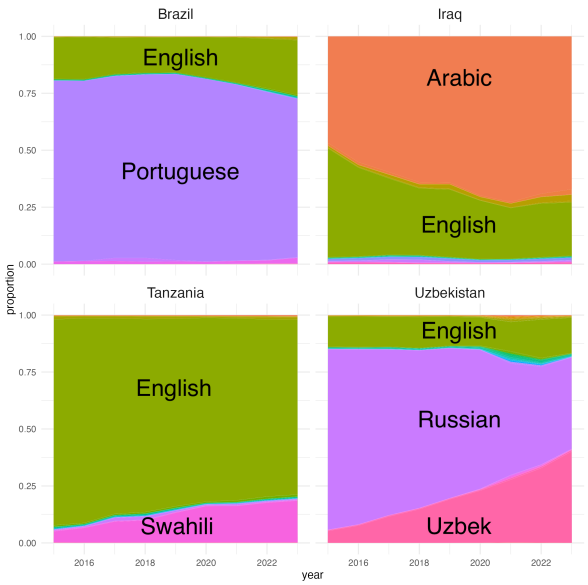

Figure 3: A streamgraph showing the proportional trajectories of dominant languages in Brazil, Iraq, Tanzania und Uzbekistan from 2015 to 2023.

is clearly some spatial dependence, with a substantial increase in diversity across most of Latin America, Central and South Asia, while the countries of the Middle East have seen a reduction. By investigating the relative abundance trajectories of the dominant languages within the respective countries, we can discern different patterns of interplay between the competing languages.

For example, across Latin America, exemplified by Brazil in Figure 3, we see a clear dominance of the national language—in the Brazil case, Portuguese—which has been declining in recent years, with increasing attention given to English, resulting in a diversification. This highlights an important aspect of diversity: what is perceived as diversification on the country level can be the result of a process yielding a concentration of languages on a global level. English is by far the most widely used language digitally, and as such an increase in its usage represents a decline in global diversity, while on a local level—as in Brazil—it is perceived as a diversification.

In Iraq and across the Middle East, on the other hand, the reverse pattern is observed: the English and standard Arabic wikis were on equal footing in 2015, after which increasing attention has been given to the Arabic one, making it dominant, thus resulting in a measured decline in diversity locally. This would, however, contribute to a global increase in diversity. In Tanzania, a similar pattern is observed, with English in steady decline and attention being directed towards the Wikipedia in the local language of Swahili. And since English was heavily dominating in 2015, this resulted in an increase in measured diversity. In Uzbekistan, a third pattern is observed, with English remaining fairly constant across time, while the dominant Russian language loses ground to the national language of Uzbek, which only constituted 5.1% of page views in 2015, gradually increasing its share to 40.0% in 2023, just barely behind Russian with 40.1%.

While we can use WikiLingDiv to describe these patterns in language attention, the data in itself cannot explain why they occur. For this reason, other dimensions of digital linguistic diversity are equally important to consider, e.g., how the digital support for Arabic, Swahili and Uzbek has changed relative to English and Russian. The addition of such data could yield insights into how extending the technological resources of a language impacts its usage. For this purpose, intricate measures of language support, such as that of the DLE (Gaspari et al., 2022; Grützner-Zahn and Rehm, 2022) for European languages, or Digital Language Support (DLS) of (Simons et al., 2022), building on (Kornai, 2013), for global estimates. To extend the analysis presented here, the most straightforward way would be to construct a metric quantifying the extent of the respective Wikipedia language edition, using, e.g., the number of Wikipedia articles and their sizes. However, equally important would be to also consider socioeconomic factors such as education level and second language acquisition, e.g., the increased use of the English Wikipedia in Latin America could very well be reflective of an increase in English proficiency. Further analyses could for example include the EF English Proficiency Index as a covariate to account for such heterogenity.

5 Discussion and Conclusion

In this paper, we have introduced WikiLingDiv, a dataset for quantifying linguistic diversity in the digital space using Wikipedia page views as a proxy for interest in retrieving information in different languages. We have showcased its extensiveness with respect to spatial and temporal coverage, as well as linguistic breadth. In doing so, we illustrated how previously observed linguistic biases in available digital content—to a degree—manifest as patterns of digital knowledge retrieval. However, information retrieval and digital language consumption is

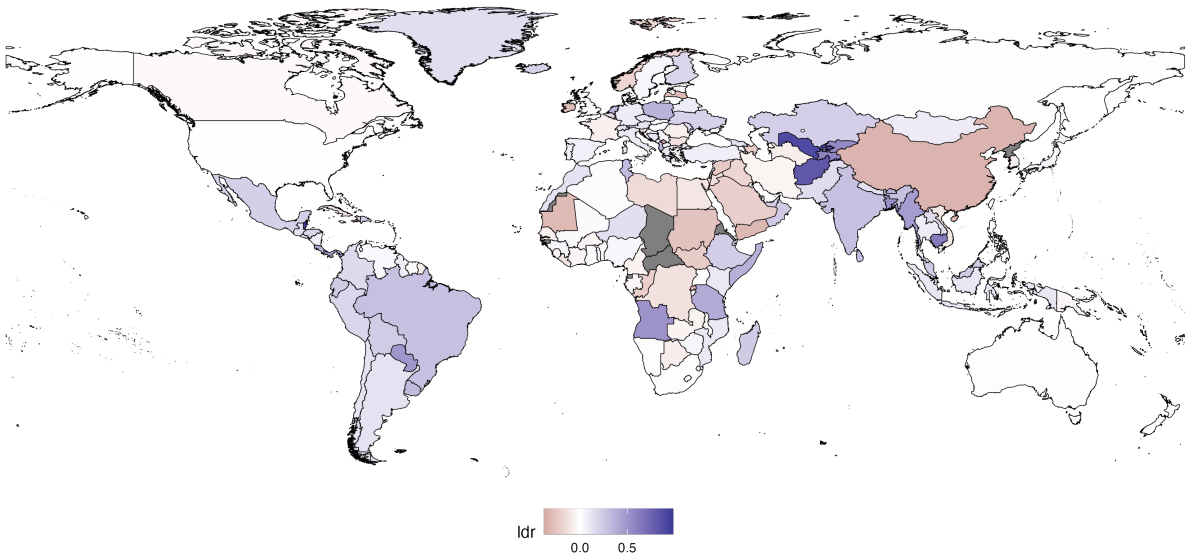

Figure 4: A coropleth map visualizing country-level change in the log diversity ratio (ldr) from 2015 to 2023. A blue color indicates an increase in diversity while a red color indicates a decrease, with more intense colors corresponding to a larger magnitude of change. An ldr of 1 is equivalent to a doubling of diversity. Grey countries indicate missing values.

only one dimension of digital linguistic diversity, and further research should aim at combining the dataset with other operationalizations, such as digital language support metrics.

While our approach is somewhat limited by the inability to differentiate between human and bot-generated page views, we have exemplified strategies to mitigate the impact of automatic scraping while still capturing signals of human linguistic preference. Furthermore, the dataset is potentially limited by the use of translation tools; it could very well be that traffic is directed towards the extensive English wikipedia, but then machine translated into a different language. This potentially yields a bias where the dominance of English is overestimated and the attention to languages with machine translation support is underestimated. In addition, while Wikipedia is one of the most widely used websites globally, its usage can be expected to vary within and across populations; e.g., men use Wikipedia to a larger extent than women (Johnson et al., 2021), there is a notorious rural-urban divide in internet access (International Telecommunication Union, 2024), and in some countries, access to Wikipedia has purposefully been restricted by the government (Zhang et al., 2017; Pan and Roberts, 2020; Yang and Roberts, 2021). Still, Wikipedia is arguably the closest approximation to a global and universilly used digital encyclopedia there is.

Through our example analysis, we have demonstrated how attention towards different language editions of Wikipedia varies overtime and across countries, resulting in cases of measurable increase as well as decrease in linguistic diversity. Most strikingly, however, we observe signs of an overall tendency towards *increasing* linguistic diversity, as per our operationalization, which to a certain extent challenges the widely-spread notion that linguistic diversity is steadily declining (Harmon and Loh, 2010; Bromham, 2023). However, since the increase is an average across countries, the observed diversification could result from English increasing its relative usage compared to already dominating language such as Spanish and Portuguese. Importantly, the longitudinal and spatial resolution of the data enables the identification of regional patterns of attention shifts from dominating global languages, such as English, to local languages—an interaction that ultimately plays a central role in shaping digital linguistic diversity. Explaining why these patterns emerge requires extending the analysis to inferential models, accounting for, e.g., the digital support of the languages, extent of internet access and digitalization across countries, and the linguistic diversity in the non-digital linguistic landscape, which could yield valuable insights into how diversity in digital language usage is shaped by the linguistic diversity of technology.

Ultimately, understanding the interaction between language, technology, and culture is cru-

cial as societies become increasingly digital, and NLP tools are further integrated into everyday life. While this endeavor is inherently complex, WikiLingDiv is a reusable and easy-to-integrate dataset that can support further research in areas of study where NLP, digital humanities, and linguistic diversity meet.

Acknowledgments

This research was funded by WWTF (grant number ICT23-012).

References

2023. Iso 639:2023 — code for individual languages and language groups. Accessed: 2025-12-18.

Thayer Alshaabi, David Rushing Dewhurst, Joshua R. Minot, Michael V. Arnold, Jane L. Adams, Christopher M. Danforth, and Peter Sheridan Dodds. 2021. The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. *EPJ Data Science*, 10(1):1–28.

Juliane Benson, Katharina Zeh, Hannes Essfors, Hannes Fellner, Julia Neidhardt, and Andreas Baumann. 2025. Linguistic diversity and digitalization: An ambivalent relationship. In *Digital Humanism: First Interdisciplinary Science and Research Conference, DIGHUM 2025, Vienna, Austria, November 20–21, 2025, Proceedings*, page 358–365, Berlin, Heidelberg. Springer-Verlag.

Elizabeth Blakey. 2024. The day data transparency died: How twitter/x cut off access for social research. *Contexts*, 23(2):30–35.

Lindell Bromham. 2023. Language endangerment: Using analytical methods from conservation biology to illuminate loss of linguistic diversity. *Cambridge Prisms: Extinction*, 1:e3.

Marco Civico. 2025. Measuring linguistic diversity: Limits and extensions of the greenberg index. *Journal of Quantitative Linguistics*, 0(0):1–28.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. Ethnologue: Languages of the world.

Hannes Essfors. 2025. Global linguistic diversity: Adapting the leinster–cobbold framework from ecology for humanities research. In *Anthology of Computers and the Humanities*, volume 3, Austin, Texas, U.S.A. Association for Computers and the Humanities.

Mikkel Flyverbom. 2019. *Digital and Datafied Spaces*, page 25–38. Cambridge University Press, Cambridge.

Federico Gaspari, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way. 2022. Introducing the digital language equality metric: Technological factors. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, page 1–12, Marseille, France. European Language Resources Association.

Victor Ginsburgh and Shlomo Weber. 2020. The economics of language. *Journal of Economic Literature*, 58(2):348–404.

Joseph H. Greenberg. 1956. The measurement of linguistic diversity. *Language*, 32(1):109–115.

François Grin and Guillaume Fürst. 2022. Measuring linguistic diversity: A multi-level metric. *Social Indicators Research*, 164(2):601–621.

Annika Grützner-Zahn and Georg Rehm. 2022. Introducing the digital language equality metric: Contextual factors. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, page 13–26, Marseille, France. European Language Resources Association.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. Glottolog 5.2. Available online at http://glottolog.org. Accessed: 2025-12-18.

David Harmon and Jonathan Loh. 2010. The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages. *Language Documentation & Conservation*, 4:97–151. Retrieved from ScholarSpace, University of Hawaii at Mānoa.

Tuomo Hiippala, Anna Hausmann, Henrikki Tenkanen, and Tuuli Toivonen. 2019. Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities*, 34(2):290–309.

International Telecommunication Union. 2024. Internet use in urban and rural areas. Facts and Figures 2024. Accessed: 9 February 2026.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Isaac Johnson, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia. 2021. Global gender differences in wikipedia readership. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:254–265.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+:

Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

András Kornai. 2013. Digital language death. *PLOS ONE*, 8(10).

Tom Leinster and Christina A. Cobbold. 2012. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489.

Amr Magdy, Thanaa M. Ghanem, Mashaal Musleh, and Mohamed Mokbel. 2014. Exploiting geo-tagged tweets to understand localized language diversity. In *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data*, GeoRich'14, New York, NY, USA. Association for Computing Machinery.

Elizabeth Marsh, Elvira Perez Vallejos, and Alexa Spence. 2022. The digital workplace and its dark side: An integrative review. *Computers in Human Behavior*, 128:107118.

Ulises A. Mejias and Nick Couldry. 2019. Datafication. *Internet Policy Review*, 8(4).

Jennifer Pan and Margaret E. Roberts. 2020. Censorship's effect on incidental exposure to information: Evidence from wikipedia. *SAGE Open*, 10(1).

Jeffrey Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Anahita Mashhadi, Jonas Lasser, Dominik Assenmacher, Shuhan Wu, Dejing Yang, Christoph Brantner, Daniel M. Romero, Jan Otterbacher, Christian Schwemmer, Kiran Joseph, David Garcia, and Fred Morstatter. 2023. Just another day on twitter: A complete 24 hours of twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1073–1081.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Statista. 2025a. Internet and social media users in the world 2025. https://www.statista.com/statistics/617136/digital-population-worldwide/. Accessed: 2025-12-17.

Statista. 2025b. Most visited websites worldwide. https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/. Statistic on global website visits, based on data from November 2024; accessed 2025-20-12.

Tuomas Väisänen, Olle Järv, Tuuli Toivonen, and Tuomo Hiippala. 2022. Mapping urban linguistic diversity with social media and population register data. *Computers, Environment and Urban Systems*, 97:101857.

Hadley Wickham. 2023. *httr: Tools for Working with URLs and HTTP*. R package version 1.4.7.

Wikimedia Foundation. 2024. Data platform/aqs/pageviews/pageviews per project. Accessed: 2025-12-18.

Wikimedia Foundation. 2025. Legal: Wikimedia foundation country and territory protection list. Last edited 19 May 2025; accessed 23 December 2025.

Wikimedia Foundation, Inc. 2025. List of wikipedias. Meta-Wiki page, accessed 20 Dec 2025.

Eddie Yang and Margaret E. Roberts. 2021. Censorship of online encyclopedias: Implications for nlp models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 537–548, New York, NY, USA. Association for Computing Machinery.

Ark Fangzhou Zhang, Danielle Livneh, Ceren Budak, Lionel Robert, and Daniel Romero. 2017. Shocking the crowd: The effect of censorship shocks on chinese wikipedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):367–376.