GlobLingDiv: A global dataset linking linguistic diversity and digital support to reveal landscapes with under-resourced languages for NLP

Katharina Zeh
Department of German Studies
University of Vienna
Vienna, Austria
katharina.zeh@univie.ac.at

Hannes Essfors
Faculty of Informatics
TU Wien
Vienna, Austria

Juliane Benson
Department of German Studies
University of Vienna
Vienna, Austria

Lale Tüver
Department of History
University of Vienna
Vienna, Austria

Andreas Baumann
Department of German Studies
University of Vienna
Vienna, Austria

Hannes Fellner
Department of
Comparative Literature
and Language Studies
University of Vienna
Vienna, Austria

Abstract

Linguistic diversity is increasingly under pressure globally and is becoming ever more relevant in digital contexts, where many languages remain structurally under-resourced, limiting access to language technologies and inhibiting equitable NLP development. To support linguistic diversity, publicly available data are needed that capture both the number of languages spoken and the distribution of speakers across them. We introduce GlobLingDiv, a database that uses country-level speaker distributions to derive language richness and entropy-based diversity measures, alongside a population-weighted digital language support measure. Applying these metrics globally, we examine the association between linguistic diversity and digital support conditions. The results reveal a substantial imbalance: highly diverse linguistic landscapes show comparatively low digital support, underscoring the need for more inclusive NLP environments.

1 Introduction

Linguistic diversity refers to the variety of spoken languages in a given community and the balance of speakers among those languages (Grin and Fürst, 2022). In recent years, linguistic diversity has decreased considerably, and languages are under threat of extinction. According to Ethnologue, a total of 7,164 languages are estimated worldwide, and by the end of the century, nearly half of them could disappear (Hutson et al., 2024; Kandler and Unger, 2023). Beyond cultural significance, linguistic diversity is increasingly relevant in digital contexts, where many languages remain structurally under-resourced and digitally marginal(Benson et al., 2025; Blasi et al., 2022). Recent initiatives have begun to systematically examine inequalities in digital language support—often focusing on specific regions or subsets of languages, such as the European Language Equality initiative (Gaspari et al., 2022; Grützner-Zahn and Rehm, 2022), yet comparable global-scale, speaker-based diversity estimates remain scarce.

To protect linguistic diversity worldwide and to work towards more inclusive digital and NLP environments, it is important not only to study the factors affecting it but also to provide reliable diversity estimates to begin with. However, a significant challenge is the lack of a comprehensive and publicly available dataset that quantifies linguistic diversity through both the number of languages and the distribution of speakers across them. An effective method of quantifying diversity is Shannon entropy (Shannon, 1948), which captures both richness and evenness based on speaker counts (see Section 3). Although entropy-based measures are increasingly used in linguistic diversity research (Grin and Fürst, 2022), they have been applied mainly to smaller regions or individual countries.

To address this gap, we have constructed a comprehensive database, GlobLingDiv, that quantifies linguistic diversity globally using Shannon entropy. The dataset draws on the Joshua Project's ethnolinguistic data (Joshua Project, 2025), as well as information from Ethnologue (Eberhard et al., 2022).[1]

[1] We acknowledge that Joshua Project and Ethnologue have missionary origins and are subject to ethical debate. Our use

Continent	langs	fams	spk %	spk M
Africa	2180	45	17.7	1345
Americas	1141	89	13.4	1019
Asia	2315	57	58.8	4469
Europe	414	25	9.6	727
Oceania	1348	77	0.5	405
Global	6745	214	100.0	7601

Table 1: Dataset summary aggregated by continent. *langs*: number of languages; *fams*: language families; *spk %*: global speakers share; *spk M*: speakers in Millions.

In addition, we provide a speaker-weighted digital support measure based on the Digital Language Support scale (Simons et al., 2022), enabling linguistic diversity to be considered alongside the degree of digital representation. As an application, we examine country-level correlations between diversity and digital support conditions.

Our contribution is twofold: (1) we introduce a publicly available dataset that quantifies linguistic diversity using richness and entropy-based measures and adds a abundance-weighted digital language support indicator at the country level; and (2) we demonstrate the value of combining these measures by examining global patterns and the association between linguistic diversity and digital support, offering a basis for identifying regions with under-resourced linguistic settings that we suggest future NLP research to focus on.

2 Constructing the dataset

Since our aim is to provide a fine-grained measure of country-level linguistic diversity on a global scale, we constructed a dataset containing language distributions based on speaker numbers per country. As our objective was to approximate overall linguistic landscapes rather than to capture individual acquisition patterns, we did not differentiate between L1 and L2 speakers. Data processing and integration were carried out in Python.[2]

First, the Joshua Project provides detailed demographic information on people groups within each country, including spoken language and population size. To address the fact that multiple people groups might speak the same language, the populations were in such cases aggregated to produce total speaker counts per language per country. Second, information from Ethnologue (Eberhard et al., 2022) was also integrated into the final dataset. In some cases, speaker counts were embedded in text fields, requiring the use of regular expressions to extract numeric values.

The data were aligned so that each row represents a unique language–country pair. When speaker counts were available from more than one source, the minimum value was selected to avoid overestimation of speaker counts. Entries were merged using ISO language and country codes (ISO, 2007, 2013), enabling integration with other datasets. Speaker counts were transformed to fractions (i.e., probabilities) by using total speaker counts per country as the basis for normalization. That is, for each country, all entries, representing the distribution of languages in the country's linguistic landscape, sum to 1.

The final dataset comprises 6,745 languages and 7,600,502,492 speakers in 239 countries, with a total of 12,249 language–country pairs (see Table 1), plus country-level totals. Approximately 95% of the linguistic richness estimated by Ethnologue is covered. GlobLingDiv is structured as follows: we provide five CSV tables containing: 1) a country-language-probability triplet. 2) country-level diversity metrics across three columns: richness, H, $exp(H)$; 3) a look-up table for each country with four columns: country ISO-code, country name, continent, abundance-weighted digital support score, total speaker count; 4) a look-up table for all languages included, consisting of two columns: language ISO-code and language name.

The limited coverage in our dataset can partly be attributed to the deliberate exclusion of extinct languages, which are included in Ethnologue. Additionally, inherent uncertainty associated with language data contributes to coverage issues: there is no universally accepted distinction between languages and dialects, such that closely related varieties may be treated as separate languages in some sources but as dialects in others, which often results in conflicting figures regarding both the number of languages and their speaker populations (Jolad and Agarwal, n.d.). Speaker estimates also rely heavily on national census data (Boissonneault et al., 2025), which have been criticized for political bias (Duchêne and Humbert, 2018), thereby reducing their ability to capture the true underlying language

of their data is limited to quantitative information and does not imply endorsement. We encourage future work to integrate additional linguistic data sources to broaden perspectives and reduce reliance on any single provider.

[2] Data and code are made available at https://gitlab.tuwien.ac.at/digilingdiv/globlingdiv

ISO	Lang.	H	exp(H)
PG	845	4.57	96.65
CM	289	3.98	53.59
CA	240	2.13	8.41
VU	115	4.00	54.34
DE	94	1.60	4.94
CH	45	1.75	5.74
AT	43	1.05	2.85
GL	3	0.62	1.86

Table 2: An excerpted GlobLingDiv table showing richness (Lang.), Shannon entropy (H), and Exponent Shannon ($exp(H)$) for selected countries. Cameroon appears diversier than Vanuatu according to richness, but Vanuatu is more diverse once relative abundance is considered. Likewise, Germany has nearly twice the richness of Switzerland, yet Switzerland is more diverse according to $exp(H)$ due to a more even speaker distribution.

and speaker distribution. However, addressing potential political biases in census-based speaker estimates at the country-level lies beyond the scope of the present study. Overall, the dataset spans 214 language families and 153 isolates, covering most of the global linguistic and phylogenetic diversity currently documented (Hammarström et al., 2024).

3 Entropy-based linguistic diversity

Linguistic diversity can be understood and measured in multiple ways. One of the most common approaches is language richness, i.e., the total number of languages within a given area. Other measures focus on linguistic abundance—how languages are distributed within a population. A third category, phylogenetic diversity, considers evolutionary relationships (Gavin et al., 2013). While phylogenetic approaches yield valuable insights, this study focuses on abundance, that is, the distribution of speakers across languages. To capture this, we compute Shannon entropy H of the probability distribution of speakers across all languages spoken in a country. Originating from information theory, the exponent of Shannon entropy $exp(H)$ is employed in biodiversity research to measure the 'effective number of species' and is here interpreted as the 'effective number of languages' (it is equivalent to the 'Hill number of order 1'; Tuomisto, 2010). It is a measure of complexity, and its values represent the level of uncertainty in the language distribution: higher (exp-)entropy indicates a more even (diverse) distribution of speakers across languages, while lower (exp-)entropy suggests a rather skewed distribution with a small set of dominant languages. In our context, $exp(H)$ can be interpreted as the number of languages in a given region that one would encounter if all of them were equally frequent. This approach offers a more nuanced perspective than simple language counts (Grin and Fürst, 2022).

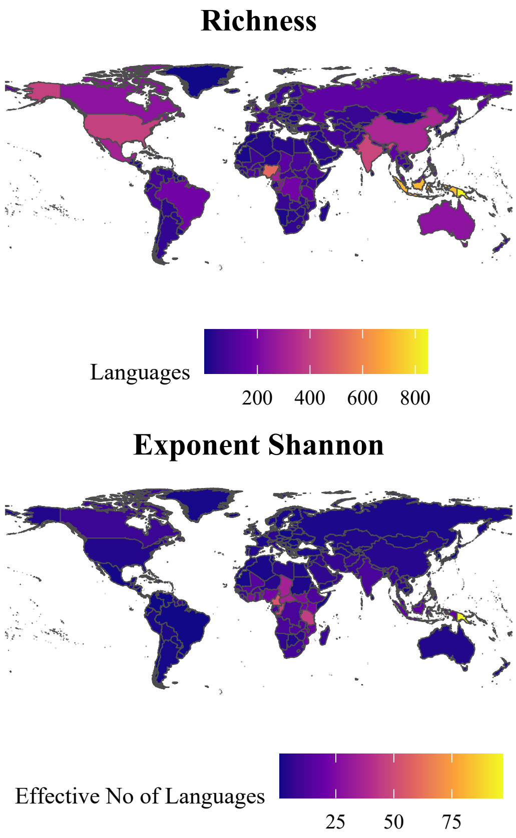

Figure 1: Global linguistic diversity visualized using both richness (language counts) and $exp(H)$, representing the effective number of equally abundant languages per country. Higher values reflect both greater balance and richness in a country's linguistic landscape.

While both measures in Figure 1 reveal substantial cross-national variation, $exp(H)$ highlights countries with more balanced speaker distributions as diversity hotspots. Cameroon and Vanuatu, ranked 8th and 20th by richness (289 and 115 languages), appear as the second and third most diverse countries once relative abundance is considered (53.6 and 54.3 effective languages), while Papua New Guinea is unequivocally perceived as the global hotspot of linguistic diversity (richness = 289, $exp(H)$ = 96.65). $exp(H)$ thus enables a relative abundance aware comparison of linguistic diversity across countries, which can—and should—be used in addition to richness, both of

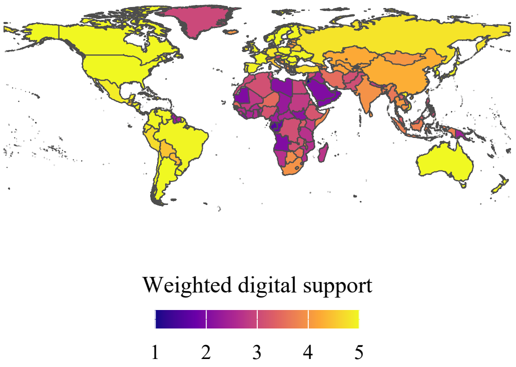

Figure 2: Country-level weighted digital language support. Higher values reflect digital representation aligned with national language distributions, while lower scores indicate limited support; values are comparatively low across much of Africa.

which can be derived using GlobLingDiv.

4 Country-level digital support

Several attempts have been made to quantify the digital vitality of languages (Gibson, 2015; Kornai, 2015). One of the most recent measures is the Digital Language Support (DLS) scale provided by Ethnologue (Simons et al., 2022). The scale captures the presence of languages across 216 digital tools and platforms and groups them into five ordered categories that reflect differing degrees of digital support (Still, Emerging, Ascending, Vital, Thriving). For our study, we recoded these categories to numeric values from 1 to 5. To obtain a country-level value, we combined these language-specific scores with the speaker information from our dataset and calculated a population-weighted average per country. In this way, languages with larger speaker populations contribute more strongly to the country score than those with fewer speakers. The resulting values provide a single, interpretable measure of how well the linguistic landscape of a country is supported in digital environments.

Countries such as Australia, the United States and the Netherlands show some of the highest digital support values (all 4.99), reflecting strong digital support for the languages spoken by the majority of the population. In contrast, several sub-Saharan African countries score toward the lower end of the scale (e.g., Republic of the Congo: 1.77). Vanuatu, previously highlighted as one of the most linguistically diverse countries, also displays a relatively low digital support value (1.66). These results (see Figure 2), illustrate that linguistic diversity and digital support do not necessarily align: countries with rich linguistic landscapes may have comparatively weak digital representation.

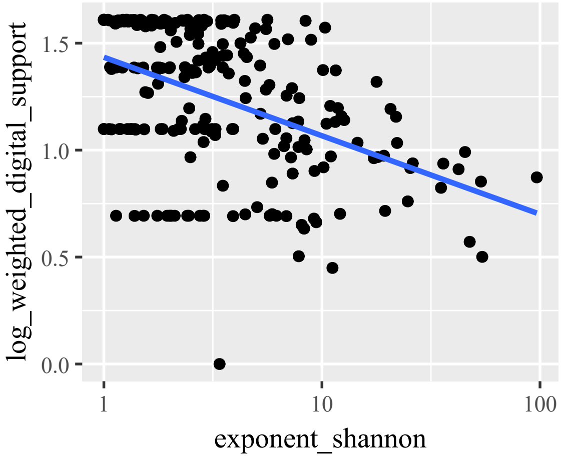

Figure 3: Scatter plot showing the association between $exp(H)$ (the x-axis has been scaled using log_{10}) and country-level digital language support (log_e-transformed). Higher linguistic diversity is generally associated with lower digital support.

5 Correlation analysis

To examine the relationship between linguistic diversity and digital language support, we conducted a country-level correlation analysis using $exp(H)$ and the country-level digital support values (log-transformed to account for skewness). The analysis was performed using Spearman's rank correlation coefficient due to the skewed distribution of $exp(H)$ (Kendall and Gibbons, 1990).

The result shows a statistically significant moderate to strong negative effect between the two measures ($r = -.47$, $p < .001$). This pattern indicates that countries with higher linguistic diversity tend to have lower levels of digital support, whereas countries with lower levels of linguistic diversity are more likely to exhibit higher digital support values.

While correlation does not imply causation, the strength and direction of the association highlight a substantial global imbalance: linguistic diversity and digital presence are unevenly distributed across countries. Indeed, a potential causal link between the two domains could go in both directions: highly diverse linguistic landscapes appear more vulnerable to digital underrepresentation, suggesting that linguistic diversity is not necessarily mirrored in the digital sphere (Mikami and Kodama, 2012); or it could be that lack of digital language resources further promotes biased language distributions in the non-digital domain.

6 Conclusion

To conclude, counteracting threats to linguistic diversity requires a clearer understanding of the factors that shape global linguistic landscapes, which in turn depends on extensive and reliable data (Bromham et al., 2022). The dataset presented here supports this effort by providing entropy-based diversity measures and a country-level indicator of digital language support. These measures reveal substantial cross-national variation in linguistic diversity and digital representation and show that the two do not necessarily align (Bella et al., 2023; Simons et al., 2022). Correlation patterns further suggest that linguistically diverse contexts may face disproportionate limitations in digital environments, pointing to structural imbalances.

At the same time, the results should be interpreted with care. Country-level aggregation inevitably obscures sub-national patterns, and the precision of speaker counts varies across sources and regions. Because the entropy-based diversity measures are derived from these speaker distributions, their interpretation is sensitive to such uncertainties. Future extensions could further draw on complementary data sources where available, such as national census data, to refine speaker estimates and language classifications. Future work could also complement the entropy-based results with qualitative validation, for example, by drawing on expert knowledge of regional and country-specific linguistic situations. In addition, future work could relate these patterns to those identified by existing initiatives and projects addressing inequalities in digital language support. Digital Language Support values offer a useful proxy, but do not capture all dimensions of digital presence or language use, particularly in informal or platform-specific settings (Gibson, 2015). The correlation analysis reflects association rather than causation. These limitations point to opportunities for future refinement, including sub-national modelling, multivariate approaches, and the integration of additional digital indicators and temporal data.

Beyond linguistic research, the dataset offers value for HSS and NLP. By providing structured and comparable measures, it can support quantitative analyses and help identify linguistic setups with limited digital support (Anastasopoulos et al., 2020; Blasi et al., 2022). The resource will be made publicly available in CSV format to support reuse and integration with other datasets.

Acknowledgments

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

References

Antonios Anastasopoulos, Christopher Cox, Hilaria Cruz, and Graham Neubig. 2020. Endangered languages meet modern nlp. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 39–45, Online. International Committee on Computational Linguistics.

Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. 2023. Towards bridging the digital language divide. *Preprint*, arXiv:2307.13405.

Juliane Benson, Katharina Zeh, Hannes Essfors, Hannes Fellner, Julia Neidhardt, and Andreas Baumann. 2025. Linguistic diversity and digitalization: An ambivalent relationship. In *Digital Humanism*, number 16319 in Lecture Notes in Computer Science, pages 358–365. Springer Nature.

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.

Michaël Boissonneault, Adam Tallman, Volker Gast, and Simon J Greenhill. 2025. Projected speaker numbers and dormancy risks of canada's indigenous languages. *Royal Society Open Science*, 12(2):241091.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature ecology & evolution*, 6(2):163–173.

Alexandre Duchêne and Philippe N Humbert. 2018. Surveying languages: the art of governing speakers with numbers. *International Journal of the Sociology of Language*, 2018(252):1–20.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World*, 25 edition. Dallas: SIL International, Dallas, TX.

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 (TDLE 2022), co-located with LREC 2022*, pages 1–12, Marseille, France.

Michael C Gavin, Carlos A Botero, Claire Bowern, Robert K Colwell, Michael Dunn, Robert R Dunn, Russell D Gray, Kathryn R Kirby, Joe McCarter, Adam Powell, and 1 others. 2013. Toward a mechanistic understanding of linguistic diversity. *BioScience*, 63(7):524–535.

Megan 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, Russia. Interregional Library Cooperation Centre.

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 (TDLE 2022), co-located with LREC 2022*, pages 13–26, Marseille, France.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1. https://glottolog.org. Max Planck Institute for Evolutionary Anthropology, Leipzig.

James Hutson, Pace Ellsworth, and Matt Ellsworth. 2024. Preserving linguistic diversity in the digital age: a scalable model for cultural heritage continuity. *Journal of Contemporary Language Research*, 3(1).

ISO. 2007. *ISO 639-3:2007 – Codes for the Representation of Names of Languages – Part 3: Alpha-3 Code for Comprehensive Coverage of Languages*. International Organization for Standardization, Geneva, Switzerland. [Accessed: 2025-04-08].

ISO. 2013. *Codes for the Representation of Names of Countries and Their Subdivisions – Part 1: Country Codes (ISO 3166-1)*. International Organization for Standardization, Geneva, Switzerland.

Sreenivasan Jolad and Anubha Agarwal. n.d. Mapping india's language and mother tongue diversity and its exclusion in the indian census. https://osf.io/sjxc6. OSF Preprints.

Joshua Project. 2025. Joshua project: People groups of the world. Accessed: 2025-07-15.

Anne Kandler and Roman Unger. 2023. Modeling language shift. In *Diffusive spreading in nature, technology and society*, pages 365–387. Springer.

Maurice G Kendall and Jean Dickinson Gibbons. 1990. *Rank correlation methods*, 5. ed. edition. A Charles Griffin title. Arnold, London [u.a.].

András 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, Russia. Interregional Library Cooperation Centre.

Yoshiki Mikami and Shigeaki Kodama. 2012. Measuring linguistic diversity on the web. In Laurent Vannini and Hervé Le Crosnier, editors, *Net.lang (version en anglais): Toward the Multilingual Cyberspace*, pages 118–139. C&F Éditions. Chapter in edited volume on multilingual cyberspace.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Gary F. Simons, Abbey L. Thomas, and Chad K. White. 2022. Assessing digital language support on a global scale. *Preprint*, arXiv:2209.13515.

Hanna Tuomisto. 2010. A diversity of beta diversities: Straightening up a concept gone awry. part 1. defining beta diversity as a function of alpha and gamma diversity. *Ecography*, 33(1):2–22.

A Appendix

This appendix summarizes the quantitative measures used to capture linguistic diversity and digital language support at the country level.

A.1 Shannon Entropy

Shannon entropy captures the diversity of languages spoken within a country, taking into account both richness and evenness.

$$H = -\sum_{i=1}^{N} p_i \log p_i \tag{1}$$

where p_i is the proportion of speakers of language i, N the total number of languages in the country, and where $\log$ refers to the natural logarithm (Shannon, 1948).

A.2 Exponentiated Shannon Entropy

To improve interpretability, Shannon entropy is exponentiated, yielding the Hill number of order $q = 1$, which corresponds to the effective number of equally abundant languages:

$$\exp(H). \tag{2}$$

This measure reflects the number of languages that would be present if all languages in a country were spoken by equal proportions of speakers.

A.3 Country-Level Weighted Digital Language Support

Country-level digital language support is computed as a speaker-weighted average of language-specific digital support scores:

$$D = \frac{\sum_i s_i n_i}{\sum_i n_i}, \tag{3}$$

where s_i denotes the digital support score associated with language i, and n_i the number of speakers of language i in the country. This measure represents the average level of digital language support experienced by speakers within a country.

For analyses sensitive to skewness, this measure is log-transformed prior to statistical modeling and visualization.