# Global Linguistic Diversity - Adapting the Leinster-Cobbold Framework from Ecology for Humanities Research

Hannes Essfors[1]

[1] TU Wien Informatics, TU Wien, Favoritenstraße 9-11, 1040 Vienna, Austria

**Abstract**

In this paper, we adapt the Leinster-Cobbold Framework from ecology and apply it to linguistic diversity. Thereby, we enable a unified framework of linguistic diversity that can account for richness, relative abundance and similarity between languages. By analyzing global linguistic diversity at the country and continent levels using the framework, we demonstrate how diversity can be interpreted as an effective number from three different perspectives. By doing this, we show that accounting for different aspects of diversity substantially influences how diversity is perceived, and conclude that a multivariate view is crucial for future endeavors to model linguistic diversity statistically.

**Keywords:** Linguistic Diversity, Leinster-Cobbold Framework, Levenshtein Distance, Lexical Similarity, Diversity Profiles

## 1 Introduction

Linguistic diversity is declining. This has been asserted at least twice in the last 15 years. First by Harmon and Loh [11], who described a 20% decline in linguistic diversity between 1990-2005, and more recently by Bromham et al. [2] who predict a loss of more than 20% of languages by the turn of the century. While both papers investigate linguistic diversity, they take two quite different approaches. [2] take the quite radical view of equating diversity to richness: the total number of languages found. This is not unproblematic for two reasons. First, this approach is prominently susceptible to the quite frankly unsolvable issue of what makes two linguistic varieties qualify as different languages; second, it effectively ignores relative abundance, i.e., how many individuals speak a particular language, as a legitimate factor in diversity. This, however, does not always reflect our intuitive understanding of diversity; imagine a city where 3 languages are spoken: A, B, and C. If the distribution is completely even, i.e., each language is spoken by 33% of the population, then it should be considered more diverse than a city where A is spoken by 90% of the population. Harmon & Loh take this into account by measuring the change in speaker concentration into dominant languages. As such, the city going from an even speaker distribution to being dominated by one language would constitute a decline in diversity. From a viewpoint equating diversity to richness, it would be argued that the diversity is unchanged.

We argue that there is a third legitimate addition to diversity, namely similarity. Consider again the city with 3 evenly distributed languages, A, B and C. Now imagine that A, B and C are highly similar, perhaps even mutually intelligible languages, such as Swedish, Danish and Norwegian. Intuitively, we would perceive this city as less diverse than if the languages spoken were highly dissimilar, such as English, Mandarin, and Swahili. We are not the first to make this argument.

Hannes Essfors. "Global Linguistic Diversity - Adapting the Leinster-Cobbold Framework from Ecology for Humanities Research." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 618–634. https://doi.org/10.63744/srhQaCwGo5mj.

Hormon & Loh state that phylogenetic and structural diversity is important, but do not incorporate it into their index. Grin & Fürst [10] too acknowledge that similarity is a legitimate ingredient in determining diversity, but perhaps the first expression of this idea goes back to the founder of linguistic Typology, Joseph Greenberg. In his 1956 work, *The Measure of Linguistic Diversity* [9], he introduces two indices, his A- and B-methods, that account for relative abundance and similarity. The A-method defines linguistic diversity as the probability of randomly selecting two speakers of different languages. Formally, this is defined as

$$A = 1 - \sum_{i}^{S} p_i^2 \tag{1}$$

where $S$ is the total number of languages, i.e., richness and $p_i$ the relative abundance of the $i^{th}$ language, i.e., the proportion of speakers of that language expressed as a fraction.[1] Assuming that the similarity between any two languages can be modeled as a resemblance factor $r \in [0, 1]$ where 1 indicates complete resemblance and 0 no resemblance, the B-method is formally defined as

$$B = 1 - \sum_{ij}^{S} p_i p_j r_{ij} \tag{2}$$

where $r_{ij}$ is the resemblance between the languages $i$ and $j$. This makes A a special case of B, where $r_{ij} = 1$ for all $i = j$ and $r_{ij} = 0$ for all $i \neq j$. In principle, the A-method model is a naive model assuming that all different languages are completely dissimilar. Both of these measures have independently been discovered and applied in ecology, where they are commonly known as the Simpson-index [24] and Rao's Quadratic entropy [22]. These, together with a multitude of other indices, have been part of a prolonged and heated debate in ecology about what diversity is and how it is best measured, in line with the example we gave in the earlier paragraph [16]. Some of this debate was, however, settled in 2012 when Leinster & Cobbold [15] introduced the *Leinster-Cobbold Framework* which offers a unified approach to diversity, incorporating richness, relative abundance, and similarity, focusing on how diversity is perceived from different viewpoints. Although the framework was developed for ecology, it is general and applicable to any domain, including linguistics. In this paper, we introduce the framework in a linguistic setting. First, the theory behind the framework is explained, and we show how Greenberg's A and B indices can be derived from it. Then, we showcase the framework empirically by measuring global linguistic diversity, thereby isolating the effect of accounting for richness, relative abundance, and similarity to determine its impact and relevance as an addition to the field of linguistic diversity.

### 1.1 The Leinster-Cobbold Framework

The Leinster-Cobbold Framework [15] is a family of diversity indices parameterized by a 'sensitivity' variable $q \in [0, \infty]$, adjusting the measure's sensitivity to dominant languages. As such, $q = 0$ assigns maximal importance to rare languages, while $q = \infty$ only takes dominant languages into account. Therefore, adjusting $q$ only changes how diversity is perceived, not the underlying diversity of the setting. The underlying diversity is given by the relative abundance of languages, summarized in a relative abundance vector $p$ ($p_1$, $p_2$,...,$p_n$) such that $\sum_i^n p_i = 1$ and an $n \times n$ similarity matrix $Z$ with all pairwise similarities. How these similarity measures are derived has to depend on what aspects of linguistic similarity are of interest, e.g., phylogenetic, lexical, syntactic etc., but formally, the only requirement is that they are bound between 0 and 1. Then, diversity is defined according to [15, eq.(1)] as

[1] Greenberg uses a slightly different notation in his original publication, but we have streamlined it for clarity.

$$ {}^{q}D^{Z}(p) = (\sum_{i}^{S} p_i (Zp)_i^{q-1})^{\frac{1}{1-q}}, q \geq 0, q \neq 1 \tag{3} $$

where

$$ (Zp)_i = \sum_{j}^{S} Z_{ij} p_j. $$

$S$ is the total number of languages, with relative abundances described by $p$. $(Zp)_i$ therefore denotes the average similarity between the $i^{th}$ language and all other languages, weighted by their relative abundance.

We exemplify this as follows: consider a country where three languages are spoken: A by 40%, B by 50%, and C by 10%. Let A be very similar to B with 75% similarity, while C is dissimilar to both with 25% similarity. The diversity of the linguistic setting is now defined by the relative abundance vector $p = [0.4, 0.5, 0.1]$ and the similarity matrix

$$ \begin{aligned} Z = [&[1, 0.75, 0.25] \\ &[0.75, 1, 0.25] \\ &[0.25, 0.25, 1]]. \end{aligned} $$

$Zp$ is then calculated as $Z \cdot p\top$ such that

$$ \begin{aligned} Zp = [&1 * 0.4 + 0.75 * 0.5 + 0.25 * 0.1, \\ &0.75 * 0.4 + 1 * 0.5 + 0.25 * 0.1, \\ &0.25 * 0.4 + 0.25 * 0.5 + 1 * 0.1] \\ &= [0.8, 0.825, 0.325]. \end{aligned} $$

One or more values for $q$ is decided upon given the desired sensitivity to dominant languages, say $q = 0$, i.e., minimal importance. Then, the diversity of the setting is calculated as

$$ {}^{0}D^{Z}(p) = \sum_{i}^{3} \frac{p_i}{(Zp)_i} = \frac{0.4}{0.8} + \frac{0.5}{0.825} + \frac{0.1}{0.325} = 1.4. $$

Thus, given the relative abundance and similarity of languages, the number of languages is *effectively* 1.4, since the value generated is equivalent to the value of a hypothetical setting with 1.4 completely dissimilar and equally abundant languages. This is referred to as the effective number of languages. The interpretation of diversity as effectively that of a hypothetical setting is crucial to understanding the measure. For a given $q$, the diversity is reasonably maximized when all languages have nothing in common, i.e., their similarity is 0, and their relative abundance is even, i.e., $p_i = \frac{1}{S}$. If we therefore let $Z$ be the identity matrix $I$, then $Zp = Ip = p$. Thus

$$ {}^{q}D^{I}(p) = \sum_{i}^{S} p_i (p_i)^{q-1})^{\frac{1}{1-q}} = (\sum_{i}^{S} p_i^q)^{\frac{1}{1-q}} = (S \cdot \frac{1}{S^q})^{\frac{1}{1-q}} = (S^{1-q})^{\frac{1}{1-q}} = S. $$

As such, assuming that all languages are completely dissimilar and equally abundant maximizes the measure, yielding richness. Therefore, a setting with three different languages and a similarity and relative abundance structure that yields a diversity of 1.4 is equivalent to a theoretical setting

with 1.4 completely dissimilar and equally abundant languages. By setting $Z = I$, the similarity between languages is disregarded, giving what we refer to as *naive* diversity. If we set $q = 0$ in the naive case, then

$$ {}^0D^I(p) = (\sum_i^S p^0)^{\frac{1}{1-0}} = S $$

independent of the relative abundance vector $p$.

There are two limits we have not introduced here, namely $q = 1$ and $q = \infty$. They are defined in the original paper by Leinster and Cobbold in [15]. However, for the sake of brevity, we omit a thorough explanation, as it is not essential to our argument. As $q$ tends towards $\infty$, however, only the most dominant languages are considered. Therefore, by plotting diversity as a function of $q$ in both the naive and non-naive case, information about richness, relative abundance, and similarity is conveyed, which is referred to as the diversity profile of a setting by Leinster and Cobbold. By selecting specific values for $q$, established measures can be derived, and for $q = 2$

$$ {}^2D^z = (\sum_{ij}^S p_i(Z_{ij}p_j)^1)^{-1} = \frac{1}{\sum_{ij}^S p_i Z_{ij} p_j} = \frac{1}{1-B}. \tag{4} $$

Thus, Greenberg's B-method is a transformation of diversity for $q = 2$. Similarly, in the non-naive case

$$ {}^2D^I = (\sum_i^S (p_i^2)^1)^{-1} = \frac{1}{\sum_i^S p_i^2} = \frac{1}{1-A}. \tag{5} $$

As such, analysis according to the Leinster-Cobbold framework incorporates and extends established quantitative views on diversity in the linguistic tradition. In the rest of the paper, we shall exemplify such an analysis and evaluate its impact on how diversity is perceived.

## 2 Materials and Methods

### 2.1 Data

In this work, we take a global approach to linguistic diversity, similar to Bromham et al. [2]. To account for richness and relative abundance, we use language and speaker data per country from the 22nd edition of Ethnologue, enriched with speaker data from the Joshua Project [5; 13]. As such, the definition of a language in our data is any variety that has an ISO639-3 code [23]. To avoid double-counting languages and speakers, we remove any macrolanguage identifiers, such as Serbo-Croatian (hbs). This gives us a dataset with 239 countries and territories, 6,745 languages, and 7,600,502,492 speakers. This dataset is to be released in a further publication.

To account for similarity, we follow the suggestion of Greenberg and turn to core vocabulary in established word lists [9, p. 110]. Multiple cross-linguistic word-list collections have been published in the past decades [18], the most extensive of which is the *Automated Similarity Judgment Program (ASJP)* [27]. The ASJP is a collection of 100- and 40-item wordlists on basic vocabulary transcribed into phonetic ASJP-code, and is most prominently used to infer phylogenetic relationships according to the lexicostatistical method [3; 12]. The database is extensive, containing 10,168 wordlists on 5,590 distinct ISO639-3 languages, making it suitable for our purposes [27]. Since the 40-item lists are the most abundant, complete for 86% of languages in the database [18], we exclusively use these. To strike a balance between coverage and data completeness, we use the same data completeness requirement as the software associated with the ASJP, excluding any wordlists with fewer than 28 words [27].

Intersecting the speaker and lexical data yields a dataset covering 4565 languages, 237 countries, and 7,210,257,212 speakers. To enable geospatial analysis, we enhance the dataset by assigning each country to a continent using the R package *rnaturalearth* [19], which we use to generate all maps in the paper.

| Continent | Countries | Languages | Speakers |
|---|---|---|---|
| Africa | 58 | 1562 | 1,245,426,292 |
| Asia | 53 | 1316 | 4,213,326,095 |
| Europe | 49 | 302 | 704,300,887 |
| North America | 38 | 582 | 585,908,491 |
| Oceania | 25 | 1100 | 39,702,975 |
| South America | 14 | 349 | 421,592,472 |
| Sum Distinct | 237 | 4565 | 7,210,257,212 |

**Table 1:** A table displaying aggregates of data available for diversity calculations after intersecting speaker and lexical data from the ASJP. Note that the sum of distinct languages does not equal the sum of languages on all continents due to languages overlapping.

### 2.2 Calculating Diversity

We calculate the diversity in each country based on the Leinster-Cobbold framework according to Equation (3). We adapt the nomenclature of Leinster and Cobbold and refer to measures of diversity where $Z = I$ as *naive diversity*. In cases where $Z \neq I$, we write *non-naive diversity* or *diversity*. Since we want to investigate the impact of accounting for relative abundance and similarity on how diversity is perceived in a global linguistic context, we calculate three measures for each country and continent: naive diversity with $q = 0$, yielding richness; naive diversity with $q = 2$, yielding Greenberg's A-method as an effective number; and diversity with $q = 2$, yielding Greenberg's B-method as an effective number. Thus, we stay as close to the linguistic tradition as possible, with the difference between each measure being first relative abundance, and then similarity, thereby isolating their effects.

For each country and continent, we generate a similarity matrix with pairwise similarities between each language. The similarities are derived using normalized Levenshtein distance (LDN) [1; 20]. Let the concept $c$ in two languages A and B be represented by the strings $a_c$ and $b_c$ in ASJP-code. We define the distance between the strings as the fewest number of substitutions, deletions, and insertions that turn $a_c$ into $b_c$, i.e., the Levenshtein distance [17]. We denote this distance as $d_l(a_c, b_c)$. The distance is then normalized through division by the longest string, i.e., the maximal Levenshtein distance possible. Formally, the distance between two languages w.r.t. one concept is thus

$$LDN(a_c, b_c) = \frac{d_l(a_c, b_c)}{max(l(a_c), l(b_c))} \tag{6}$$

where *l($a_c$)* and *l($b_c$)* are the lengths of the strings. For a word list with A and B defined for the concepts $C$, the similarity between A and B, $s(A, B)$ is then given as the average distance across the word list subtracted from 1, formally:

$$s(A, B) = 1 - \frac{1}{|C|} \sum_{c \in C} LDN(a_c, b_c). \tag{7}$$

If two languages have similar core vocabulary, the similarity will be closer to 1. If it is very different, it will be closer to zero.

Although we here derive the similarity measures using word lists as suggested by Greenberg, measures based on other aspects of language—e.g., syntax or phonology—could theoretically be used as well. If such measures are independent of each other, the choice could significantly impact the results and would add a further view on diversity: which aspect of linguistic similarity to consider. In such a case, the scarcity of cross-linguistic data might present a substantial bottleneck: Grambank, which would be a potential source of data on morphosyntax [25], only covers 2467 languages. However, we restrain our scope to compare naive, non-similarity-aware diversity with similarity-aware diversity based on core vocabulary.

### 2.3 Statistical analysis

As described in the data section, we derive our data by intersecting Ethnologue and ASJP, failing to account for approximately 400,000,000 speakers. The disproportionate spread of these missing speakers results in some countries lacking substantial coverage. As seen in Figure 1, about 11% of countries have a speaker coverage lower than 75%. Since our employed diversity measures depend on the relative abundance of languages, a dominating language missing might heavily distort the measured diversity. To circumvent this, we use a threshold for speaker coverage at 75% for country-level analysis, ensuring reliability while still including 89% of countries.

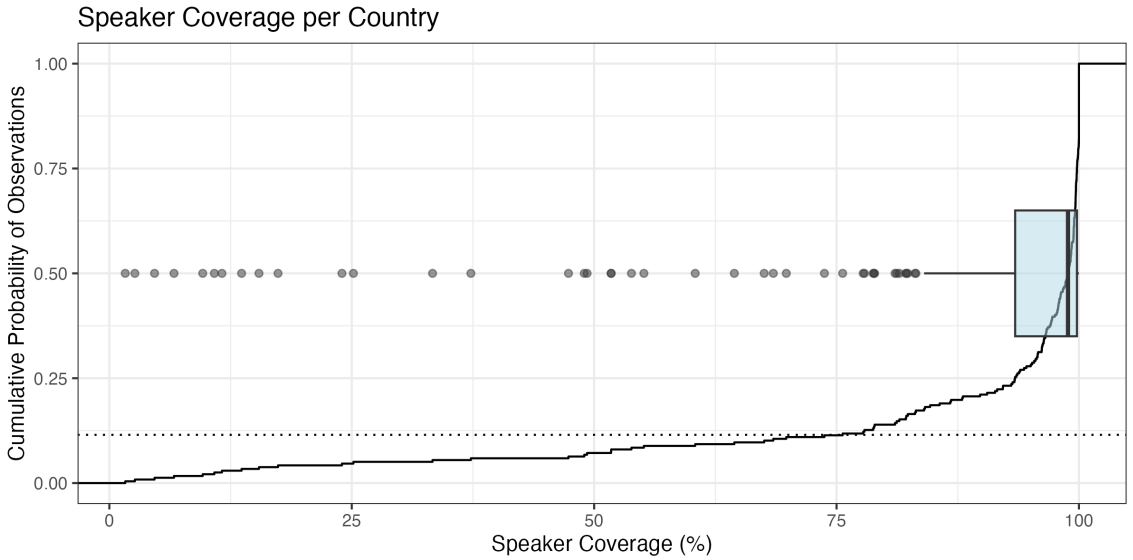


**Figure 1:** An ECDF-plot showing cumulative probability of randomly selecting a country with a certain speaker coverage in the ASJP indicated by the solid black line. The dashed line indicates the applied cutoff at 75% coverage used in the study, excluding 11% of data points. The ECDF-plot is overlaid with a boxplot in blue showing the interquartile ranges.

We carry out all analyses using the statistical programming language R [21].[2] First, we conduct spatial analysis by generating choropleth maps using a continuous viridis palette [7] to indicate the relative magnitude of diversity to spot patterns in its distribution. Then, we investigate how the measure of diversity is impacted by accounting for relative abundance and similarity. For this purpose, each country is ranked according to its naive diversity with $q = 0$, naive diversity with $q = 2$, and diversity with $q = 2$ such that a low rank indicates low diversity. By correlating these ranks using Spearman's $\rho$ [26], we get a measure of association where $\rho = 1$ indicates complete agreement on diversity rankings and $\rho = -1$ complete disagreement, i.e., a large effect

[2] All code written for the paper is made available at https://github.com/Eszettfors/Global-Linguistic-Diversity—Adapting-the-Leinster-Cobbold-Framework-from-Ecology

of increasing $q$ from 0 to 2. 95% confidence intervals around the estimates are then constructed using percentile bootstrapping with 1000 resamples to assess the uncertainty around the estimate. To make interpretation easier, we rescale the estimates as $\rho_t = \frac{1-\rho}{2}$ such that 0 equals no effect and 1 equals maximum effect. We then plot country ranks against each other in a scatter plot, studying deviation from the identity line to identify outliers of interest. Finally, we demonstrate the use of the Leinster-Cobbold framework to its fullest extent by calculating the diversity profiles of each continent with varying values for $q$. For that purpose, we treat each continent as a setting, aggregate the speaker numbers per language, and derive a similarity matrix for each continent as explained in Section 2.2.

## 3 Results

### 3.1 Spatial Analysis

In Figure 2, we find the geographical distribution of diversity given by a) language counts, b) the effective number of equally abundant languages, and c) the effective number of equally abundant and completely dissimilar languages. Considering a), linguistic diversity appears relatively evenly distributed across the globe. A high concentration of languages is found in North America, with the United States (359 languages), Canada (217 languages), and Mexico (182 languages) housing the most languages. In Table 2, the 10 most diverse countries and the number of languages according to the measures can be found. In South America, Brazil appears notably more diverse (157 languages) compared to other South American countries. In Africa, the most diverse countries are found in sub-Saharan Africa, with Nigeria (346 languages), Cameroon (218 languages), and Congo (189 languages). On the same latitude, another hotspot of linguistic diversity is found in Oceania and Asia, with Papua New Guinea housing the most languages of any country (689 languages), followed by Indonesia (490 languages). Other countries being highlighted as diverse through this measure are Australia (194 languages), India (194 languages), China (180 languages), and Russia (128 languages). Unsurprisingly, the countries perceived as the most diverse under this measure tend to span a large area with a large population.

| | Naive Diversity, q = 0 | Naive Diversity, q = 2 | Diversity, q = 2 |
|---|---|---|---|
| 1 | Papua New Guinea (689) | Vanuatu (37.23) | Cameroon (6.09) |
| 2 | Indonesia (490) | Solomon Islands (29.01) | Papua New Guinea (4.86) |
| 3 | United States (359) | Congo (21.93) | Chad (4.82) |
| 4 | Nigeria (346) | Central African R. (19.75) | Côte D'Ivoire (4.69) |
| 5 | Cameroon (218) | Cameroon (18.39) | Central African R. (4.46) |
| 6 | Canada (217) | Tanzania (18.36) | Nigeria (4.22) |
| 7 | Australia (194) | South Sudan (13.53) | Kenya (4.05) |
| 8 | India (194) | Chad (13.24) | South Sudan (3.98) |
| 9 | Congo (189) | Uganda (12.28) | Liberia (3.85) |
| 10 | Mexico (182) | Papua New Guinea (12.03) | Ghana (3.82) |

**Table 2:** A table showing the 10 most diverse countries and their effective numbers according to three different views on diversity, accounting for richness, relative abundance, and similarity. Note how the small oceanic island nations of Vanuatu and Solomon Islands are by far analyzed as the most diverse when accounting for relative abundance, but not when accounting for similarity. In both cases, predominantly African countries are perceived as diverse compared to the case of only considering richness.

However, setting $q = 2$ to make the measure aware of speaker distributions and penalizing the presence of dominant languages as seen in b), reduces the correlation between country size and

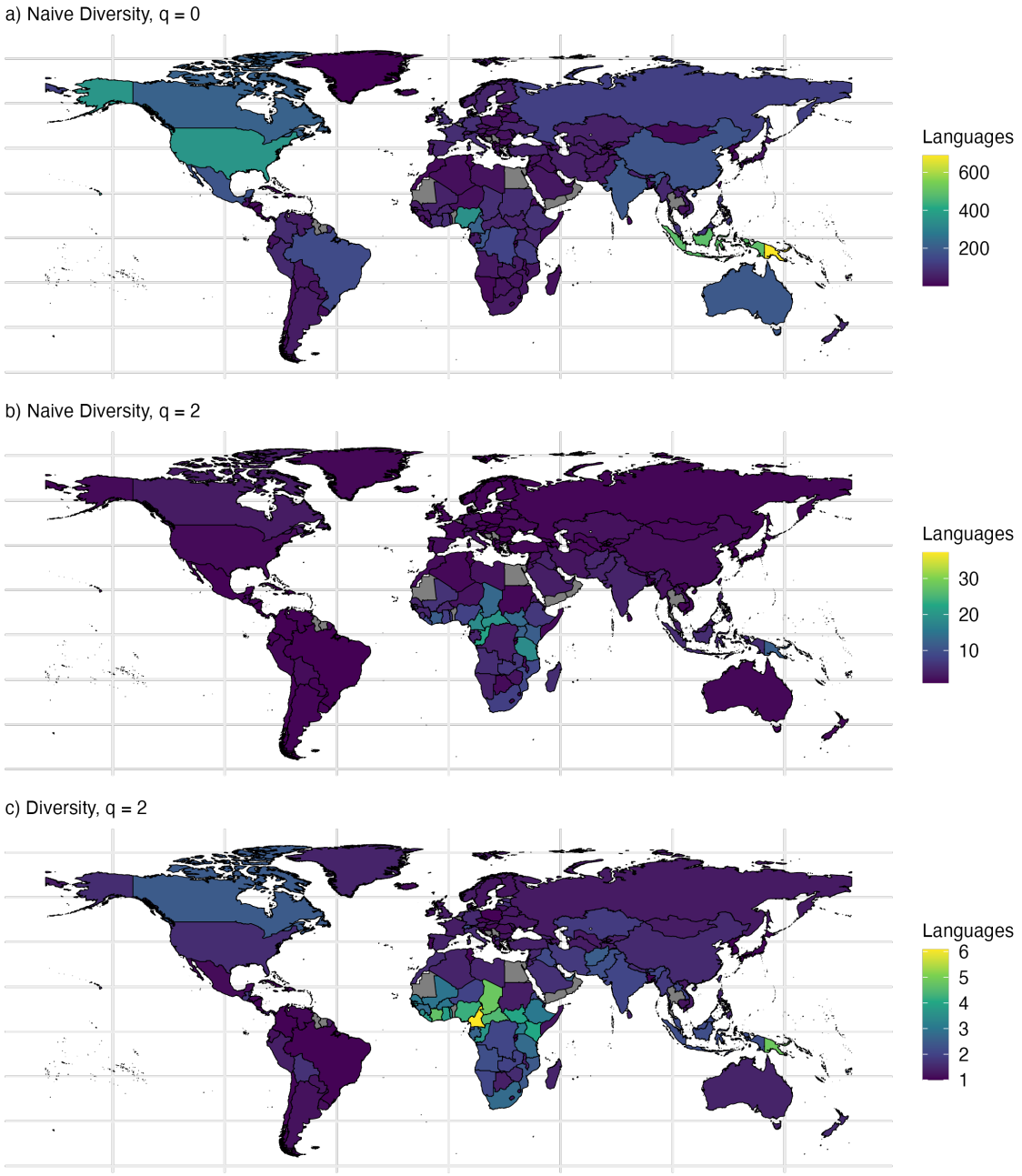


**Figure 2:** Three maps showing the spatial distribution of diversity according to three points of view accounting for a) richness, b) relative abundance, and c) relative abundance together with similarity. The color indicates the relative magnitude of diversity, measured as the effective number of languages, with yellow indicating high diversity. Greyed countries are excluded due to insufficient speaker coverage in the data.

diversity. Naturally, the range of the effective number of languages is across the board lower, since the measure is no longer indifferent to the relative abundance of languages, and in many parts of the world, diversity discrepancies are now gone. E.g., Brazil (1.02 effective languages) is not perceived as particularly more diverse than the rest of South America, due to the relative dominance of Portuguese. The measured 1.02 implies that effectively only one language is present: Portuguese. Similarly, the United States (1.74 effective languages) and Mexico (1.12 effective languages) are

perceived as much less diverse, due to the dominance of English and Spanish. On the other hand, Canada is still perceived as somewhat more diverse, effectively housing 3.2 languages, due to the strong presence of, e.g., French next to English. A clear shift towards Sub-Saharan Africa being perceived as more diverse can also be noticed: Nigeria is perceived as somewhat less diverse (7.96 effective languages) compared to other countries in the region, such as Congo (21.92 effective languages), Cameroon (18.39 effective languages) or Tanzania (18.35 effective languages). In Oceania and South East Asia, the countries with the most prominent linguistic richness, Indonesia and Papua New Guinea, are under scrutiny of speaker number distributions perceived as substantially less diverse. Indonesia effectively only houses 5.19 languages, while the diversity in Papua New Guinea is substantially higher (12.03 effective languages). As such, Papua New Guinea is still perceived as diverse relative to most of the world, but quite a bit less diverse than the countries in Sub-Saharan Africa. Oceania is, however, still a hotspot of linguistic diversity according to this measure, as the most remarkable shift in perceived diversity is noted for the small island nations of Vanuatu and the Solomon Islands, which, according to the measure, are perceived as the most diverse countries of the world. In Vanuatu, 113 languages are found in the data, but the distribution of speakers is equivalent to 37.23 effective languages, making Vanuatu by far the most diverse country in the data, almost twice as diverse as, for example, the Republic of the Congo. The Solomon Islands are a step down from Vanuatu, with 29.01 effective number of languages, which still makes it appear substantially more diverse than the rest of the world.

Considering the similarity weighted measure in c), changes in the global distribution of diversity are not as radical as when accounting for relative abundance. Generally, the effective number of languages in each country goes down as expected, but relative to other countries, Sub-Saharan Africa is even more pronounced as the epicenter of global linguistic diversity. Most prominently, Cameroon stands out as significantly diverse, with effectively 6.09 equally distributed and completely dissimilar languages; this is an entire language more than the second most diverse country, Papua New Guinea (4.89 effective languages). As such, we notice that when accounting for both similarity and relative abundance, Papua New Guinea again appears very diverse compared to most other countries. The perceived diversity of Vanuatu and the Solomon Islands, on the other hand, falls substantially when considering similarity, to effectively only 3.78 and 3.34 languages. This is unsurprising given that the languages of Vanuatu form a network of related, and to varying degrees mutually intelligible languages [6; 8], resulting in substantial similarity between the languages. As a reference, the mean similarity between languages in Vanuatu in the data is 0.26, which can be compared to the median mean similarity in Oceania of 0.17 and the global median mean similarity of 0.15, as can be seen in Figure 3. Generally, Asia, Europe, and North America have similar distributions of mean similarities, with most values concentrated around the median at 0.15 and a standard deviation of 0.05. The distributions of Africa and Oceania deviate somewhat with fat tails and a somewhat higher median of 0.17. The variability in Africa is comparable to other continents, with a standard deviation of 0.05, while it is substantially larger in Oceania (0.11). The countries of South America, on the other hand, seem to be composed of more dissimilar languages, yielding a median mean similarity of 0.13 with lower variability (SD = 0.03).

### 3.2 Rank correlation

Moving away from the exact interpretation of the measures as effective number of languages, and instead considering the country ranks according to the diversity measures, the impact of accounting for relative abundance and similarity can be observed in Figure 4. The point estimate of the global effect of accounting for relative abundance in a) is moderate at $\rho_t = 0.23$, suggesting disagreement between the measures. By considering the impact of setting $q = 2$ for different continents, some differences can be found. The point estimate is substantially larger for South America ($\rho_t = 0.54$), Asia ($\rho_t = 0.32$), Europe ($\rho_t = 0.31$), and North America ($\rho_t = 0.34$), compared to

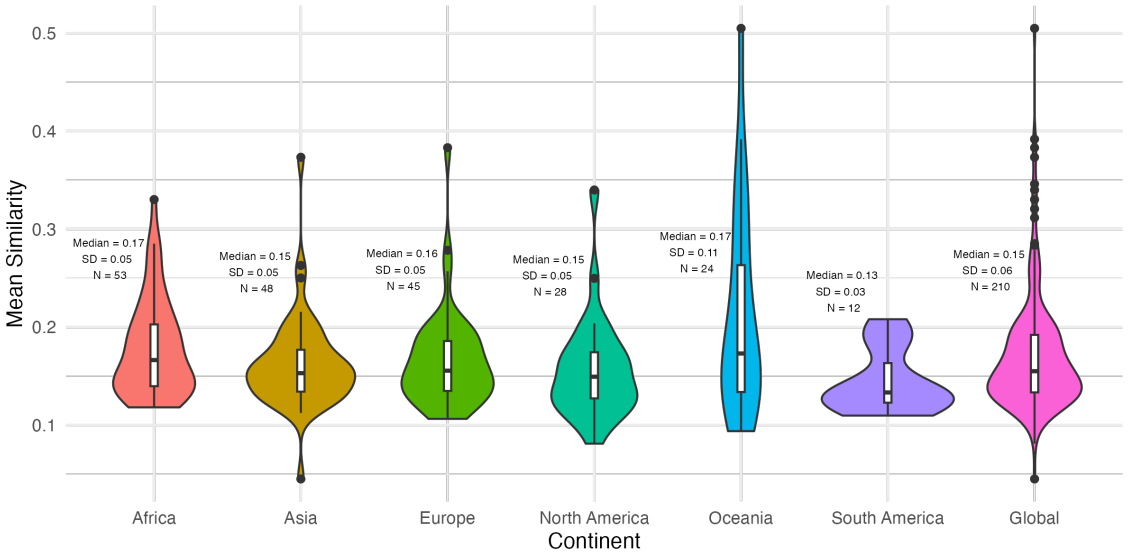


**Figure 3:** Violinplots showing the distribution of mean pairwise similarity per country across continents. Within each violin, the interquartile ranges of mean similarities are displayed using boxplots. Africa, Asia, Europe, and North America show similarities to the global distribution, while Oceania and South America are deviating.

Africa ($\rho_t = 0.106$) and Oceania ($\rho_t = 0.09$). This suggests that countries in Africa and Oceania are characterized by less dominant languages and more even speaker distributions. These results are coherent with the spatial analysis, which saw a shift in focus towards Africa and Oceania as linguistic hotspots when pivoting from $q = 0$ to $q = 2$.

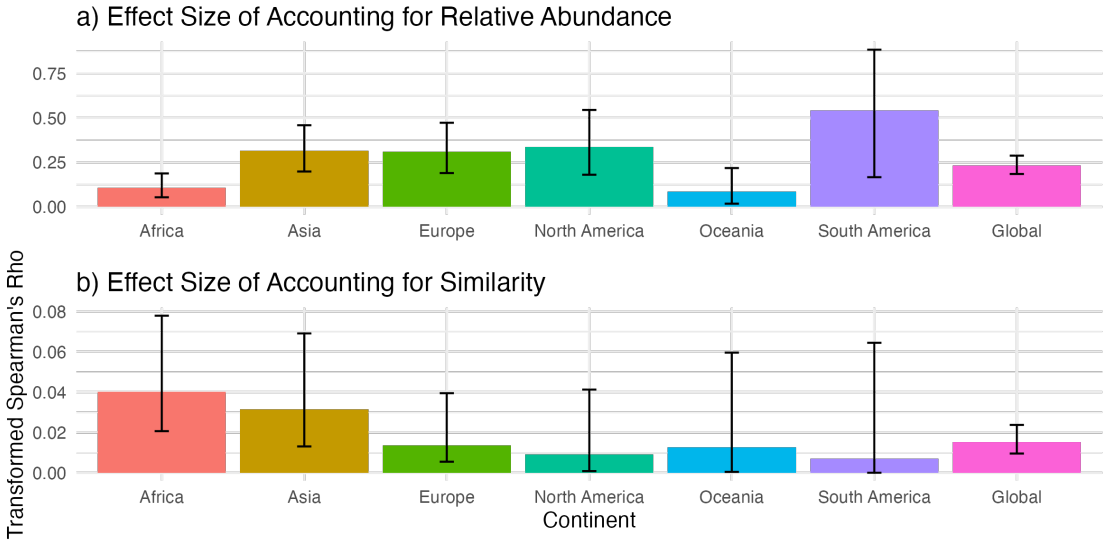


**Figure 4:** Barplots showing the effect of accounting for a) relative abundance and b) similarity globally as well as on a continent basis. The effect size is measured as a transformed Spearman's rho, thus indicating disagreement on rank assignment. 0 equals no effect, 1 equals max effect. Black lines mark 95% confidence intervals.

The same analysis, but correlating the ranks of naive and non-naive diversity for $q = 2$, reveals an overall much smaller impact of accounting for similarity compared to relative abundance.

Globally, Spearman's $\rho_t$ amounts to an abysmal 0.015, signaling a very small impact of accounting for similarity. Some variation between point estimates across continents is found, with the effect suggested to be the largest for Africa ($\rho_t = 0.049$). Therefore, it is generally more important to account for relative abundance compared to similarity. This does however not mean that there aren't cases where accounting for similarity is important. We see that the countries losing the most rankings when accounting for similarity are those where dialect continuums are present, e.g., Madagascar. We also see that the countries gaining the most rankings when accounting for similarity generally are small insular nations, such as Guam and Palau, while larger countries such as Brazil and Colombia lose rankings (see detailed analysis in Appendix 6).

### 3.3 Diversity profiles

Up until now, we have only looked at diversity from three points of view: naive with maximal importance of rare languages; naive with high importance of dominant languages; and non-naive with high importance of dominant languages. The full potential of the Leinster-Cobbold framework lies, however, in the fact that diversity can be considered from an infinite number of points of view, depending on how $q$ is chosen. By calculating diversity for varying values of $q$, and plotting the change, the full diversity structure can be shown, which can be seen for the continents of the world in Figure 5.[3] Considering the naive case in a) and b), for $q = 0$, i.e., richness, Africa is the most diverse continent, followed by Asia, Oceania, North America, South America, and Europe. For all values of $q$, Africa is perceived as the most diverse continent, something that is even more pronounced for larger values of $q$, due to many dominating languages. Europe is perceived as the least diverse continent for $q = 0$ due to the relatively few languages found there, but as the importance of dominant languages increases, it overtakes most continents and is perceived as the second most diverse continent for larger $q$, since multiple dominating languages are evenly distributed, such as German, French, Italian etc. Both South and North America fall rapidly, due to the overwhelming dominance of two languages (Spanish-Portuguese and Spanish-English) converging at approximately 2 languages. Oceania falls slower than Asia for very small $q$, but continues falling sharply for $q > 1$, while Asia starts to flatten, converging on approximately 4.4 languages. For large $q$, Oceania is without a doubt perceived as the least diverse continent, since it is dominated by English, the language of 63% of speakers, resulting in Oceania converging at approximately 1.6 languages. As such, the question, "what continent is more diverse?" can have multiple different answers, each depending on the chosen value of $q$, i.e., how important should dominating languages be, and these answers are only made available by plotting diversity profiles.

Considering similarity as well as seen in c), the diversity of the continents can be understood even better. Africa is still by far perceived as the most diverse continent, but now, Europe does not overtake Asia since the dominant languages in Europe, being Indo-European languages, are relatively similar. In Asia, the dominant languages are not related and quite different, e.g., Mandarin, Hindi, and Japanese. Consequently, a fairly equal dominance structure makes Asia perceived as more diverse for all values of $q$ due to the difference in similarity. While South America and North America both quickly converged to a diversity equivalent to 2 languages in the naive case for large $q$, North America is perceived as substantially more diverse in the non-naive case for all values of $q$. The reason for this is that Portuguese and Spanish are very similar languages relative to English and Spanish, reducing the perceived diversity with an entire language relative to North America, where English and Spanish are contrasted as different languages. Furthermore, also in the non-naive case is Oceania for small $q$ perceived as quite diverse, but falls even more radically for increasing $q$ than in the naive case. The perceived diversity quickly drops below North America, then follows South America, and again converges below South America. This is an unsurprising behavior given

[3] A visualization of the underlying speaker distribution can be seen in Figure 7 in the Appendix.

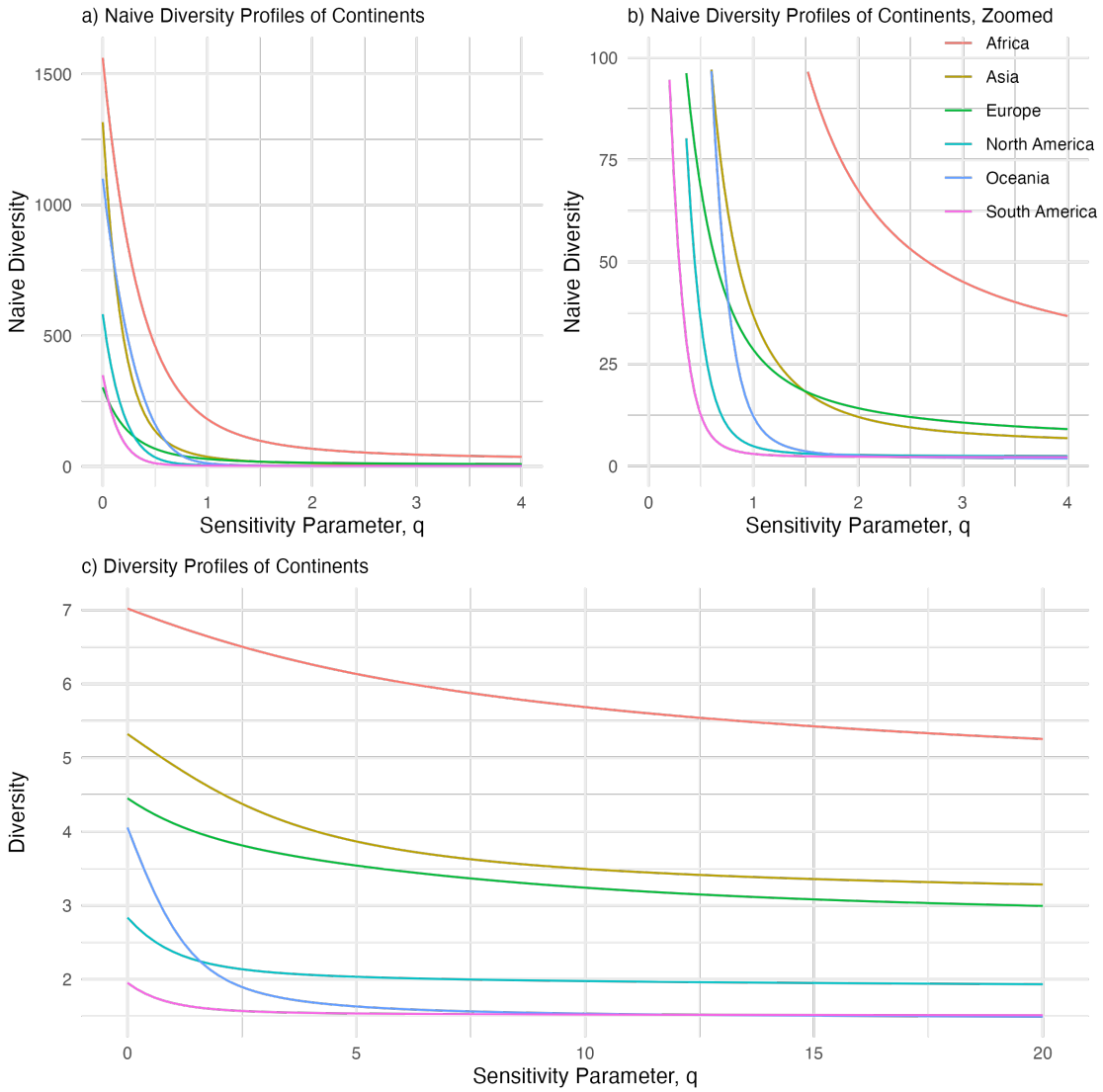


**Figure 5:** Diversity profiles of all continents for increasing $q$ based on naive diversity in a) and b) and based on non-naive diversity in c).

that most of the languages contributing to its incredible richness are small indigenous languages found on the island nations. These languages tend to be related and belong to the Austronesian family, which could be contributing to the sharp decline in perceived diversity in the non-naive case. Furthermore, a possible explanation of why Oceania is perceived as lacking diversity is the incredible dominance of English, which also reflects itself in other dominant languages in the area being English-based Creole languages, such as the second most widely spoken language Tok Pisin. Since the Pidgins are based on English, they will have a high lexical similarity with English, which thus appears even more dominant, and subsequently reduces the diversity for large $q$.

## 4 Discussion

From the analysis, it is clear that how diversity is perceived crucially depends on how it is measured, since there is an observable effect of accounting for both relative abundance and similarity in the context of linguistic diversity. However, this does not mean that one measure is necessarily better

than the other; for example, our analysis showed that when accounting for similarity, Palau is perceived as more diverse than Brazil, even though far more languages are spoken in Brazil. If a researcher is only interested in the range of languages found in a country or any other setting, then there is merit to claiming that Brazil is a more diverse country than Palau. However, we would, similar to Leinster and Cobbold [15], argue that not taking similarity or relative abundance into account when the data exists is equivalent to throwing away information that is relevant to understanding diversity. Therefore, the best course of action is to draw the diversity profiles of settings to give the most accurate description of diversity possible; something that can't be captured by a single measure. That is, the point of this framework is not to identify some optimal value for $q$, but rather to raise awareness for the effect that changing $q$ has on how we perceive diversity.

There are some limitations to this approach, namely that data can be scarce. Drawing diversity profiles requires both speaker and similarity measures, which forced us to disregard more than 2000 languages, 400,000,000 speakers, and 27 countries to enable the analysis. Still, considering the impact of accounting for relative abundance, we are convinced that this modeling step is necessary. What it shows is that we, the linguistic and computational humanities community, need to increase our effort in collecting and making cross-linguistic data available, as this constitutes the upper limit for quantitative modeling of linguistic phenomena. Since similarity data creates the largest constraint, and its impact on perceived diversity was shown to be relatively small except for a few important cases, we would consider it motivated to not account for similarity in a global context and only consider naive diversity. If sufficient data is available, however, and the analysis pertains to countries or areas where dialect continuums exist, the results conclusively point to similarity-aware measures being necessary. Furthermore, the Leinster-Cobbold framework is only one approach to addressing the similarity aspect of diversity—albeit a generalizable and mathematically sound one. Other frameworks, such as [4], exist and could prove relevant for linguistics as well (cf. [14]).

Deriving relative abundance and similarity-aware measures of linguistic diversity allows us to describe how diversity differs across space, as shown in this paper, and potentially—given diachronic data—across time. They can, however, never on their own explain why such differences exist. For that purpose, we need to conduct statistical modeling, such as regression modeling using explanatory variables to better understand linguistic diversity. Bromham et al. [2] showed how different sociodemographic variables can partially explain why languages die, which would equal a reduction in richness. However, this only accounts for one aspect of diversity. As such, a promising avenue for future research would be to utilize the Leinster-Cobbold framework in a regression modeling context and explore how sociodemographic variables can explain variation in different aspects of linguistic diversity, not only richness.

## 5 Conclusion

In this paper, we have explored global linguistic diversity accounting for three important aspects of diversity –richness, relative abundance, and similarity –using the Leinster-Cobbold framework. We have shown that accounting for relative abundance drastically changes how diversity is perceived, accentuating countries in Oceania and Asia as hotspots of linguistic diversity. Accounting for similarity does not impact how diversity is perceived as sharply, but underscores Sub-Saharan Africa as the linguistically most diverse area on the planet, and most importantly, corrects for cases where dialect continua define the linguistic landscape. Hence, limiting an analysis to only richness or only similarity disregards many important aspects of diversity. As such, we advocate for a view on linguistic diversity as a multivariate phenomenon, and suggest that it is appropriately modeled as such.

## Acknowledgements

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

## References

[1] Bakker, Dik, Müller, André, Velupillai, Viveka, Wichmann, Søren, Brown, Cecil H., Brown, Pamela, Egorov, Dmitry, Mailhammer, Robert, Grant, Anthony, and Holman, Eric W. "Adding Typology to Lexicostatistics: A Combined Approach to Language Classification". In: *Linguistic Typology 13*, no. 1 (May 2009), pp. 169–181. DOI: 10.1515/LITY.2009.009. URL: https://doi.org/10.1515/LITY.2009.009.

[2] Bromham, Lindell, Dinnage, Russell, Skirgård, Hedvig, Ritchie, Andrew, Cardillo, Marcel, Meakins, Felicity, Greenhill, Simon, and Hua, Xia. "Global predictors of language endangerment and the future of linguistic diversity". en. In: *Nature Ecology & Evolution 6*, no. 2 (2022), pp. 163–173. ISSN: 2397-334X. DOI: 10.1038/s41559-021-01604-y.

[3] Brown, Cecil, Holman, Eric, Wichmann, Søren, and Velupillai, Viveka. "Automated Classification of the World's Languages: A Description of the Method and Preliminary Results". In: *STUF - Language Typology and Universals 61*, no. 4 (Nov. 2008), pp. 285–308. DOI: 10.1524/stuf.2008.0026.

[4] Chao, Anne, Chiu, Chun-Huo, and Jost, Lou. "Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers". en. In: *Annual Review of Ecology, Evolution, and Systematics 45* (Nov. 2014), pp. 297–324. ISSN: 1543-592X, 1545-2069. DOI: 10.1146/annurev-ecolsys-120213-091540.

[5] Eberhard, David M., Simons, Gary F., and Fennig, Charles D. *Ethnologue: Languages of the World*. 25th ed. Dallas, Texas: SIL International, 2022. URL: https://www.ethnologue.com.

[6] Alexandre François, Sébastien Lacrampe, Michael Franjieh, and Stefan Schnell, edited by. *The Languages of Vanuatu: Unity and Diversity*. Vol. 5. Studies in the Languages of Island Melanesia. Canberra, Australia: Asia-Pacific Linguistics, School of Culture, History, Language, College of Asia, and the Pacific, The Australian National University, 2015. ISBN: 9781922185235. URL: http://hdl.handle.net/1885/14819.

[7] Garnier et al. "viridis(Lite) - Colorblind-Friendly Color Maps for R". viridis package version 0.6.5. 2024. DOI: 10.5281/zenodo.4679423. URL: https://sjmgarnier.github.io/viridis/.

[8] Gooskens, Charlotte and Schneider, Cindy. "Linguistic and non-linguistic factors affecting intelligibility across closely related varieties in Pentecost Island, Vanuatu". In: *Dialectologia 23* (2019), pp. 61–85. ISSN: 2013-2247.

[9] Greenberg, Joseph H. "The Measurement of Linguistic Diversity". In: *Language 32*, no. 1 (1956), pp. 109–115. ISSN: 0097-8507. DOI: 10.2307/410659.

[10] Grin, François and Fürst, Guillaume. "Measuring Linguistic Diversity: A Multi-level Metric". en. In: *Social Indicators Research 164*, no. 2 (Nov. 2022), pp. 601–621. ISSN: 1573-0921. DOI: 10.1007/s11205-022-02934-5.

[11] Harmon, David and Loh, Jonathan. "The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages". en. In: *Language Documentation & Conservation 4* (2010), pp. 97–151.

[12] Holman, Eric, Wichmann, Søren, Brown, Cecil, Velupillai, Viveka, Müller, André, and Bakker, Dik. “Explorations in automated language classification”. In: *Folia Linguistica 42*, no. 3-4 (Nov. 2008), pp. 331–354. DOI: 10.1515/FLIN.2008.331.

[13] Joshua Project. “Joshua Project: People Groups of the World”. Accessed: 2025-04-08. 2025.

[14] Karsdorp, F.B., Manjavacas, Enrique, and Fonteyn, Lauren. “Introducing Functional Diversity: A Novel Approach to Lexical Diversity in (Historical) Corpora”. English. In: *Proceedings of the Computational Humanities Research Conference 2022*, ed. by Folgert Karsdorp, Alie Lassche, and Kristoffer Nielbo. Vol. 3290. CEUR Workshop Proceedings, Nov. 2022, pp. 114–126.

[15] Leinster, Tom and Cobbold, Christina A. “Measuring diversity: the importance of species similarity”. en. In: *Ecology 93*, no. 3 (2012), pp. 477–489. ISSN: 1939-9170. DOI: 10.1890/10-2402.1.

[16] Leinster, Tom and Meckes, Mark W. “Maximizing Diversity in Biology and Beyond”. en. In: *Entropy 18*, no. 3 (Mar. 2016), p. 88. ISSN: 1099-4300. DOI: 10.3390/e18030088.

[17] Levenshtein, Vladimir I et al. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. Soviet Union. 1966, pp. 707–710.

[18] List, Johann-Mattis, Forkel, Robert, Greenhill, Simon J., Rzymski, Christoph, Englisch, Johannes, and Gray, Russell D. “Lexibank, a public repository of standardized wordlists with computed phonological and lexical features”. en. In: *Scientific Data 9*, no. 9 (2022), p. 316. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01432-0.

[19] Massicotte, Philippe and South, Andy. “rnaturalearth: World Map Data from Natural Earth”. R package version 1.0.1. 2023. URL: https://CRAN.R-project.org/package=rnaturalearth.

[20] Petroni, Filippo and Serva, Maurizio. “Measures of lexical distance between languages”. In: *Physica A: Statistical Mechanics and its Applications 389*, no. 11 (June 2010), pp. 2280–2283. ISSN: 0378-4371. DOI: 10.1016/j.physa.2010.02.004.

[21] R Core Team. “R: A Language and Environment for Statistical Computing”. R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[22] Rao, C. Radhakrishna. “Diversity and dissimilarity coefficients: A unified approach”. In: *Theoretical Population Biology 21*, no. 1 (Feb. 1982), pp. 24–43. ISSN: 0040-5809. DOI: 10.1016/0040-5809(82)90004-1.

[23] SIL International. “About ISO 639-3: Codes for the representation of names of languages – Part 3”. https://iso639-3.sil.org/about. Accessed: 2025-06-13.

[24] Simpson, E. H. “Measurement of Diversity”. en. In: *Nature 163*, no. 688 (Apr. 1949), pp. 688–688. ISSN: 1476-4687. DOI: 10.1038/163688a0.

[25] Skirgård, Hedvig et al. “Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss”. In: *Science Advances 9*, no. 16 (2023). DOI: 10.1126/sciadv.adg6175.

[26] Spearman, Charles. “The Proof and Measurement of Association Between Two Things”. In: *The American Journal of Psychology 15*, no. 1 (1904), pp. 72–101. URL: http://www.jstor.org/stable/1412159.

[27] Wichmann, Søren, Holman, Eric W., and Brown, Cecil H. “The ASJP Database (version 20)”, ed. by Søren Wichmann, Eric W. Holman, and Cecil H. Brown. https://asjp.clld.org/. Accessed: 2025-04-14. 2022.

## 6 First Appendix Section

In Figure 6b), we can see that the countries are concentrated around the identity line, suggesting high agreement, but there are some outliers. The general trend is that some countries lose many ranks when accounting for similarity, e.g., Madagascar, which loses 71 ranks due to accounting for similarity. The reason for this is the dominating languages on Madagascar being varieties of Malagasy, creating a dialect continuum, with the languages being to varying degrees mutually intelligible. Therefore, in countries or regions where languages or linguistic varieties are closely related, it is crucial to account for similarity to model linguistic diversity adequately. Still, however, this is less important than accounting for relative abundance, as can be seen in a) where countries are more sparsely distributed, further away from the identity line. Here, it is interesting to note that the countries losing the most rankings are generally countries with relatively large populations, such as Brazil, which loses 182 rankings, going from being perceived as one of the most diverse countries in the world to one of the least diverse countries. On the other hand, the countries gaining the most rankings are generally small insular nations such as Guam and Palau. Reasonably, due to their small sizes, comparatively few languages are found, but these languages are more evenly distributed, resulting in them being perceived as more diverse than countries with many more languages, such as Brazil or Colombia.

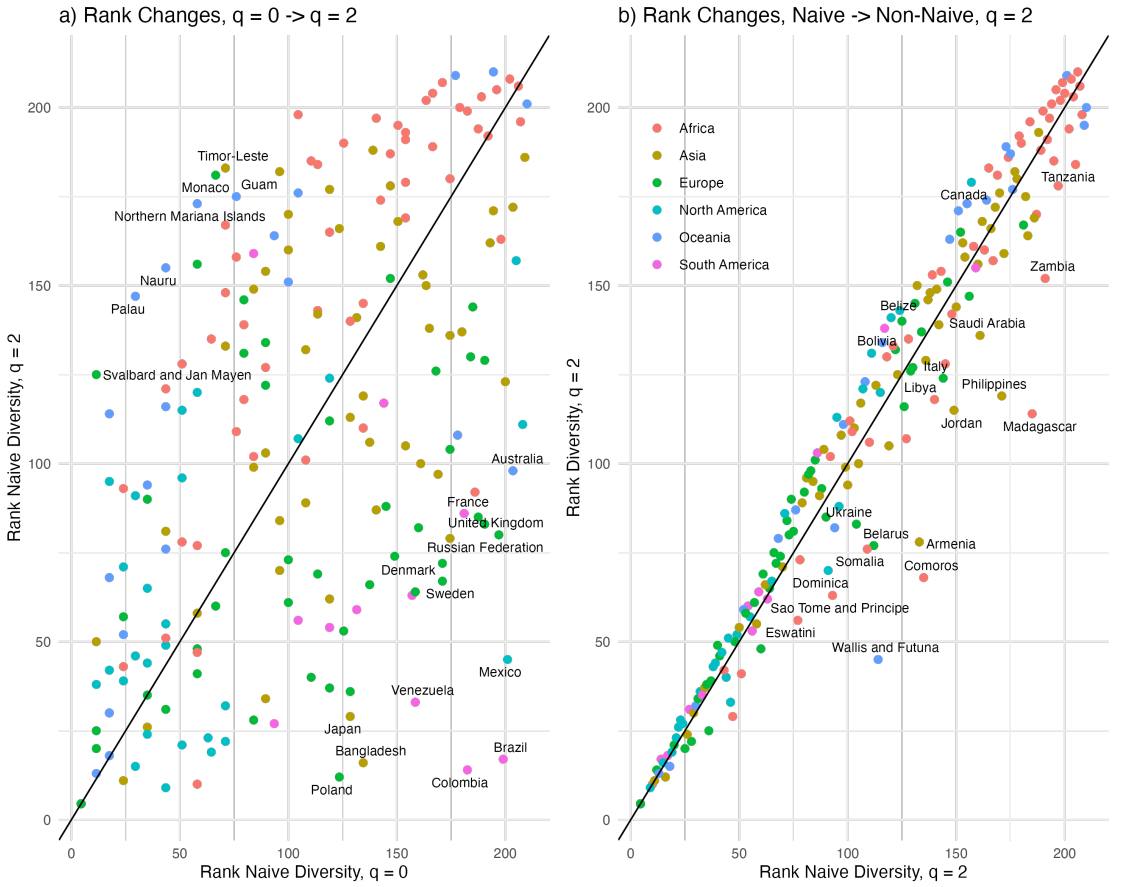


**Figure 6:** Scatterplots of diversity rankings according to a) naive diversity for $q = 0$ and $q = 2$, and b) naive diversity $q = 2$ and diversity $q = 2$. The black line is the identity line, and the further away from this line a country is, the larger the rank shift between the measures.

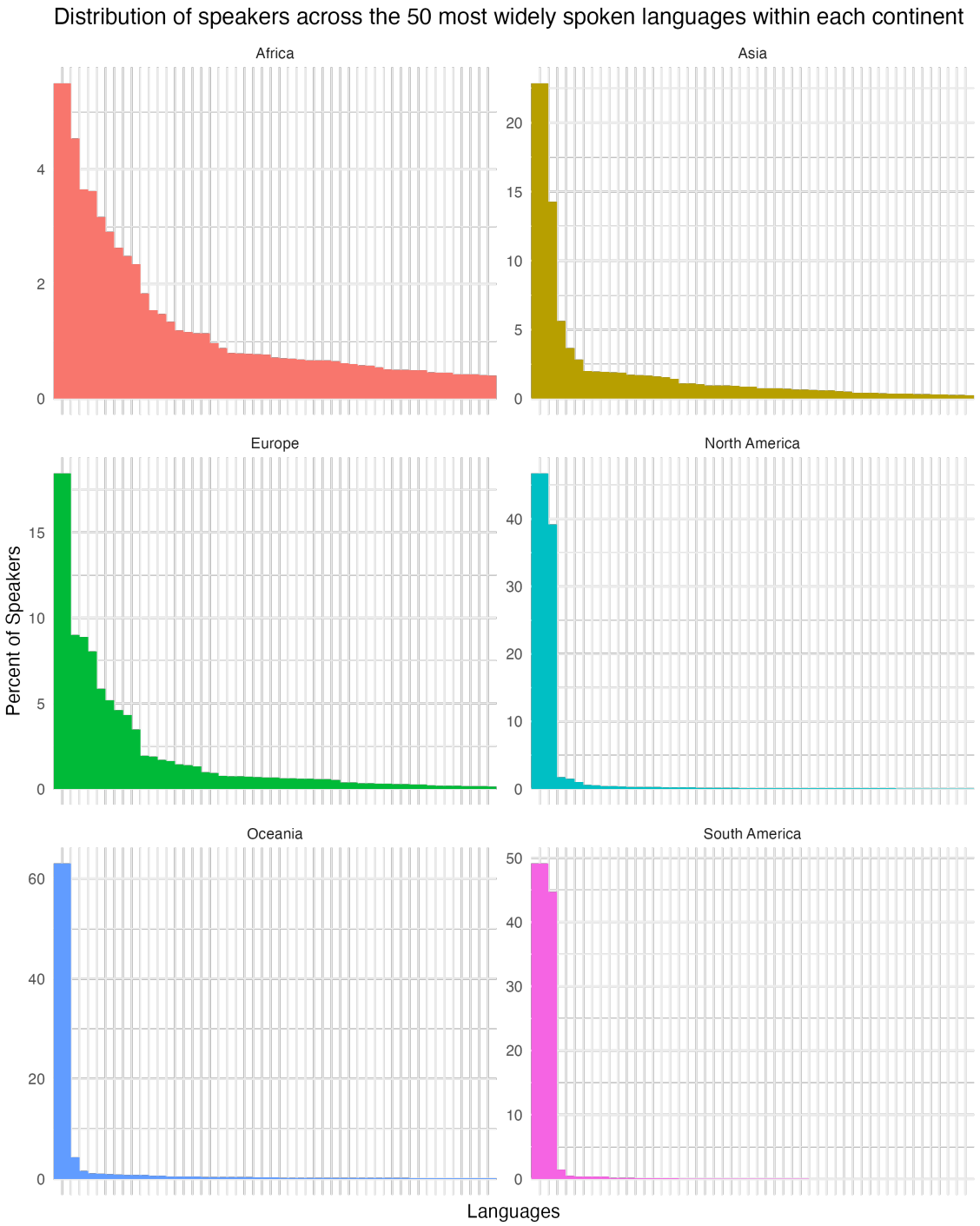


**Figure 7:** Barplots showing the distribution of speakers across the 50 most widely spoken languages for each continent. A bar represents the percent of all speakers on a given continent speaking a given language, sorted from most abundant to least abundant. The largest concentration of speakers into a few languages is found in the Americas and Oceania.