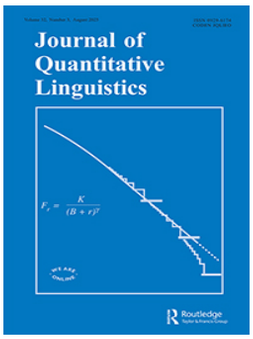
Measuring Linguistic Diversity: Limits and Extensions of the Greenberg Index

Civico, Marco

How to cite

CIVICO, Marco. Measuring Linguistic Diversity: Limits and Extensions of the Greenberg Index. In: Journal of quantitative linguistics, 2025, p. 1–28. doi: 10.1080/09296174.2025.2567699

This publication URL: https://archive-ouverte.unige.ch/unige:188159
Publication DOI: 10.1080/09296174.2025.2567699



Routledge
Taylor & Francis Group

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: www.tandfonline.com/journals/njql20

Measuring Linguistic Diversity: Limits and Extensions of the Greenberg Index

Marco Civico

To cite this article: Marco Civico (06 Oct 2025): Measuring Linguistic Diversity: Limits and Extensions of the Greenberg Index, Journal of Quantitative Linguistics, DOI: 10.1080/09296174.2025.2567699

To link to this article: https://doi.org/10.1080/09296174.2025.2567699

Published online: 06 Oct 2025.

JOURNAL OF QUANTITATIVE LINGUISTICS
https://doi.org/10.1080/09296174.2025.2567699

Routledge
Taylor & Francis Group

Measuring Linguistic Diversity: Limits and Extensions of the Greenberg Index

Marco Civico

Faculty of Translation and Interpreting, University of Geneva, Geneva, Switzerland

ABSTRACT

This theoretical paper proposes a refined framework for measuring linguistic diversity in multilingual societies, with a focus on the Greenberg Diversity Index (GDI), a widely used indicator based on the probability that two randomly selected individuals speak different languages. After deriving exact upper and lower bounds for GDI under varying language distributions, the paper introduces several theoretical extensions that capture more realistic sociolinguistic configurations. These include competence-weighted indices to account for individual multilingualism, co-competence matrices to model overlapping language repertoires, distance-weighted formulations based on interlinguistic similarity, and entropy-based metrics that reflect the unpredictability of language identity. Each index is analysed in terms of its mathematical structure, interpretability, and alignment with different types of linguistic data. A comparative framework is proposed to support the selection and interpretation of these metrics across research and policy applications. Together, these contributions advance the quantitative modelling of linguistic diversity and offer tools for more nuanced sociolinguistic analysis.

1. Introduction

Linguistic diversity is a central feature of many contemporary societies, shaping patterns of communication, social integration, and institutional design. Understanding and quantifying this diversity is essential not only for descriptive purposes, but also for informing language policies, educational planning, and the management of multilingual public services. Despite its significance, there is no single way to define or measure linguistic diversity: different metrics capture different aspects, each grounded in distinct theoretical and empirical assumptions.

At its core, linguistic diversity involves at least three interrelated dimensions: *richness* (the number of distinct languages), *evenness* (the balance of speaker shares across those languages), and *distance* (the degree of difference between the languages). This tripartite structure mirrors the general theory

CONTACT Marco Civico marco.civico@unige.ch

of diversity proposed by Stirling (2007), who argues that any measure of diversity must be sensitive to the *variety*, *balance*, and *disparity* of the elements in a system. In the linguistic domain, this translates into the need for tools that go beyond simple counts or proportions and incorporate relations among languages and speaker repertoires.

Formal indices of linguistic diversity aim to operationalize these dimensions, possibly in an integrated way. Some are grounded in probabilistic models of mutual comprehension or information exchange (Gazzola et al., 2020), others draw from ecology or information theory (Grin & Fürst, 2022; Gullifer & Titone, 2020; Jost, 2007), and still others exploit language genealogies or typological distances (Desmet et al., 2012).

Among the earliest and most widely used indices is the *Greenberg Diversity Index* (GDI) (Greenberg, 1956), which estimates the probability that two randomly selected individuals from a population do not share the same language. Closely related to the Simpson index in ecology and the Herfindahl index in economics, GDI provides a compact and interpretable measure of linguistic heterogeneity. However, it also embodies several simplifying assumptions: that each person is assigned to exactly one language group (typically based on their primary or native language);[1] that languages are treated as mutually exclusive categories; and that all language differences are considered equally salient.

These assumptions become problematic in contemporary multilingual contexts, where bilingualism and overlapping repertoires are common, and where linguistic distances may matter more than categorical differences. While the classical GDI and related indices (e.g. Herfindahl, Simpson) are typically applied to L1 distributions in policy contexts (such as language-of-instruction planning or teacher deployment), this paper addresses a broader question: how to capture the full diversity of a linguistic landscape shaped by L2s, L3s, and other forms of multilingual competence.[2]

This theoretical paper offers a comprehensive examination of the structure and limitations of the GDI and proposes several conceptual and operational extensions. After deriving theoretical bounds on linguistic diversity for a given number of languages, I introduce alternative formulations that:

- incorporate *individual multilingualism* through competence-weighted indices;
- measure *relational diversity* via pairwise *co-competence* across repertoires, capturing the extent to which individuals share languages with others in the population;
- account for *linguistic similarity* through distance-weighted diversity measures;

- use *Shannon entropy* to reflect the informational diversity of the language distribution, offering greater sensitivity to smaller groups than the Greenberg index.

Each of these extensions retains the probabilistic intuition behind GDI while relaxing its restrictive assumptions. Taken together, they offer a more flexible and realistic framework for measuring linguistic diversity in complex societies.

The remainder of the paper is organized as follows. Section 2 revisits the theoretical properties of the Greenberg index, establishing its exact upper and lower bounds and clarifying its interpretation. Section 3 develops several generalizations of GDI adapted to multilingual settings, including competence-weighted, distance-sensitive, and entropy-based variants. Section 4 compares these metrics empirically and theoretically, and Section 5 concludes by outlining criteria for selecting diversity measures according to specific research goals and available data.

2. The Original Greenberg Diversity Index

The Greenberg Diversity Index, introduced by Joseph Greenberg in 1956 (Greenberg, 1956), is a widely used measure of linguistic diversity that captures two key components:

- *richness*, the number of distinct languages n spoken in a population; and
- *evenness*, the distribution of individuals across these languages.

Definition 1 (Greenberg Diversity Index). *Let* $\mathbf{p} = (p_1, \ldots, p_n) \in \Delta_n$ *denote a probability distribution over n languages*,[3] *where each individual is assumed to speak exactly one native language. The component* $p_i \in [0, 1]$ *represents the proportion of individuals who speak language i. The probability that two randomly selected individuals speak the same language is given by the squared* ℓ_2*-norm of the distribution*:

$$\sum_{i=1}^{n} p_i^2 = \| \mathbf{p} \|_2^2 . \tag{1}$$

The Greenberg Diversity Index is defined as:

$$GDI(\mathbf{p}) = 1 - \sum_{i=1}^{n} p_i^2 = 1 - \| \mathbf{p} \|_2^2 . \tag{2}$$

This expression gives the probability that two randomly chosen individuals *do not* speak the same language. Notably, this measure was independently formulated in other fields as well. Indeed, it is algebraically equivalent

to the Gini – Simpson index of biodiversity (Simpson, 1949) and to the complement of the Herfindahl – Hirschman concentration index used in economics.[4]

Example 1. *As an example, consider a population where three languages are spoken with proportions* $\mathbf{p} = (0.5, 0.3, 0.2)$. *Then*:

$$GDI = 1 - (0.5^2 + 0.3^2 + 0.2^2) = 1 - (0.25 + 0.09 + 0.04) = 1 - 0.38 = 0.62.$$

This result provides a numerical reflection of linguistic diversity, in this case with some imbalance in speaker shares.

Proposition 1 (Theoretical bounds of the Greenberg Diversity Index). *Let* $\mathbf{p} = (p_1, \ldots, p_n) \in \Delta_n$ *be a probability distribution over* $n \geq 1$ *languages. Then*:

$$0 \leq GDI(\mathbf{p}) \leq 1 - \frac{1}{n} \tag{3}$$

The lower bound is achieved when all mass is on a single language. The upper bound is achieved when all languages are equally represented. A detailed proof is provided in Appendix A.

The upper bound of the Greenberg Diversity Index increases with the number of languages n, but approaches 1 asymptotically. For example:

$$n = 2 \Rightarrow \mathrm{GDI}_{\max} = 0.5, \quad n = 5 \Rightarrow \mathrm{GDI}_{\max} = 0.8, \quad n = 10 \Rightarrow \mathrm{GDI}_{\max} = 0.9.$$

This limiting behaviour reflects a situation of extreme linguistic fragmentation, such as when each individual in the population speaks a different language (i.e. the number of languages equals the population size).

These theoretical bounds offer both descriptive and normative guidance for interpreting linguistic diversity:

- They enable the normalization of observed GDI values by comparing them to the maximum achievable value for a given n, thereby quantifying the degree of imbalance in the distribution of speakers.
- They clarify that linguistic diversity is jointly determined by the number of languages (*richness*) and their relative proportions (*evenness*). A population with many languages may nonetheless exhibit low GDI if one language dominates demographically.
- They highlight a pattern of diminishing marginal returns: the addition of new languages yields progressively smaller increases in maximal diversity as n grows. That is, $\mathrm{GDI}_{\max} = 1 - \frac{1}{n}$ approaches 1 at

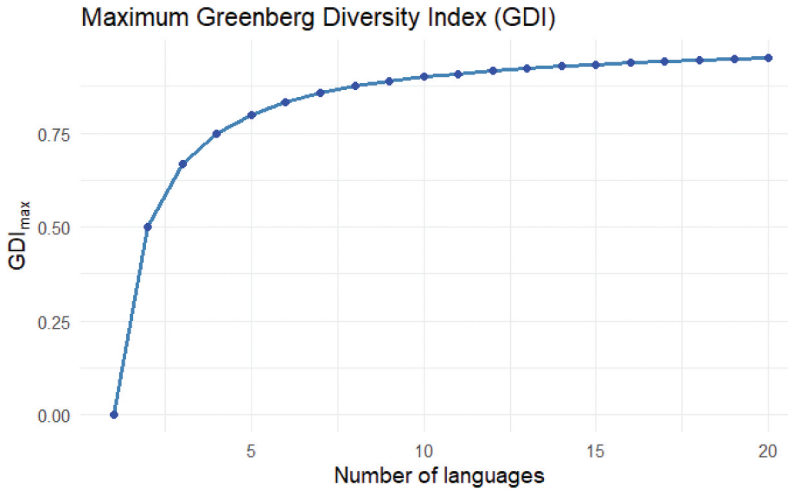


Figure 1. Maximum theoretical values of the Greenberg diversity index (GDI) as a function of the number of languages. The curve follows the formula $GDI_{\max} = 1 - \frac{1}{n}$.

a decreasing rate, and additional languages only meaningfully raise diversity when they are spoken by non-negligible segments of the population.

This asymptotic behaviour has implications for both cross-national comparisons and policy assessment. When comparing linguistic diversity across settings with different values of n, it is advisable to normalize the GDI (e.g. by dividing by $1 - \frac{1}{n}$) to avoid misleading conclusions due purely to the size of the language inventory.

Figure 1 shows how the theoretical upper bound of the Greenberg Diversity Index increases as the number of equally represented languages (n) grows. The relationship is asymptotic: while the index approaches 1 as $n \to \infty$, the marginal increase in diversity diminishes with each additional language. This highlights that increasing linguistic diversity in practice depends not only on the number of languages present, but also on their relative balance, an aspect that can prove relevant for several types of policy development.

Figure 2 illustrates how the Greenberg Diversity Index behaves when only two languages are present in a population. In this case, letting p_1 denote the proportion of speakers of the first language and $p_2 = 1 - p_1$, the index simplifies to:

$$\mathrm{GDI}(p_1) = 1 - \left(p_1^2 + (1 - p_1)^2\right) = 2p_1(1 - p_1),$$

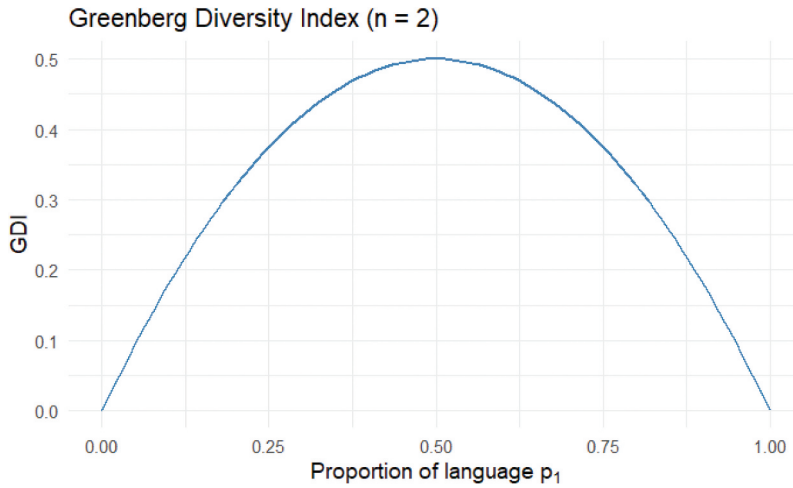


Figure 2. GDI as a function of the proportion p_1 of speakers of language 1, in the case of exactly two languages. Since $n = 2$, the proportion of speakers of the second language is given by $p_2 = 1 - p_1$. The GDI is maximized when both languages are spoken equally, i.e. when $p_1 = p_2 = 0.5$.

which is a symmetric quadratic function with maximum value 0.5 when $p_1 = 0.5$. As soon as one language becomes dominant, the diversity index declines symmetrically, reflecting the growing linguistic concentration. This minimal setting exemplifies how the GDI captures both *richness* and *evenness* in a simple analytical form, making its behaviour easy to interpret even without graphical tools.

2.1. Interpretation and Implications

The bounds derived in the previous section provide a clear reference frame for assessing linguistic diversity in real-world populations. For any given number of languages n, the GDI cannot exceed $1 - \frac{1}{n}$, even in the case of perfect evenness. This means that the number of languages places a structural ceiling on the potential diversity of a population. At the same time, the actual value of the GDI is sensitive to the distribution of speakers. A population with many languages may nonetheless display low diversity if one language dominates the distribution. Conversely, a small number of languages can yield relatively high GDI values if their shares are balanced.

This highlights a crucial point: diversity is not only a matter of how many languages are present (*richness*), but also of how equitably they are represented (*evenness*). This distinction is especially relevant in multilingual societies with large dominant groups and smaller minorities. Two countries with the same number of languages may have very different GDI values

depending on their demographic balance. Moreover, aggregate indices like the GDI do not account for the spatial distribution of language communities. In countries such as Canada or Switzerland, linguistic groups tend to be regionally concentrated, with French in Quebec, German in central and eastern Switzerland, and so on. As a result, while national GDI values may suggest a low average probability of communicative overlap, the likelihood that any two individuals within the same region share a language can be quite high. A similar situation holds in Finland, where Swedish speakers are concentrated in specific coastal areas. India presents an even more complex case: national-level diversity is high, but common knowledge of Hindi and English among the educated and urban populations facilitates communication across regions, despite the vast number of local and regional languages. These cases illustrate how linguistic diversity, as captured by global indices, may not fully reflect the communication experience, which is often shaped by regional clustering and overlapping repertoires. Incorporating spatial or interaction-based perspectives, such as co-competence or network measures, may provide a more nuanced understanding of diversity in such contexts.

The upper bound $1 - \frac{1}{n}$ also implies diminishing returns to linguistic diversity: each additional language contributes less to the maximum possible GDI. For example, increasing n from 2 to 3 increases the upper bound from 0.5 to 0.667, but increasing n from 10 to 11 increases it only from 0.9 to approximately 0.909. This property is particularly relevant when comparing highly multilingual societies. From a policy perspective, comparing a population's actual GDI to its theoretical maximum can provide a useful index of linguistic concentration or inequality. A large gap between actual and maximal GDI suggests strong asymmetries in language presence or use. Conversely, values close to the upper bound may indicate robust coexistence or balanced multilingualism.

Finally, these observations underscore the importance of using normalized diversity metrics when comparing across countries or regions. Without accounting for the theoretical limits imposed by n, comparisons of raw GDI values may lead to misleading conclusions about the underlying linguistic landscape.

3. Applications and Extensions

The theoretical bounds derived for the Greenberg Diversity Index offer practical benchmarks for empirical research in sociolinguistics, language policy, and multilingual studies in general. By comparing observed GDI values to their maximum theoretical values for a given number of languages n, researchers can assess the degree of linguistic concentration or imbalance in a population, independent of raw language counts.

This approach enables several types of analysis. First, it facilitates meaningful cross-country or cross-regional comparisons. Since the maximum achievable GDI depends on n, normalizing observed values by their theoretical maximum (GDI/GDI_{max}) allows for more interpretable comparisons across populations with different levels of linguistic richness. Second, temporal trends in GDI or its normalized version can be used to monitor the impact of migration, language shift, education, or policy interventions. A declining GDI may indicate growing dominance by a single language, while an increasing GDI could reflect revitalization or stabilization of minoritized languages. In many contexts, such shifts would already be noticeable through qualitative observation or anecdotal evidence; the GDI thus offers a way to objectively corroborate and quantify these perceived trends. Third, the gap between observed and maximal GDI may serve as an indicator of linguistic inequality. A substantial gap could point to asymmetries in language prestige, institutional support, or intergenerational transmission, and can be interpreted alongside other measures of linguistic justice or inclusion.

The classical GDI provides a simple yet powerful measure of linguistic diversity, but its assumptions limit its ability to capture the full complexity of multilingual societies. In particular, GDI assumes mutually exclusive language groups and does not account for individuals who use multiple languages, for differences in linguistic similarity, or for the functional roles of languages within a society. To address these limitations, various extensions and refinements have been proposed or can be developed. In the rest of the paper, I shall present these possible extensions, illustrating how they expand the analytical scope of linguistic diversity measurement.

3.1. Competence Vs. Frequency of Use

Before introducing formal extensions of the Greenberg Diversity Index, it is important to clarify a key conceptual distinction that affects how linguistic diversity is measured: the difference between *competence* in a language and its *frequency of use*. In traditional applications of the GDI, individuals are typically assigned to a single language group (usually their native or most used language) not because multilingualism was denied, but because this reflected the research focus of the time. This simplifying assumption facilitates aggregation, but overlooks the widespread presence of multilingual speakers and the diversity of language use across social domains. More general formulations must therefore decide whether to model language knowledge in terms of:

- *competence*, that is, the ability to understand, speak, read, or write in a language, regardless of context or frequency; or
- *use*, i.e. the actual frequency or proportion of time a language is used by an individual across contexts.

Competence-based profiles allow individuals to score highly in more than one language (e.g. a speaker may be fully competent in both Spanish and Catalan), and can be useful for modelling latent linguistic resources, education policy, or intergenerational transmission. However, they typically require normalization when aggregated, as individuals may contribute more than one unit of weight to the population totals.[5] Use-based profiles, by contrast, generally constrain the sum of weights for each individual to 1, reflecting the distribution of their linguistic behaviour across languages. This approach is well suited to sociolinguistic fieldwork and descriptive statistics, as it captures actual usage patterns. However, because it reflects relatively short-term, context-specific behaviours, it may underrepresent latent multilingual competences and is not always an adequate basis for long-term policy planning.

In this paper, I adopt a competence-based modelling framework for both conceptual and analytical reasons. While use-based indices can be derived as a special case, a competence-based approach is more suitable for capturing the potential for mutual intelligibility between individuals. Since the original GDI estimates the probability that two randomly selected individuals speak different languages, it is natural to extend it in a way that reflects their ability to communicate not only through shared mother tongues, but also via additional languages in which they have some level of competence. In multilingual societies, speakers often shift across languages depending on context, and communication may occur in a language that is not dominant for either party. For this reason, modelling competence rather than usage better aligns with the communicative logic underlying the original GDI.[6]

3.2. Accounting for Individual Multilingualism

3.2.1. Competence-Weighted GDI

In many multilingual societies, individuals may be able to speak more than one language across different domains of life. This raises the question of how GDI might be extended to reflect individual multilingualism more accurately. One intuitive extension is to replace categorical speaker counts with weighted language profiles. For each individual k, we define a vector of weights $w_{ki} \in [0, 1]$, where each entry represents the *degree of competence* in each language (e.g. fluency, literacy), where no constraint is placed on the sum; for example, a fully bilingual individual in a trilingual context might be represented by the vector $(1, 1, 0)$, where each value indicates competence in one language.[7]

Definition 2 (Competence-weighted Greenberg Index). *Let* $\mathbf{W} \in \mathbb{R}^{\mathbb{N}\times\kappa}$ *be a matrix of individual language competences in n languages for a population of*

N individuals. Each row $\mathbf{w}_k \in \mathbb{R}^{\ltimes}$ *denotes the competence vector of individual k, with entries* $w_{ki} \in [0, 1]$.

Define the total competence mass as:

$$Z = \sum_{k=1}^{N} \sum_{i=1}^{n} w_{ki} = \sum_{k=1}^{N} \| \mathbf{w}_k \|_1. \tag{4}$$

Then the Competence-weighted Greenberg Diversity Index is:

$$GDI^{comp}(\mathbf{W}) = 1 - \left\| \frac{1}{Z} \sum_{k=1}^{N} \mathbf{w}_k \right\|_2^2. \tag{5}$$

Equivalently, if $\mathbf{w}_{\text{tot}} = \sum_{k=1}^{N} \mathbf{w}_k \in \mathbb{R}^{\ltimes}$ *is the aggregate competence vector, and* $\hat{\mathbf{d}} = \mathbf{w}_{\text{tot}}/Z$, *then*:

$$GDI^{comp}(\mathbf{W}) = 1 - \| \hat{\mathbf{d}} \|_2^2 . \tag{6}$$

This formulation emphasizes that GDI^{comp} is a function of the aggregate competence distribution, normalized to form a probability vector. It generalizes the classical GDI to continuous competence levels and multilingual profiles.

3.2.2. Properties

• If each individual is monolingual and has full competence in exactly one language (i.e. each $\mathbf{w}_k$ is a unit vector), then $\text{GDI}^{comp}(\mathbf{W})$ reduces to the classical GDI.

• The index satisfies the same bounds as the classical GDI, given in Equation (3).

Example 2. *Consider a population of* $N = 4$ *individuals and* $n = 3$ *languages, labelled A, B, and C. Each individual has a competence profile represented as a row vector in the competence matrix* $\mathbf{W} \in \mathbb{R}^{4\times3}$, *with the following entries*:

$$\mathbf{W} = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 1.0 & 0.4 & 0.0 \\ 0.0 & 1.0 & 0.5 \end{bmatrix}$$

Each row corresponds to an individual's competences in languages A, B, and C respectively. The total competence vector is obtained by summing across the individual competence vectors, i.e. summing each column of $\mathbf{W}$ *across all individuals*:

$$\mathbf{w}_{\text{tot}} = \sum_{k=1}^{4} \mathbf{w}_k = [\,1.0 + 0.0 + 1.0 + 0.0, \quad 0.0 + 1.0 + 0.4 + 1.0, \quad 0.0 + 0.0 + 0.0 + 0.5\,] \\ = [\,2.0 \quad 2.4 \quad 0.5\,]$$

The total competence mass is:

$$Z = \sum_{i=1}^{3} w_{\text{tot},i} = 2.0 + 2.4 + 0.5 = 4.9.$$

Normalizing yields the distribution:

$$\hat{\mathbf{p}} = \frac{1}{Z}\mathbf{w}_{\text{tot}} = \left(\frac{2.0}{4.9}, \frac{2.4}{4.9}, \frac{0.5}{4.9}\right) \approx (0.4082,\ 0.4898,\ 0.1020).$$

The competence-weighted Greenberg Index is then:

$$GDI^{comp}(\mathbf{W}) = 1 - \| \hat{\mathbf{p}} \|_2^2 = 1 - (0.4082^2 + 0.4898^2 + 0.1020^2) \\ \approx 1 - (0.1666 + 0.2409 + 0.0104) = 0.582.$$

For comparison, if each individual were assigned to a single language (namely, their native language, for which competence = 1) the resulting counts would be: A (2), B (2), C (0), giving:

$$\text{GDI} = 1 - (0.5^2 + 0.5^2) = 0.5.$$

This illustrates how the competence-based extension captures additional linguistic diversity contributed by multilingual individuals, compared to the classical GDI. Beyond its conceptual and formal interest, the competence-weighted formulation of the Greenberg index offers practical advantages for language policy. Its linear and differentiable structure allows for integration into optimization models, enabling policymakers to simulate and evaluate the impact of interventions such as targeted language education or translation services. In the Swiss context, for example, this index could help assess how shifts in second-language competence (say, increased German proficiency in French-speaking cantons) affect overall linguistic cohesion and mutual understanding. By incorporating partial competences rather than relying solely on declared native language, the index provides a more realistic picture of communicative potential in a multilingual society. Moreover, its mathematical tractability facilitates empirical estimation from microdata (e.g. Swiss censuses or PISA language self-assessments) , and makes it suitable for tracking changes in linguistic diversity over time in response to federal and cantonal policy initiatives.

3.2.3. Alternative Approach: Co-Competence Matrix

A second complementary extension to the classical GDI shifts the focus from aggregated language distributions to interpersonal similarity in linguistic

competence. Rather than treating diversity as a property of group-level distributions, this formulation considers it as a function of how easily individuals can communicate with one another based on the overlap in their repertoires.

Let each individual $i \in \{1, \ldots, N\}$ be associated with a vector of language competences $\mathbf{w}_i = (w_{i1}, \ldots, w_{in}) \in [0,1]^n$, with at least one language at full competence ($w_{ik} = 1$) to reflect their native or dominant language.

We define the *similarity* $s_{ij} \in [0,1]$ between individuals i and j as:

$$s_{ij} = \frac{\sum_{k=1}^{n} \min(w_{ik}, w_{jk})}{\min\left(\sum_{k=1}^{n} w_{ik},\ \sum_{k=1}^{n} w_{jk}\right)}. \tag{7}$$

This score reflects the extent to which the smaller of the two repertoires is covered by the overlap between them. It can be interpreted as the proportion of the less linguistically equipped individual's competence that is shared with their interlocutor, and thus indicates the potential for mutual understanding. A value of $s_{ij} = 1$ implies that the weaker repertoire is entirely included in the stronger one, signalling *maximal communicative potential*. Conversely, a value of 0 implies that there is no linguistic overlap.

Definition 3 (Co-Competence Greenberg Index). *Let* $\mathbf{W} \in \mathbb{R}^{\mathbb{N}\times\kappa}$ *be a matrix of individual language competences, and* s_{ij} *be the similarity between individuals i and j as defined above. Then the Co-Competence Greenberg Index is given by*:

$$GDI^{\text{co}}(\mathbf{W}) = 1 - \frac{1}{\binom{N}{2}} \sum_{i<j} s_{ij}.$$

This formulation mirrors the logic of the classical GDI, which captures the expected difference between two individuals sampled at random. However, rather than aggregating over speaker proportions per language, it computes interpersonal similarity scores derived from overlap in competence profiles. The final index reflects the *average dissimilarity* in language repertoires across the population.

Example 3. *Consider a population of* $N = 4$ *individuals and* $n = 3$ *languages (A, B, C), and the same competence matrix* $\mathbf{W}$ *used is the same as in Example 2. We compute the similarity scores* s_{ij} *for each unordered pair* (i, j) *as follows*:

$$s_{12} = \frac{\min(1.0, 0.0) + \min(0.0, 1.0) + \min(0.0, 0.0)}{\min(1.0,\ 1.0)} = 0$$

$$s_{13} = \frac{\min(1.0, 1.0) + \min(0.0, 0.4) + \min(0.0, 0.0)}{\min(1.0,\ 1.4)} = \frac{1.0}{1.0} = 1.0$$

$$s_{14} = \frac{\min(1.0, 0.0) + \min(0.0, 1.0) + \min(0.0, 0.5)}{\min(1.0,\ 1.5)} = 0$$

$$s_{23} = \frac{\min(0.0, 1.0) + \min(1.0, 0.4) + \min(0.0, 0.0)}{\min(1.0,\ 1.4)} = \frac{0.4}{1.0} = 0.4$$

$$s_{24} = \frac{\min(0.0, 0.0) + \min(1.0, 1.0) + \min(0.0, 0.5)}{\min(1.0,\ 1.5)} = \frac{1.0}{1.0} = 1.0$$

$$s_{34} = \frac{\min(1.0, 0.0) + \min(0.4, 1.0) + \min(0.0, 0.5)}{\min(1.4,\ 1.5)} = \frac{0.4}{1.4} \approx 0.2857$$

Summing the similarities:

$$\sum_{i<j} s_{ij} = 0 + 1.0 + 0 + 0.4 + 1.0 + 0.2857 = 2.6857.$$

Since $\binom{N}{2} = 6$ *(i.e. the number of unique pairs among* $N = 4$ *individuals), the co-competence Greenberg Index is*:

$$GDI^{co}(\mathbf{W}) = 1 - \frac{2.6857}{6} \approx 1 - 0.4476 = 0.5524.$$

The resulting value of $GDI^{co} \approx 0.552$ reflects the average pairwise dissimilarity of language repertoires across the population. This means that, on average, a randomly selected pair of individuals shares just under half of the linguistic competence of the less linguistically equipped person. In this specific example, some individuals (e.g. individuals 1 and 3, or 2 and 4) share a substantial portion of their repertoires, while others (e.g. 1 and 2, or 1 and 4) have no linguistic overlap at all. The index captures this variation by computing the average interpersonal dissimilarity, where a value closer to 1 would indicate little shared competence across the population, and a value closer to 0 would reflect highly overlapping repertoires. For comparison, the classical GDI for this population is exactly 0.5, based on the distribution of native languages alone. The fact that GDI^{co} is slightly higher than the classical value suggests that relational diversity (as captured by repertoire mismatch) is greater than what is visible from dominant language categories alone. The value of 0.552 therefore indicates a moderate level of linguistic diversity from a relational perspective: some communicative bridges exist, but linguistic repertoires are far from fully shared.

This formulation captures linguistic diversity through interpersonal variation in competence. While the classical GDI operates on group proportions

and the competence-weighted GDI works on aggregate totals, the co-competence index emphasizes how similar or dissimilar speakers' repertoires are. In multilingual populations, it offers a social-relational perspective on diversity, where shared competences reduce diversity and distinct repertoires increase it. This perspective echoes the 'Swiss communication model' proposed by Grin et al. (2015), where the probability of successful communication across groups is modelled based on shared receptive competences (B2 level or higher). Rather than focusing on individual group sizes or average linguistic knowledge, the model estimates the likelihood that randomly paired individuals from different linguistic communities can understand one another. The resulting probabilities (e.g. for Germanophone – Francophone – Italophone trios) depend on the overlap between their repertoires. This operationalization of communicative potential aligns with the co-competence logic of the relational GDI: diversity is not merely a matter of aggregate counts, but of how repertoires intersect or diverge across individuals.

This approach resonates with recent developments in the measurement of multilingual communication. Gazzola et al. (2020) propose a family of indices to evaluate the probability that individuals with heterogeneous repertoires can communicate under different models: via a shared common language, via mixed active-receptive skills (polyglottism), or via receptive multilingualism. Their work similarly moves beyond categorical group membership and emphasizes the probabilistic nature of understanding in multilingual settings. However, while the co-competence matrix developed here shares conceptual ground with the probabilistic models proposed by Gazzola et al. (2020), the two approaches differ in structure and interpretation. The co-competence matrix computes a continuous similarity score between each pair of individuals, based on how much of the weaker repertoire is covered by the stronger, and aggregates these scores into a population-level index of relational diversity. By contrast, Gazzola et al. (2020) define several binary communication models (e.g. shared language, polyglottism, receptive multilingualism) and derive the probability that two randomly selected individuals can communicate under each model. Their framework focuses on expected communicative success, while the co-competence index emphasizes interpersonal repertoire overlap and relative inclusion. Both approaches highlight the importance of multilingual competence in enabling communication, but the co-competence matrix provides a interaction-sensitive view of diversity that can be applied to micro-level competence data.

3.2.4. Comparison

Compared to the classical GDI, which assumes that each individual belongs to a single linguistic group, the two extensions discussed so far allow for

individual multilingualism and offer more nuanced representations of diversity. The competence-weighted GDI aggregates the strength of language repertoires across individuals and normalizes the totals to yield a population-level probability distribution. It captures the overall linguistic capital of a population, accounting for both dominant and partial competences. By contrast, the co-competence formulation (GDI^{co}) takes an interactional perspective. It measures the average similarity between individuals' language profiles, based on how much of the weaker repertoire is shared by the other. This approach is grounded in communicative potential: a high similarity score indicates that two individuals are likely able to understand each other, even if their repertoires are not identical. It thus captures not only which languages are present in a population, but also how they are distributed and intersect socially.

Both extensions improve on the classical GDI by incorporating multilingual repertoires, but they emphasize different aspects:

- The **competence-weighted** approach is computationally straightforward and suitable when only speaker-level data is available. It generalizes the GDI while maintaining a familiar population-based structure.
- The **co-competence matrix** approach emphasizes relational diversity, that is, how repertoires overlap between individuals. It is particularly informative in contexts where communication, contact, or language mediation are of central interest.

While both extensions of the GDI account for individual multilingualism, they differ in how they interpret it. The competence-weighted GDI captures diversity at the population level by pooling individual language competences, but treats these contributions as additive and independent. In contrast, the co-competence formulation evaluates how competences are distributed across individuals, placing emphasis on shared repertoires and interactional potential. As a result, GDIco tends to yield higher diversity scores when multilingualism is widespread but unevenly distributed, i.e. when individuals know different combinations of languages, and mutual understanding is not guaranteed. For example, in a population where all individuals are multilingual but speak non-overlapping sets of languages, the competence-weighted GDI may suggest low concentration, whereas GDI^{co} will highlight the lack of shared repertoires. This distinction makes the co-competence index particularly well suited to analysing relational linguistic diversity in socially complex or fragmented multilingual settings. Consider for example a population of four individuals and three languages (A, B, C). Two individuals (1 and 2) are bilingual in A and B; the other two (3 and 4) are bilingual in A and C. All individuals share competence in language A, while B and C are each known by half the population. At the aggregate level, the competence-weighted GDI is relatively high

($GDI^{comp} = 0.625$), reflecting balanced linguistic diversity in the population as a whole. At the interpersonal level, however, there is substantial overlap in repertoires: everyone speaks A, and each pair shares at least one language. The co-competence GDI is therefore lower ($GDI^{co} = 0.333$), indicating high average similarity and strong communicative potential within the population. Although individuals differ in their secondary competences, the shared core of language A ensures cohesion, which the co-competence index captures.

This example underscores the value of using both indices in tandem: while GDIcomp reflects how much linguistic knowledge is present in a population, GDIco reveals how that knowledge is distributed across individuals. Depending on the analytical goal, whether assessing aggregate capacity or relational fragmentation, each index offers complementary insights.

3.3. Incorporating Linguistic Distance

A further refinement in the measurement of linguistic diversity involves incorporating the degree of structural or typological similarity between languages. The standard GDI treats all languages as categorically distinct, implicitly assigning maximal dissimilarity to any pair of languages that are not identical. However, this assumption may overstate diversity in contexts where many languages are closely related or mutually intelligible.

This concern was already addressed in Greenberg's original framework: he proposed a second measure, Index B, which accounts for resemblance between language pairs (Greenberg, 1956). In his formulation:

$$B = 1 - \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \cdot r_{ij}, \tag{8}$$

where p_i and p_j denote the proportions of speakers of languages i and j, and $r_{ij} \in [0, 1]$ is a resemblance coefficient, with $r_{ii} = 1$ by convention. The more similar two languages are, the larger r_{ij}, and the lower the overall index.

This framework can be naturally reinterpreted in terms of *linguistic distance* rather than similarity. Let $\delta_{ij} \in [0, 1]$ represent the distance between languages i and j, with $\delta_{ii} = 0$. Then the resemblance term becomes $r_{ij} = 1 - \delta_{ij}$, and the diversity index becomes:

Definition 4 (Distance-Weighted Greenberg Index). *Let* $\mathbf{p} = (p_1, \ldots, p_n) \in \mathbb{R}^{\ltimes}$ *be a normalized distribution over languages, and let* $\delta \in [0, 1]^{n \times n}$ *be a symmetric matrix of pairwise linguistic distances with* $\delta_{ii} = 0$. *Then the distance-weighted Greenberg Index is*:

$$GDI^{\text{dist}} = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \cdot \delta_{ij}. \tag{9}$$

This formulation generalizes the classical GDI by incorporating the magnitude of differences between languages. The index captures not only how frequently different languages are spoken, but also how *dissimilar* those languages are from one another. If all languages are maximally distant (i.e. $\delta_{ij} = 1$ for $i \neq j$, and $\delta_{ii} = 0$), then the index collapses to the classical GDI of Equation (2).[8]

Example 4. *Consider a population where three languages are spoken as L1 with proportions*:

$$\mathbf{p} = (0.5,\ 0.3,\ 0.2).$$

Suppose their linguistic distance matrix is:

$$\delta = \begin{bmatrix} 0 & 0.3 & 0.9 \\ 0.3 & 0 & 0.8 \\ 0.9 & 0.8 & 0 \end{bmatrix}$$

We compute the distance-weighted Greenberg Index as:

$$GDI^{\text{dist}} = \sum_{i=1}^{3} \sum_{j=1}^{3} p_i p_j \delta_{ij}$$

$$= 2 \cdot (p_1 p_2 \delta_{12} + p_1 p_3 \delta_{13} + p_2 p_3 \delta_{23})$$

$$= 2 \cdot (0.5 \cdot 0.3 \cdot 0.3 + 0.5 \cdot 0.2 \cdot 0.9 + 0.3 \cdot 0.2 \cdot 0.8)$$

$$= 2 \cdot (0.045 + 0.09 + 0.048) = 2 \cdot 0.183 = 0.366.$$

This value reflects both the distribution of speakers and the structural dissimilarity of the languages they speak. For comparison, the classical Greenberg Index with the same language shares is:

$$GDI = 1 - \sum_{i=1}^{3} p_i^2 = 1 - (0.5^2 + 0.3^2 + 0.2^2) = 1 - (0.25 + 0.09 + 0.04)$$
$$= 0.62.$$

Here, $\text{GDI}^{\text{dist}} < \text{GDI}$, reflecting the fact that while the speaker distribution is relatively even, some of the languages are structurally close (e.g. $\delta_{12} = 0.3$). The diversity observed may therefore be less substantial than what a purely categorical measure would suggest.

3.3.1. Distance Data Sources

The main challenge in implementing distance-weighted diversity measures lies in defining a reliable and justifiable distance matrix. Two prominent resources support this task.

The *Automated Similarity Judgment Program (ASJP)*[9] (Wichmann et al., 2022) provides distance scores based on phonetic similarity of short Swadesh vocabulary lists across thousands of languages. ASJP offers broad coverage and replicability, but focuses narrowly on phonetic-lexical similarity and may not reflect deeper grammatical or typological relationships.

In contrast, the *URIEL* (Littell et al., 2017) and *URIEL+* (Khan et al., 2025) databases provide multidimensional, vector-based representations of languages across features such as phonology, morphology, syntax, typology, and language inventories. The associated *lang2vec* framework enables users to compute pairwise distances tailored to specific linguistic domains or composite representations. The updated URIEL+ resource improves data quality, expands coverage (particularly for underdocumented languages), and allows for flexible weighting and dimensional selection.

Depending on the analytical goal (e.g. capturing genealogical relatedness, phonological proximity, or typological divergence) researchers can choose between ASJP, URIEL+, or a hybrid approach to construct the distance matrix δ. While distance-based extensions to GDI offer greater realism and cross-linguistic sensitivity, they rely heavily on the availability and validity of underlying similarity data.

3.4. Entropy-Based Diversity Index

Another well-established approach to measuring linguistic diversity draws on Shannon entropy, a foundational concept in information theory. Entropy quantifies the uncertainty or unpredictability of a random draw from a distribution and can be interpreted, in this context, as the heterogeneity of language use in a population.

Definition 5 (Shannon Entropy). *Let* $\mathbf{p} = (p_1, \ldots, p_n) \in \Delta_n$ *be a probability distribution over n languages, where each* p_i *denotes the proportion of speakers of language i. Then the Shannon entropy of* $\mathbf{p}$ *is defined as*:

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i. \tag{10}$$

This metric captures how evenly linguistic resources are distributed across categories. A high entropy indicates a more balanced (less predictable)

distribution, while a low entropy reflects dominance by one or a few languages.

3.4.1. Properties

- $H = 0$ when the population is monolingual (i.e. $p_i = 1$ for some i, and $p_j = 0$ for all $j \neq i$);
- H is maximized when $p_i = \frac{1}{n}$ for all i, yielding $H_{\max} = \ln n$;[10]
- H increases with both the number of languages and the evenness of their distribution.

3.4.2. Normalized Entropy

To facilitate comparisons across populations of different sizes (i.e. different n), entropy is often normalized by its theoretical maximum:

$$\hat{H} = \frac{H}{\ln n}. \tag{11}$$

This normalized entropy $\hat{H} \in [0, 1]$ allows for direct interpretation on a common scale:

- $\hat{H} = 0$ indicates perfect concentration (e.g. full dominance by a single language);
- $\hat{H} = 1$ indicates maximal diversity (i.e. a uniform language distribution).

Entropy-based metrics are particularly valuable in settings where the focus is on probabilistic uncertainty or the distributional balance of languages, such as modelling exposure in multilingual environments or assessing the risk of language shift. In this context, entropy can serve as an indicator of linguistic vulnerability: low entropy values signal strong concentration in one language, which may foreshadow the erosion of minority languages, while high entropy suggests a more balanced distribution that supports the coexistence of multiple languages. Unlike the standard GDI, which captures the expected pairwise dissimilarity between individuals, entropy quantifies the overall unpredictability of language identity in the population. In other words, while GDI reflects how likely it is that two people speak different languages, entropy reflects how surprising or uncertain a random individual's language affiliation would be.

Example 5. Consider two language distributions over three languages:

- Case 1 (balanced): $\mathbf{p} = (0.33,\ 0.33,\ 0.34)$
- Case 2 (skewed): $\mathbf{q} = (0.90,\ 0.05,\ 0.05)$

We compute both GDI and normalized Shannon entropy for each case.

Case 1: Balanced distribution

$$\text{GDI} = 1 - (0.33^2 + 0.33^2 + 0.34^2) \approx 1 - 0.3334 = 0.6666,$$

$$H = -(0.33 \ln 0.33 + 0.33 \ln 0.33 + 0.34 \ln 0.34) \approx 1.0984,$$

$$\hat{H} = \frac{H}{\ln 3} \approx \frac{1.0984}{1.0986} \approx 0.9998.$$

Case 2: Skewed distribution

$$\text{GDI} = 1 - (0.90^2 + 0.05^2 + 0.05^2) = 1 - (0.81 + 0.0025 + 0.0025) = 0.185,$$

$$H = -(0.90 \ln 0.90 + 0.05 \ln 0.05 + 0.05 \ln 0.05) \approx 0.3945,$$

$$\hat{H} = \frac{H}{\ln 3} \approx \frac{0.3945}{1.0986} \approx 0.359.$$

These results illustrate a key difference between the two metrics. GDI drops sharply when most individuals speak the same language, because the probability that two randomly selected individuals speak different languages becomes small. Entropy also decreases, but remains comparatively higher when rare languages are still present, because it captures the unpredictability of language identity. As such, entropy-based indices are often preferred when linguistic balance or exposure to minority languages is of analytic interest.

Recent work by Gullifer and Titone (2020) adapts entropy to characterize *individual-level* variation in bilingual and multilingual usage across social contexts. Their notion of *language entropy* captures how integrated or compartmentalized language use is within domains like home, work, or social life. For example, an individual using multiple languages in balanced proportions at work would have a higher entropy than one who uses only one language in that context. Crucially, they show that language entropy is predictive of L2 proficiency and accentedness, beyond classic measures like age of acquisition or overall exposure. While the focus here remains on population-level linguistic diversity, the underlying logic is shared: entropy is a robust tool for capturing the balance, spread, and unpredictability of language usage, whether aggregated across speakers or disaggregated across contexts.

While the classical GDI offers an intuitive interpretation rooted in interpersonal dissimilarity (namely, the probability that two randomly selected individuals speak different languages), Shannon entropy provides a complementary perspective by quantifying the unpredictability of the language distribution. One may prefer entropy when the goal is to capture

the richness and evenness of linguistic diversity in probabilistic terms, especially in settings where individual language affiliations are fuzzy or overlapping. Unlike GDI, which is quadratic and sensitive primarily to the concentration of dominant groups, entropy increases more sharply with the presence of smaller linguistic groups, giving more weight to rare languages. This makes entropy particularly well suited to multilingual contexts in which the uncertainty or variability of language choice itself is of analytical interest.

4. Comparing Diversity Indices

The extensions of the Greenberg Diversity Index proposed in this paper respond to different conceptual and empirical challenges in measuring linguistic diversity. Each formulation adds specific informational dimensions, such as graded multilingual competences, interpersonal repertoire overlap, or structural similarities between languages, to refine how diversity is captured in multilingual contexts. The classical GDI, while simple and interpretable, assumes mutually exclusive language groups. The weighted version incorporates partial competences, making it suitable for aggregate representations of multilingual populations. The co-competence index shifts focus to the dissimilarity between individual repertoires, offering a relational view of communicative potential. The distance-weighted GDI adjusts for inter-language similarity, preventing inflation of diversity scores when related languages dominate, and corresponds to Greenberg's Index B (Greenberg, 1956). Finally, the entropy-based index provides a probabilistic measure of unpredictability in language affiliation, particularly sensitive to richness and balance in the distribution. Table 1 summarizes their main properties. Their comparative use can yield complementary insights, adapted to different data sources and analytical goals.

The choice of diversity index should reflect both the research goal and the granularity of available data. Rather than selecting a single index in isolation, researchers may benefit from reporting multiple complementary measures. Together, they offer a more comprehensive and interpretable portrait of linguistic diversity, one that accounts not only for who speaks what, but also for how people communicate and how languages relate. In addition, competence- and distance-based approaches can be combined into hybrid indices that incorporate both the distribution of multilingual competences and the structural similarity between languages. Such indices allow for more fine-grained modelling of communicative potential in multilingual populations, especially in contexts where related languages coexist and mutual understanding may depend on partial overlap.

Table 1. Comparison of diversity indices according to key analytical features.

Index	Captures multilingualism?	Considers language similarity?	Data requirements	Best suited for
Classical GDI	No	No	Speaker proportions	Simple comparisons across monolingual groups
Weighted GDI	Yes	No	Individual-level language weights	Aggregate multilingual settings
Co-Competence GDI	Yes	No	Individual-level repertoires	Interaction-based analyses, communicative potential
Distance-weighted GDI	Yes (if weighted)	Yes	Distribution and distance matrix	Typologically-aware or structural diversity comparisons
Entropy	Yes	No	Weighted distributions	Richness-sensitive contexts

5. Conclusions and Future Research

The theoretical structure and potential extensions of the Greenberg Diversity Index, one of the simplest and most widely used measures of linguistic diversity, have been examined. After deriving the upper and lower bounds of the GDI as a function of the number of languages and their distribution, several extensions were proposed to capture more complex sociolinguistic configurations, particularly those involving individual multilingualism and interlinguistic similarity. Concrete examples demonstrated that these variants often yield different numerical values and interpretations, especially in multilingual or typologically dense contexts. Each index reflects distinct theoretical and empirical priorities, such as classification accuracy, social overlap, richness, or structural contrast. Rather than privileging a single metric, a comparative approach is recommended, whereby multiple indices are applied in tandem to offer complementary perspectives on linguistic diversity.

These theoretical insights provide not only normative benchmarks and analytical clarity, but also practical tools for researchers and policymakers engaged in the study and management of multilingualism. A promising direction for future research lies in the development of empirical applications grounded in the formal indices introduced in this paper. Such applications may include the use of survey microdata, administrative records, or network-based models of language competence to estimate diversity in concrete settings. In particular, network-based approaches offer a way to

account for the structure of communicative interactions within a population, capturing not only who speaks what, but also who can understand whom. By modelling linguistic repertoires as nodes and communicative compatibility as links, such frameworks can reveal patterns of cohesion, fragmentation, or bridging that remain invisible to aggregate indices.

While the formulation of the indices here assumes individual-level linguistic profiles, their application in practice may rely on more tractable approximations. One practical approach is to partition the population into groups with similar language repertoires, such as census categories, educational profiles, or regional clusters, and assign weights to these groups instead of individuals. This facilitates implementation in large-scale studies, particularly when individual-level data are unavailable or infeasible to collect, while still capturing relevant patterns of linguistic composition and inequality.

In multilingual societies, where language policy, education, and institutional communication often rely on accurate assessments of linguistic composition, these metrics may serve as valuable instruments for monitoring, evaluation, and planning. Further work could also focus on implementing the proposed indices in open-source computational frameworks to support replicable and scalable analyses.

Notes

1. This does not imply that multilingualism is absent from the population, but rather that it is not captured in the classical formulation of the index.
2. The assumption that languages form discrete, mutually exclusive categories can be sociolinguistically problematic. As the case of Bosnia and Herzegovina illustrates, distinctions between languages may reflect political boundaries more than linguistic distance. In such settings, raw categorical approaches may misrepresent the actual structure of linguistic repertoires and communicative realities.
3. Δ_n denotes the n-dimensional probability simplex: the set of all vectors $\mathbf{p} = (p_1, \ldots, p_n) \in \mathbb{R}^\aleph$ such that $p_i \geq 0$ for all i and $\sum_{i=1}^{n} p_i = 1$. Intuitively, it represents all possible ways to distribute one unit of probability (or population share) across n categories (in this case, languages). For example, in Δ_3, the simplex is a triangle in 3D space whose interior contains all valid language distributions over three languages.
4. A related idea appears in the work of Sankoff and Laberge (1978), who constructed an 'index of participation in the linguistic market' based on speakers' roles in social interaction and their exposure to legitimated language forms. While grounded in sociolinguistics and stratification theory, their approach also quantifies communication potential across a population.
5. Consider, for example, three individuals that report the following competences over two languages: person A speaks only language 1, person B only language 2, and person C speaks both equally. Their competence profiles can be written as vectors: $(1, 0)$, $(0, 1)$, and $(1, 1)$, where the first and second components

represent competence in language 1 and language 2, respectively. Summing these vectors yields $(1+0+1, 0+1+1) = (2, 2)$, which means language 1 has a total competence of 2, and language 2 also has 2. But since the population only includes 3 individuals, the total weight is 4 (and not 3) because the third individual contributes two units. To form a valid distribution, this vector must be normalized to sum to 1.

6. It should be noted that, while competence provides a more flexible and communicatively relevant foundation for modelling diversity, it also introduces empirical challenges. Unlike frequency of use, which can be observed or reported with reference to specific contexts, competence is often self-assessed or inferred, and may vary by domain (e.g. literacy vs. oral fluency). Nevertheless, it remains a key dimension for understanding linguistic capacity, especially in policy-relevant domains such as education, administration, or public service delivery.
7. Alternatively, and in light of the observations made earlier, each vector entry can be defined to represent the *relative frequency of language use*, such that the weights sum to one: $\sum_i w_{ki} = 1$; this reflects the proportion of time an individual uses each language across a given period (e.g. a day, a week, or a year), and is thus independent of the overall time frame.
8. For a proof, see Appendix B.
9. https://asjp.clld.org/https://asjp.clld.org/
10. For a proof, see Cover and Thomas (2006, p. 29)

Acknowledgments

The author wishes to thank Prof. François Grin (University of Geneva) for his insightful comments and suggestions on earlier versions of this work. Special thanks are also due to Mr Antonio Au-Yeung for his careful review of the mathematical notation and formalism. The author is also grateful to the two anonymous reviewers for their constructive feedback, which helped improve the clarity and precision of the manuscript. Their contributions have been greatly appreciated.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Swiss National Science Foundation under Grant 226064 (Project GLAD).

ORCID

Marco Civico http://orcid.org/0000-0001-8486-118X

References

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley-Interscience.

Desmet, K., Ortuño-Ortín, I., & Wacziarg, R. (2012). The political economy of linguistic cleavages. *Journal of Development Economics*, *97*(2), 322–338. https://doi.org/10.1016/j.jdeveco.2011.02.003

Gazzola, M., Templin, T., & McEntee-Atalianis, L. J. (2020). Measuring diversity in multilingual communication. *Social Indicators Research*, *147*(2), 545–566. https://doi.org/10.1007/s11205-019-02161-5

Greenberg, J. H. (1956). The measurement of linguistic diversity. *Language*, *32*(1), 109–115. https://doi.org/10.2307/410659

Grin, F., Amos, J., Faniko, K., Fürst, G., Lurin, J., & Schwob, I. (2015). *Suisse-Société multiculturelle. Ce qu'en font les jeunes aujourd'hui*. Edition Rüegger.

Grin, F., & Fürst, G. (2022). Measuring linguistic diversity: A multi-level metric. *Social Indicators Research*, *164*(2), 601–621. https://doi.org/10.1007/s11205-022-02934-5

Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition*, *23*(2), 283–294. https://doi.org/10.1017/S1366728919000026

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, *88*(10), 2427–2439. https://doi.org/10.1890/06-1736.1

Khan, A., Shipton, M., Anugraha, D., Duan, K., Hoang, P. H., Khiu, E., Doğruöz, A. S., & Lee, E.-S. A. (2025). Uriel+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 6937–6952). Abu Dhabi, UAE (Association for Computational Linguistics). https://aclanthology.org/2025.coling-main.463/

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., & Levin, L. (2017). Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Lapata, M. Blunsom, P. and Koller, A. (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 8–14). Association for Computational Linguistics, Valencia, Spain. https://aclanthology.org/E17-2002.pdf

Sankoff, D., & Laberge, S. (1978). The linguistic market and the statistical explanation of variability. In D. Sankoff (Ed.), *Linguistic variation: Models and methods* (pp. 239–250). Academic Press.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*(4148), 688. https://doi.org/10.1038/163688a0

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707–719. https://doi.org/10.1098/rsif.2007.0213

Wichmann, S., Holman, E. W., & Brown, C. H. (2022). The ASJP database (version 20). Retrieved July, 2025, from https://asjp.clld.org/

Appendix

Appendix A. Proof of Proposition 1

We now show that the bounds of the Greenberg Diversity Index stated in Equation (3) hold for any probability distribution $\mathbf{p} \in \Delta_n$.

From Equation (2), the Greenberg Diversity Index can be written as

$$\mathrm{GDI}(\mathbf{p}) = 1 - \sum_{i=1}^{n} p_i^2.$$

Lower bound: Since $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$, we have $\sum_{i=1}^{n} p_i^2 \leq 1$. Therefore,

$$\mathrm{GDI}(\mathbf{p}) = 1 - \sum_{i=1}^{n} p_i^2 \geq 0,$$

with equality if and only if $p_j = 1$ for some j, and $p_i = 0$ for all $i \neq j$, that is, when the entire population shares the same language.

Upper bound: To maximize $\mathrm{GDI}(\mathbf{p})$, we minimize $\sum_{i=1}^{n} p_i^2$ subject to $\sum_{i=1}^{n} p_i = 1$, which is a case of constrained optimization. Hence, we can use the method of Lagrange multipliers. Define the Lagrangian:

$$\mathcal{L}(p_1, \ldots, p_n, \lambda) = \sum_{i=1}^{n} p_i^2 - \lambda \left(\sum_{i=1}^{n} p_i - 1 \right).$$

Taking partial derivatives:

$$\frac{\partial \mathcal{L}}{\partial p_i} = 2p_i - \lambda \overset{F.O.C.}{=} 0 \quad \Rightarrow \quad p_i = \frac{\lambda}{2} \quad \text{for all } i,$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{i=1}^{n} p_i - 1 \overset{F.O.C.}{=} 0.$$

Substituting into the constraint yields:

$$n \cdot \frac{\lambda}{2} = 1 \quad \Rightarrow \quad \lambda = \frac{2}{n},$$

$$\Rightarrow \quad p_i = \frac{1}{n} \quad \text{for all } i.$$

At this point:

$$\sum_{i=1}^{n} p_i^2 = n \cdot \left(\frac{1}{n} \right)^2 = \frac{1}{n}, \quad \Rightarrow \quad \mathrm{GDI}(\mathbf{p}) = 1 - \frac{1}{n}.$$

Hence, for any $\mathbf{p} \in \Delta_n$, the GDI satisfies:

0

$$0 \leq \mathrm{GDI}(\mathbf{p}) \leq 1 - \frac{1}{n},$$

with the bounds attained respectively when the distribution is fully concentrated or uniform.

Appendix B. Equivalence of Distance-Weighted and Classical GDI

In this appendix, we show that the distance-weighted formulation of the Greenberg Diversity Index (GDI) reduces to the classical form when all distinct languages are treated as maximally dissimilar.

Let $p_i \in [0, 1]$ denote the proportion of speakers of language i in the population, with $\sum_{i=1}^{n} p_i = 1$, and let $\delta_{ij} \in [0, 1]$ denote the linguistic distance between languages i and j. The distance-weighted GDI is defined as:

$$\mathrm{GDI}^{\mathrm{dist}} = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \cdot \delta_{ij}$$

Now assume that:

$$\delta_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

We split the double sum into diagonal and off-diagonal terms:

$$\mathrm{GDI}^{\mathrm{dist}} = \sum_{i=1}^{n} p_i^2 \cdot \delta_{ii} + \sum_{i \neq j} p_i p_j \cdot \delta_{ij}$$

Since $\delta_{ii} = 0$ and $\delta_{ij} = 1$ for $i \neq j$, this simplifies to:

$$\mathrm{GDI}^{\mathrm{dist}} = \sum_{i \neq j} p_i p_j$$

We now show that this expression is equivalent to:

$$\mathrm{GDI} = 1 - \sum_{i=1}^{n} p_i^2$$

To do so, we use the identity for squaring a sum of terms:

$$\left(\sum_{i=1}^{n} p_i \right)^2 = \sum_{i=1}^{n} p_i^2 + \sum_{i \neq j} p_i p_j$$

Since $\sum_i p_i = 1$, this yields:

$$1 = \sum_{i=1}^{n} p_i^2 + \sum_{i \neq j} p_i p_j \quad \Rightarrow \quad \sum_{i \neq j} p_i p_j = 1 - \sum_{i=1}^{n} p_i^2$$

Therefore:

$$\mathrm{GDI}^{\mathrm{dist}} = \sum_{i \neq j} p_i p_j = 1 - \sum_{i=1}^{n} p_i^2 = \mathrm{GDI}$$

This result shows that the classical GDI is a special case of the distance-weighted GDI in which all language pairs are treated as equally and maximally dissimilar. This reinforces the interpretation of the classical GDI as measuring categorical linguistic difference.