universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Global Linguistic Diversity
Incorporating Similarity Measures Using Large-Scale Cross-Linguistic Databases

verfasst von | submitted by

Hannes Nils Hampus Essfors BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von | Supervisor:	Ass.-Prof. Mag. Mag. Dr. Andreas Baumann

Acknowledgements

Writing a thesis is a time-consuming effort requiring both discipline and patience, not only from the author, but also from those around him. I would therefore like to extend my deepest gratitude to first and foremost my partner, Ida, whose unwavering support allowed me to undertake this effort.

I would also like to acknowledge the importance of my supervisor, Andreas Baumann, who supported me in writing a thesis on linguistic diversity, a topic which I am deeply fascinated by. His input and encouraging thoughts on the direction of the thesis were invaluable. I would also like to thank him for getting me into the programming language R, which enabled much of the analysis part of the thesis.

Finally, I would like to extend my thanks to the rest of the DIGILINGDIV - project team: Hannes, Katharina, Julia, and Juliane, for their support. Having a supportive environment working on similar topics is invaluable for inspiration and motivation.

Abstracts

English

Linguistic diversity is under threat, with projections suggesting that over 20% of languages risk disappearing by the turn of the century. While ecology has developed statistical techniques to quantify diversity, linguists often resort to simple language counts. To further the methodology available to researchers, this thesis adapts the Leinster-Cobbold framework from ecology to create a unified notion of diversity on a global scale, taking richness, relative abundance, and similarity of languages into account.

Large-scale cross-linguistic databases (PHOIBLE, Grambank, ASJP) were assessed for their suitability in modeling global linguistic diversity, using speaker data from Ethnologue and the Joshua Project as a baseline. Measures of linguistic similarity were derived and analyzed through cluster and correlation analysis.

The results indicate that only the ASJP database offers sufficient coverage to derive similarity measures on a global scale. While lexical and morphosyntactic similarity measures both display strong phylogenetic classificatory power, correlations between different similarity measures are weak, meaning no single metric can serve as a reliable proxy for the others. Applying the Leinster-Cobbold framework in a global context using lexical similarity demonstrates that accounting for similarity is impactful in regions with dialect continua but has a modest effect globally. The study concludes that incorporating similarity is essential to accurately model linguistic diversity.

Deutsch

Sprachdiversität ist bedroht; bis zum Ende dieses Jahrhunderts könnten 20% aller Sprachen ausgestorben sein. Während in der Ökologie statistische Methoden zur Messung von Biodiversität etabliert sind, verlassen sich Sprachwissenschaftler häufig auf einfache Sprachzählungen. Diese Arbeit adaptiert daher das Leinster-Cobbold-Rahmenwerk aus der Ökologie, um einen einheitlichen Diversitätsbegriff zu schaffen, der die Gesamtzahl, relative Häufigkeit und Ähnlichkeit von Sprachen berücksichtigt.

Große sprachübergreifende Datenbanken (PHOIBLE, Grambank, ASJP) wurden unter Verwendung von Sprecherzahlen aus Ethnologue und Joshua Project als Referenz untersucht und hinsichtlich ihrer Eignung zur Modellierung globaler Sprachdiversität bewertet. Maße sprachlicher Ähnlichkeit wurden abgeleitet und mittels Cluster- und Korrelationsanalysen ausgewertet.

Die Ergebnisse zeigen, dass lediglich die ASJP-Datenbank eine ausreichende globale Abdeckung bietet, um Ähnlichkeitsmaße abzuleiten. Während die Maße zuverlässig sprachliche Ähnlichkeiten modellieren, korrelieren die Maße nur schwach miteinander, weshalb kein Maß als verlässlicher Proxy für die anderen dienen kann. Durch die Anwendung des Leinster-Cobbold-Rahmenwerks wird aufgezeigt, dass das Berücksichtigen von Ähnlichkeiten in Ländern mit Dialektkontinua erhebliche Auswirkungen hat, während der globale Effekt gering bleibt. Daraus wird die Notwendigkeit geschlossen, sprachliche Ähnlichkeiten zu berücksichtigen, um Sprachdiversität zu modellieren.

Contents

1 Introduction

Much like biodiversity, linguistic diversity is under threat. Already in 2003, the UNESCO Ad Hoc Expert Group on Endangered Languages asserted through their report *Language Vitality and Endangerment* that

> "Language diversity is essential to the human heritage. Each and every language embodies the unique cultural wisdom of a people. The loss of any language is thus a loss for all humanity.
>
> Though approximately six thousand languages still exist, many are under threat. There is an imperative need for language documentation, new policy initiatives, and new materials to enhance the vitality of these languages.
>
> The cooperative efforts of language communities, language professionals, NGOs and governments will be indispensable in countering this threat. There is a pressing need to build support for language communities in their efforts to establish meaningful new roles for their endangered languages." (p.1)

This impending loss of diversity and urgent need for action was recently brought to attention through an influential publication in Nature by Bromham et al. (2022), predicting a global loss of over 1500 languages by the end of this century. To arrive at this conclusion, Bromham et al. used 51 predictor variables to construct a statistical model to predict changes in endangerment levels. Thereby, they show that it is possible and necessary to conduct statistically rigorous analysis to understand the notion of linguistic diversity. To do exactly this, linguistic diversity needs to be measured numerically, something that has a long-standing tradition in ecology but has not been given thorough scholarly attention in linguistics. While linguistics commonly default to counting the number of languages and equating it to diversity (Bouckaert et al., 2022; Harmon & Loh, 2010), as in Bromham et al. (2022), there are two other aspects of equal importance to measure diversity: relative abundance and similarity. This was made explicit in a recent paper on the topic by Grin and Fürst (2022), where they note that while linguistic distance is a legitimate component in diversity measures, it poses conceptual and methodological challenges. They highlight the lack of established approaches for operationalizing interlinguistic distance, and ultimately set aside the problem due to its complexity.

This leaves a gap that needs to be filled, and with the recent rise of large-scale cross-linguistic data such as WALS (Dryer & Haspelmath, 2013) and the CLLD (Forkel & Haspelmath, 2023), it could very well be that the operationalization of linguistic similarity on a global scale is feasible. And if that is the case, would adding similarity impact how diversity is perceived in a way that better reflects actual diversity? These are the overarching questions that this thesis aims to address, and as such, provide important knowledge and tools to quantitatively investigate linguistic diversity further.

1.1 Purpose

The purpose of the thesis is to investigate global linguistic diversity by incorporating measurements of linguistic similarity using large-scale cross-linguistic data.

1.2 Research Questions

The research questions of the thesis are the following:

1. What data is available to calculate linguistic similarity/distance on a global scale, and how well does such data cover languages, speakers, and countries?
2. How do linguistic distance/similarity measures based on different domains of language correlate in a global context? Could any measure be used as a proxy for the other?
3. How does adding linguistic distance/similarity to diversity measures impact how global linguistic diversity is perceived?

1.3 Outline of the Thesis

To address the research questions, Chapter 2 provides a theoretical background, describing the academic landscape in which the thesis situates itself. It introduces the concept of diversity from both an ecological and linguistic point of view. Chapter 3 presents the cross-linguistic data sources that form the foundation of the thesis, together with the methods used to derive similarity measures and analyze them together with diversity. In Chapter 4, the linguistic data are described and visualized, and the coverage of the typological databases is assessed and compared. In Chapter 5, the measures of similarity are analyzed and combined with linguistic data to estimate the global linguistic diversity and the impact of accounting for similarity. Chapter 6 concludes the thesis by summarizing the main findings and giving direction for future research.

2 Theoretical Background

Having formally stated the purpose and intention of the thesis, this chapter will now proceed to describe the background and relevant literature concerning diversity. This will give the necessary understanding of where the thesis fits into the scientific landscape concerning linguistic diversity and similarity, and how it can bridge both aspects.

2.1 What is Diversity?

Certainly, the importance of diversity in Western culture has not eluded anyone, with diversity initiatives being widespread throughout organizational bodies (Leslie, Kim, & Ye, 2025). According to Leslie et al., the purpose of such actions is to support the inclusion of individuals belonging to marginalized social groups characterized by traits such as ethnicity and gender. As a result, we have a general idea of diversity associated with variability and differences, such that more variability and differences equal more diversity. This is concluded by the general definition given by Cambridge University Press (n.d.), where diversity is defined as

> "the fact of many different types of things or people being included in something; a range of different things or people:" (Cambridge University Press, n.d.)

As such, the things being included shape the notion of diversity, and depending on the field of study, these *things* will be field-specific. For example, in the context of ecology, the thing will commonly be different species, creating the notion of biodiversity as "the variety and variability of biological organisms" (National Research Council (US) Committee on Noneconomic and Economic Value of Biodiversity, 1999), and the *something* - in this thesis referred to as *setting* - will most commonly be a certain ecosystem. For the topic of this thesis, linguistics, the *thing* will instead be the notion of speech, summarized as a language, and the setting could be anywhere language is spoken: a country, a city, a school, etc. As such, diversity can be generalized to any science and objects of study where the variability of objects is of interest. From this general definition, it follows that the concept of diversity requires a few foundational definitions to be formalized. First and foremost, the primary object of interest, and a setting in which the object of interest can be present. Then, a notion of what makes two objects different is necessary, such that the objects can be categorized as distinct types. And finally, a dimension of attributes inherent to all objects, according to which differences between types of objects can be observed. This follows from our conceptual understanding of diversity, and once all of these properties are defined, they can be evaluated, allowing for an empirical approach and quantification of diversity (Mari, Wilson, & Maul, 2023). In the following sections, we shall see how this has been approached in both ecology and linguistics.

2.2 Diversity in Ecology

2.2.1 Richness

In ecology, measuring diversity has a longstanding tradition, rooted in both the empirical traditions of the natural sciences and the fundamental interdependence of species within ecosystems, making biodiversity crucial for the functioning of ecosystems (Hooper et al., 2005). The fundamental measurement used is a *diversity index*, a statistic summarizing the biological variability in a given setting into a single number, allowing the comparison of multiple settings. The oldest and simplest such measure, probably first implemented by Darwin in 1855 (Magurran, 2021), assumes an accepted taxonomy for grouping biological organisms and then counts the co-occurrence of species. This gives the key measure of richness (R), which simply states the number of unique species found in a setting, or in general terms, the number of different types of objects. This is a legitimate component of diversity, since R restricts how diverse a setting can be. If R = 1, there is only one type of object; hence, the *range of different things* is inherently low. If R is large, the range of different things is greater, implying more possible diversity. This measure is simplistic, however, since it does not take into account the distributional properties of the different objects in the setting; imagine a setting *A* containing 99 animals distributed on the species cat, dog, and mice, and assume equal inter- and intraspecific similarity. If there are 33 cats, 33 dogs, and 33 mice, then R = 3. If the distribution changes, giving setting *B*, such that there are 90 cats, 5 dogs, and 4 mice, then still, R = 3, implying that A and B are equally diverse. However, the distributional properties of A and B are different, and reasonably, one would argue that A is less diverse than B since A is dominated by cats. This means that to accurately measure diversity, one has to not only consider the number of present species, but also their relative abundance (Magurran, 2021), which can be generalized to any type of objects with an established taxonomy, such as languages.

2.2.2 Relative Abundance

In the ecology literature, a rich tradition of usage surrounds two specific diversity indices that take relative abundance of species and have been equated with diversity, namely the Gini-Simpson index and Shannon entropy (Jost, 2018). The Gini-Simpson index (λ) was developed by Simpson (1949) as a measure of diveristy and is formally defined as

$$\lambda = 1 - \sum_{i=1}^{S} p_i^2 \tag{2.1}$$

where p_i is the relative abundance of the *i*th species expressed as a fraction in a setting with S different species, such that R = S. As such, p_i can be considered the probability of randomly sampling an individual belonging to the *ith* species, and assuming that the number of individuals is so large that sampling with replacement approximates sampling without replacement, then p_i^2 is the joint probability of randomly drawing two animals belonging to the same species. The complement, which equals the Gini-Simpson index, is thus the probability of randomly sampling two individuals from different species. Jost suggests the biological interpretation as

the probability of two animals of different species belonging to an ecosystem randomly walking around meeting each other. If all species are equally abundant, the probability of an interspecific encounter is maximized for a given richness. As richness increases while the distribution of individuals across species remains constant, the index approaches 1. If the richness remains constant, but the distribution changes such that one species dominates, the index approaches 0. As such, the properties of the measure are coherent with our general understanding of diversity, albeit with a probabilistic interpretation.

Similarly, the Shannon entropy (H') possesses similar mathematical properties to the Gini-Simpson index, but differs in its interpretation and its lower sensitivity to the presence of rare species (Chao, Chiu, & Jost, 2014). The measure is simply the entropy measure developed by Shannon (1948) to quantify the uncertainty in predicting the next character in a string, formalized as

$$H' = -\sum_{i}^{S} log(p_i) * p_i \tag{2.2}$$

where again p_i is the relative abundance of the *i*th species and S is the richness of the setting. It follows that the biological interpretation is the uncertainty of which species an animal will belong to if sampled randomly (Jost, 2018). For a given richness, this uncertainty will be maximized if the distribution of animals is equal across species, and minimized if one species dominates. As such, this measure too behaves coherently with our understanding of diversity, but whereas the Gini-Simpson is a probability and restricted to values between 0 and 1, Shannon entropy is measured in bits and can take on potentially infinitely large values. Therefore, although both are measures of diversity, they are different measures with different interpretations and different mathematical properties. The question that arises is which of these measures of Diversity should be used if one is interested in assessing diversity generally? Commonly, researchers defaulted to using either one of the indices or reporting both. Furthermore, comparing different settings using these indices are problematic, as is made evident by two examples found in the sample chapter of a book written by Jost and Chao (n.d.):

> "Imagine a community with a million equally common species. Its Gini-Simpson index is 0.999999. A plague attacks this community and eliminates 99.99% of the species, leaving only a hundred species untouched. Ecologists and conservation biologists would consider this a huge drop in diversity, both in terms of the variety of interactions experienced by a constituent organism before and after the plague, and in terms of the conservation value of the pre- versus post-plague communities. The tools of diversity analysis should be able to clearly indicate the magnitude of this drop if they are to be useful in less dramatic situations. Yet the post-plague community's Gini-Simpson index is 0.99, only 1% less than the Gini-Simpson index of the pre-plague community. An ecologist equating the Gini-Simpson index with "diversity" would therefore conclude that the plague which killed almost all the species did not have a big effect on the community's "diversity"!" (p.2)

As such, while these measures agree with the general definition of diversity, their interpretations are not meaningful in a biological sense, and as such also not in the general and in

extension the linguistic sense.

2.2.3 The Hill numbers Framework

To resolve the issue stated above, a unified framework was introduced in a seminal paper by (Hill, 1973), relating richness, the Gini-Simpson index, and Shannon entropy to each other as measures of *effective number of species*. Jost often refers to this as *True Diversity*; a measure parametrized by the variable *q*, often referred to as the Hill number of order q, and is formally defined by Chao, Chiu, and Jost (2016) in their Equation (3a) as

$$^{q}D = (\sum_{i}^{S} p_i^q)^{\frac{1}{1-q}}, q \geq 0, q \neq 1 \tag{2.3}$$

where ^{q}D is the diversity of order q, p_i is the *ith* species and S the richness of the setting. For different values of q, this gives the number of equally abundant species in a hypothetical setting that would generate the same value as ^{q}D. This interpretation is motivated given that for all q for which ^{q}D is well-defined, ^{q}D = S, if all species are equally abundant. This can easily be shown by some algebraic manipulation: if a setting with richness S is equally distributed, it follows that $p_i = \frac{1}{S}$. Therefore

$$^{q}D = (\sum_{i}^{S} (\frac{1}{S})^q)^{\frac{1}{1-q}} = (\frac{1}{S^q} * S)^{\frac{1}{1-q}} = (\frac{1}{S^{(q-1)}})^{\frac{1}{1-q}} = (S^{(1-q)})^{\frac{1}{1-q}} = S$$

As such, if all species are not equally abundant, then $^{q}D < S$, and the more uneven the distribution is, the smaller ^{q}D gets. Thus, if the richness of a given setting *s1* is 5, then $max(^{q}D_{s1}) = 5$. If, however, the distribution across the species for the same setting is such that $^{q}D_{s1} = 4$, then the diversity of that setting is equivalent to that of a different setting *s2* with 4 equally abundant species such that $max(^{q}D_{s2}) = 4$; therefore, the diversity of $s1$ is *effectively* that of $s2$, and thus diversity in the Hill number framework is said to measure the *effective number of species*. This means that the unit of measure is species, making it intuitive in the same way as richness, but adjusted by taking relative abundance into account. At that point, the parameter q becomes a sensitivity parameter that can be selected to adjust how the diversity should react to the presence of dominating species. As such, for larger values of q, a given shift in concentration of relative abundance of dominant species is associated with a sharper drop in diversity compared to smaller values of q. Therefore, for q = 0, the diversity measure is insensitive to relative abundance. And indeed, setting q = 0 as an input to eq 2.3 gives

$$^{0}D = (\sum_{i}^{S} p_i^0)^{\frac{1}{1-0}} =^{q} D = (\sum_{i}^{S} 1)^1 = S$$

which is the richness. As such, richness is the diversity of order 0 in the Hill numbers framework. Selecting a larger value for q, such as q = 2 as input gives

$$^2D = (\sum_i^S p_i^2)^{\frac{1}{1-2}} = \frac{1}{\sum_i^S p_i^2}$$

a measure commonly referred to as Inverse Simpson Concentration and is the inverse of 1 - λ from Equation (2.1), allowing the Gini-Simpson index to be derived from the Hill number framework for q = 2. For q = 1 ^qD isn't defined but according to Chao et al. (2016), the limit as q appraoches 1 gives

$$^1D = \lim_{q \to 1}(^qD) = \exp(-\sum_i^n \log(p_i) * p_i) = \exp(H').$$

As such, the Shannon entropy is the natural logarithm of the Hill number of order 1. Therefore, the Hill number framework not only provides a measure that is coherent with our understanding of diversity, but unites existing measures in a single framework using the same units that are straightforward to interpret. Thus far, everything seems fine, but the framework makes one fundamental assumption that violates our understanding of Diversity, namely that all species, or languages in a linguistic context, are equally different. This is, however, not the case, and as Jost (2018) exemplifies:

> "A set of five rat species has the same diversity as a set consisting of one rat, one armadillo, one manatee, one pangolin, and one monkey, if the set of relative abundances are the same for both." (p.60)

to make this example analogous to linguistics, consider one set of 3 languages: Danish, Swedish, and Norwegian with 5 speakers of each. Consider another set of the 3 languages: Japanese, French, and Swahili with speakers 5 of each. Intuitively, the second set should be considered more diverse according to our understanding of diversity, since the languages included are more dissimilar than in the first set, but any measure derived from the Hill numbers framework will conclude that the two sets of species or languages are equally diverse.

A way of addressing this aspect of diversity is to incorporate measures of pairwise distance between the different type of objects for which diversity is measured, a fundamental contribution to which was made by Rao (1982) who adapted the Gini-Simpson index to account for both pairwise distances between species, e.g. evolutionary divergence time, giving Rao's Q, formally defined as

$$Q = \sum_{ij}^S d_{ij} p_i p_j \tag{2.4}$$

where $d_{j,i}$ is any distance between the *i*th and *j*th species with relative abundances denoted as p_i and p_j. Similar to the Gini-Simpson index, the relative abundances have probabilistic interpretations, and as such, the measure can be interpreted as the average distance between any two randomly selected species (Chao et al., 2016). However, since the Gini-Simpson index is a special case of Rao's Q, where all different species are considered to have a distance of 1 and otherwise 0, it suffers from the same setbacks that have been tackled by the Hill numbers framework.

The Leinster-Cobbold Framework

A complete unifying framework of diversity that incorporates both relative abundance and similarity was given in 2012 by Leinster and Cobbold, who take the Hill number framework of measuring diversity as the effective number of species parametrized by q as a starting point. In their framework, the diversity given by Equation (2.3) is a special case of diversity, referred to as *naive diversity*, since it assumes that all species are equally and completely dissimilar, thus becoming a special case of non-naive diversity where the similarity structure between the species are accounted for.

While the Leinster-Cobbold framework was developed in ecology, it generalizes as a measure of diversity for any objects and system, requiring two things: data on the relative abundance of the different types, and a measure of similarity between *all* types under consideration. Given data on the relative abundance of each species, pairwise similarity between all species, and a value of q signaling sensitivity to dominant species, the measure gives an effective number which is interpreted as the equivalent number of equally dissimilar species and equally abundant species of order q. Formally, diversity is thus defined as

$$ {}^{q}D^{Z}(p) = (\sum_{i}^{S} p_i (Zp)_i^{q-1})^{\frac{1}{1-q}}, q \geq 0, q \neq 1 \tag{2.5} $$

where

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

S is the total number of species, i.e., richness, p_i and p_j are the relative abundance of the i^{th} and j^{th} species in the relative abundance vector p, and Z_{ij} is the pairwise similarity between them. Z_{ij} is a measure of similarity such that $0 \leq Z_{ij} \leq 1$, where 1 indicates complete similarity and 0 complete dissimilarity. The similarity matrix Z is therefore the S x S matrix containing all pairwise similarities Z_{ij} and $(Zp)_i$ the average similarity of the i^{th} species to all other species, weighted by their relative abundances. It follows that Zp is a vector with each entry corresponding to a species, containing the expected similarity of that species to any randomly selected species. This means that if a species A is very similar to another species B that has a high relative abundance, then species A will contribute more to lowering the diversity than in the case where species B has a low relative abundance.

This is perhaps best understood through an example, and for the purpose of illustration, let's use a linguistic example, i.e., species are languages. Imagine a country where 3 languages are spoken: A by 30 %, B by 50 % and C by 20 % of the population. Assume that A and B are very similar with 80% similarity. For the sake of simplicity, assume that C is equally dissimilar to both A and B, with 20 % similarity. All languages are 100 % similar to themselves. For diversity calculation purposes, this linguistic setting can now be represented by the relative abundance vector $p = [0.3, 0.5, 0.2]$ and the similarity matrix

$$Z = [[1, 0.8, 0.2] \\ [0.8, 1, 0.2] \\ [0.2, 0.2, 1]]$$

The vector with expected similarities Zp is now easily calculated as $Z \cdot p\top$ such that

$$\begin{aligned} Zp = [&1 * 0.3 + 0.8 * 0.5 + 0.2 * 0.2, \\ &0.8 * 0.3 + 1 * 0.5 + 0.2 * 0.2, \\ &0.2 * 0.3 + 0.2 * 0.5 + 1 * 0.2] \\ &= [0.74; 0.78; 0.36] \end{aligned}$$

A value 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 will be calculated as

$${}^0D^Z(p) = (\sum_i^3 \frac{p_i}{(Z_p)_i}) = \frac{0.3}{0.74} + \frac{0.5}{0.78} + \frac{0.2}{0.36} = 1.6.$$

Thus, given the assumed relative abundance and similarity of the languages, the effective diversity is 1.6 - equivalent to a setting with 1.6 completely dissimilar and equally abundant languages. From the calculation above, it is also clear to see that as $(Zp)_i$ increases, the diversity decreases; and $(Zp)i$ is maximized by the i^{th} language being similar to highly abundant languages.

Moreover, Equation (2.5) relates to Equation (2.3) by the S x S identity matrix *I*. In the Hill numbers framework, all different species are assumed to be completely dissimilar, i.e., $Z_{ij} = 0$ for $i \neq j$, and the same species are assumed to be identical, i.e., $Z_{ij} = 1$ for $i = j$. Therefore, all Hill numbers are a special case of diversity where $Z = I$, i.e., the identity matrix.

All of this put together means that when applying the Leinster-Cobbold framework, one as a researcher has to make two decisions that will impact the measured diversity: how sensitive the measure should be to the presence of dominant species, i.e., the value of q, and if similarity should be taken into account, i.e., a similarity matrix is used instead of the identity matrix. Whatever decision is made, the measured diversity will be different for a given setting, but the underlying composition of the setting remains unchanged. Therefore, what varies is not the diversity in itself, but how it is *perceived*, which is dependent on a subjective decision made by the researcher. For example, if one measures diversity using q = 0 and an identity matrix to arrive at richness, this is a measure of the diversity that assumes that differences in relative abundance and similarity are unimportant. If one measures the same setting with q = 2 and Z instead of I, then the measured diversity will be lower than in the previous case, but the actual underlying composition of the setting has not changed. Ergo, the *diversity* remains the same, but the *perceived diversity* differs. For this reason, Leinster and Cobbold advocate characteriz-

ing settings by their diversity profile, which is the graphical representation of how the measured diversity changes as q increases. This can be done in both the naive and non-naive case, giving a full picture of the true diversity which can't be captured by any single statistic.

All in all, Leinster-Cobbold gives a unified framework coherent with the general understanding of diversity, interpreted as the effective number of species adjusted for relative abundance and similarity. This is a mathematically rigorous system that can be generalized for any field of study, including linguistics. In the next section, we shall therefore see what approaches have been taken to measuring diversity in linguistics, and how the general notion of linguistic diversity can be formalized through the Leinster-Cobbold framework, thus bridging ecology and linguistics.

2.3 Diversity in Linguistics

2.3.1 Greenberg's A and B-methods

The tradition of measuring linguistic diversity goes almost as far back in linguistics as in ecology, originating with *The Measure of Linguistic Diversity* by Greenberg (1956), in which he accounted for much of the diversity concept as discussed in the previous section. He did this to solve the problem

> "...of developing quantitative measures of this diversity in order to render such impressions more objective, allow the comparing of disparate geographical areas, and eventually to correlate varying degrees of linguistic diversity with political, economic, geographic, historic, and other non- linguistic factors." (p.109)

which perfectly corresponds with the efforts of this thesis: to develop quantitative measures of diversity to better model and understand linguistic diversity. As such, Greenberg introduced his *Monolingual non weighted index* as a diversity measure, defined as the probability of randomly picking two members of a population at random speaking different languages. One immediately notes that this is the same interpretation of diversity as the Gini-Simpson Index in the biodiversity context, with speakers instead of animals and languages instead of species. Greenberg referred to this as the A-method and he formally defined it as

$$A = 1 - \sum_i i^2 \tag{2.6}$$

where *i* takes on the values m, n, o..., each representing the proportion of speakers of the languages M, N, O...[1]. As can be seen, this is identical to the Gini-Simpson index defined in Equation (2.1), and to make Greenberg's terminology more coherent with the ecological theory on diversity, we can say that *i* is the relative abundance of i^{th} language, i.e., p_i. This is possible due to it being a monolingual index, making the very strong assumption that every person only speaks one language. Greenberg suggests multiple ways of tackling this, among others counting every

[1] This is the notation used by Greenberg in his original paper, and is suboptimal by modern academic standards. The notation is kept here for illustrational purposes.

multilingual speaker twice, thrice, etc., referred to as the split-personality method. While this conceptually makes sense, it requires data on multilingualism, which might be very scarce.

Greenberg furthermore adapts his A-method to create the *monolingual weighted index*, or the B-method. Here, Greenberg draws on the idea that not only the total range of differences, but also the magnitude of such differences, is important for the concept of diversity and states that:

> "First, with the same proportional distribution among the languages, we wish to call an area more diverse if a number of unrelated or distantly related languages are spoken there than one in which the languages are closely related descendants of the same original languages." (Greenberg, 1956, p. 109)

Thus Greenberg, similarly to the ecological literature, acknowledges that simply counting the number of languages is not sufficient, and first takes the step of incorporating relative abundance, and then similarity. His idea of the concept is that any two languages, *M* and *N*, can be related using a resemblance factor r_{mn}, a number between 0 and 1 that expresses how much the languages resemble each other, i.e., their similarity. He then proposes to weight the probability of randomly selecting speakers of m and n by their resemblance, giving

$$B = 1 - \sum_{mn}(mn) * r_{mn} \tag{2.7}$$

which in the previously used notation from the ecological literature would be equivalent to

$$B = 1 - \sum_{ij}^{S} p_i p_j Z_{ij}$$

As can be seen, the foundational ideas for measuring linguistic diversity are not very far from the ideas cultivated in ecology around biodiversity. Using the Leinster-Cobbold framework, Greenberg's B-method can be derived for q = 2 such that

$${}^{2}D^{z} = (\sum_{ij}^{S} p_i (Z_{ij} p_j)^{1})^{-1} = \frac{1}{\sum_{ij}^{S} p_i Z_{ij} p_j} = \frac{1}{1 - B} \tag{2.8}$$

What is lacking is therefore first and foremost the idea of expressing diversity as an effective number of languages, for the same reason that Gini-simpson, Shannon-entropy, and Rao's Q are insufficient as measures of biodiversity Jost (2018), which is solved by the Leinster-Cobbold framework. To the extent of my knowledge, there is only one instance of implementing the Hill numbers framework to measure linguistic diversity, namely by Grin and Fürst (2022), who operationalized linguistic diversity as naive diversity of $q = 1$, i.e., the Hill number of order 1.

This thesis thus picks up here by implementing the Leinster-Cobbold in the domain of linguistics. This would provide a unified view on diversity across both biology and linguistics, uniting the fields in both a conceptual and computational approach towards diversity.

2.3.2 Similarity

For this purpose, a way of quantifying the similarity between languages is necessary, analogous to the resemblance factor suggested by Greenberg (1956). Greenberg explicitly suggests an operationalization of similarity as

> "Using an arbitrary but fixed basic vocabulary, e.g. the most recent version of the glottochronology list, the propor- tion of resemblances between each pair of languages to the total list is given as a fraction. Since we are measuring actual resemblance, not historic relationship, it is suggested that both cognates and loans be included as resemblance." (p.110)

This approach suggested by Greenberg would thus account for the linguistic dimension of lexicon. Grin and Fürst also argues for this as a possibility, referring to Dyen, Kruskal, and Black (1992) making use of the lexicostatistical method of comparing the percentage of cognates based on core vocabulary. Other lexical approaches are given by Ginsburgh and Weber (2020), suggesting using levenshtein distances, popularized in dialectometry through e.g. Gooskens and Heuven (2020); Gooskens et al. (2018); Heeringa, Gooskens, and van Heuven (2023); Heeringa, Johnson, and Gooskens (2009); Nerbonne and Heeringa (2002); Nerbonne and Hinrichs (2006) as a way of discerning intelligibility between linguistic varieties. This approach will be explored further in Chapter 3. Apart from lexicon, Grin and Fürst suggests that other typological aspects of language, e.g., phonology and syntax, might be equally important. Finally, they raise an aspect of distance that does not take on a synchronic, but rather a diachronic view by considering cladistic distance, referring to reconstructed phylogenetic trees, and measuring the proportion of shared ancestry compared to maximal branching length, as suggested by Fearon (2003) and formalized by Ginsburgh and Weber (2020) such that

$$\tau_{jk} = 1 - (\frac{l}{m})^{\delta} \tag{2.9}$$

where *l* is the number of shared branches between the languages j and k, *m* is the maximum number of branches between any two languages, and δ is a parameter that adjusts how sensitive the measure is to languages having shared ancestry compared to languages lacking shared ancestry. The intuition behind using cladistic distance to operationalize Greenberg's resemblance factor is the assumption that two languages that have branched off later and share more ancestry will be more similar than languages that branched off early, or do not share any ancestry (Grin & Fürst, 2022). The implicit argument is thus that cladistic distances can be used as a proxy for actual similarity between languages. Whether this assumption actually holds is unclear, since it has, to the extent of my knowledge, not been tested on larger samples of languages with high variability. A study by Heeringa et al. (2023) recently undertook this effort on a sample of 16 Indo-European languages, all belonging to the Slavic, Romance, and Germanic subfamilies. Within all three subfamilies, they found strong and significant product-moment correlations between lexical and cladistic distance (r = [0.851, 0.872, 0.887], $p \leq 0.05$). In the Germanic and Slavic families, significant correlations between cladistic and phonetic distance were found (r = [0.832, 0.706], $p \leq 0.05$), as well as between cladistic and syntactic distance among the Slavic languages (r = 0.353, $p \leq 0.05$). Between the syntactic, lexical, and phonetic

distances, moderate to strong correlations were found within the subfamilies, but across the subfamilies, the authors concluded that there are no fixed relationships between the distance measures. As such, this would support the idea that cladistic distance can be used as a proxy for other linguistic distance/similarity measures. However, this is based on a very small sample of languages, not accounting for most of the variation in linguistic behavior observed globally. Furthermore, all languages included are closely related, which wouldn't be the case in many linguistic settings. As a matter of fact, languages can exhibit similarity without being genealogically related, through e.g., contact (Ponti et al., 2019). Referring back to Greenberg (1956):

> "we are interested in actual resemblance, not historical relationship." (p.110)

2.3.3 Cross-Linguistic Data

It is clear that in the ideal of worlds, implementing the Leinster-Cobbold similarity framework by accounting for linguistic similarity, as many aspects of language as possible would be considered, e.g., lexicon, phonology, and morphosyntax, to accurately account for linguistic resemblance. The operationalization of such similarity measures would need data describing these aspects for all, or at least as many languages as possible, and they would need to be efficient to calculate, since the Leinster-Cobbold similarity requires data on all pairwise similarities. In a global context with all languages of the world, this would potentially generate similarity matrices as large as $\sim 7000 * 7000 = 49.000.000$ to adequately account for diversity. Furthermore, if a language is not described in such a way that similarities can be measured, then that language can't be accounted for when measuring diversity. As such, the most crucial aspect of operationalizing linguistic similarity on a global scale will be the coverage of data sources, which is why the entire Chapter 4 is dedicated to exactly this.

To enable this endeavor, the importance of cross-linguistic open databases that emerged at the beginning of the 21st century can't be overstated. The generally considered first typological database in the modern sense was the World Atlas of Language Structures (WALS), published as a book in 2005, and online as a database in 2008 (Dryer & Haspelmath, 2013). WALS gathers features on linguistic typology concerning lexicon, phonology, and grammar, and encodes languages from across the world according to a set number of features, e.g., word order, and assigns each language a value based on available descriptions of that language. By encoding as many languages as possible, this allows researchers to study and compare languages on a global scale, as done by e.g., Dahl (2008). Since then, multiple similar databases have been developed and openly published, focusing on specific aspects of language and areas. A survey on the topic was carried out by Ponti et al. (2019), resulting in an overview of publicly available databases that is presented in Table 5.2. As such, we can see that there is a lot of variation in coverage of these databases, with some having a very extensive coverage with languages in the thousands, such as ASJP, WALS, PHOIBLE Online, and URIEL. They could all be possibly used for the purpose of assessing linguistic similarity globally, but although the language coverage is impressive, for them to be useful in a diversity context, relative abundance, i.e., the number of speakers of these languages, are equally important, and also the geographical distribution of the languages and speakers covered. Without this knowledge, it can't be determined how useful such databases indeed are to measure linguistic diversity globally. As such, this is the

key first step to analyzing linguistic diversity, relating back to RQ1. Once this is done, measures from the respective databases can be derived and analyzed to address RQ2. Based on the results in RQ1 and RQ2, the most appropriate operationalization of similarity on a global scale can be concluded, which then can be used to address RQ3. By doing this, the thesis will be able to conclude to what extent similarity-aware linguistic diversity can be implemented on a global scale, and if this impacts how diversity is perceived.

Name	Levels	Coverage	Feature Example
World Atlas of Language Structures (WALS)	Phonology, Morphosyntax, Lexical semantics	2,676 languages; 192 attributes; 17% values covered	ORDER OF OBJECT AND VERB Amele: OV (713) Gbaya Kara: VO (705)
Atlas of Pidgin and Creole Language Structures (APiCS)	Phonology, Morphosyntax	76 languages; 335 attributes	TENSE–ASPECT SYSTEMS Ternate Chabacano: purely aspectual (10) Afrikaans: purely temporal (1)
URIEL Typological Compendium	Phonology, Morphosyntax, Lexical semantics	8,070 languages; 284 attributes; ∼439,000 values	CASE IS PREFIX Berber (Middle Atlas): yes (38) Hawaaian: no (993)
Syntactic Structures of the World's Languages (SSWL)	Morphosyntax	262 languages; 148 attributes; 45% values covered STANDARD NEGATION	IS SUFFIX Amharic: yes (21) Laal: no (170)
AUTOTYP	Morphosyntax	825 languages; ∼1,000 attributes	PRESENCE OF CLUSIVITY !Kung (Ju): false Ik (Kuliak): true
Valency Patterns Leipzig (ValPaL)	Predicate–argument structures	36 languages; 80 attributes; 1,156 values	TO LAUGH Mandinka: 1 > V Sliammon: V.sbj[1] 1
Lyon–Albuquerque Phonological Systems Database (LAPSyD)	Phonology	422 languages; ∼70 attributes	â AND ú Sindhi: yes (1) Chuvash: no (421)
PHOIBLE Online	Phonology	2,155 languages; 2,160 attributes	m Vietnamese: yes (2053) Pirahã: no (102)
StressTyp2	Phonology	699 languages; 927 attributes	STRESS ON FIRST SYLLABLE Koromfé: yes (183) Cubeo: no (516)
World Loanword Database (WOLD)	Lexical semantics	41 languages; 24 attributes; ∼2,000 values	HORSE Quechua: kaballu borrowed (24) Sakha: s1lg1 no evidence (18)
Intercontinental Dictionary Series (IDS)	Lexical semantics	329 languages;	WORLD Russian: mir Tocharian A: arki´sos ¯. i
Automated Similarity Judgment Program (ASJP)	Lexical semantics	7,221 languages; 40 attributes	I; Ainu Maoka: co7okay Japanese: watashi

Table 2.1: A table covering major linguistic databases with typological information according to name, linguistic level, and coverage. The table was reprinted from table 1 in Ponti et al. (2019).

3 Method and Data

In this chapter, the methodology of the thesis is introduced. First, the sources from which linguistic and typological data were accessed and analyzed are presented one at a time, together with the linguistic similarity measures derived from each data source. Then, the methodology used to analyze and validate the linguistic similarity measures is presented. Finally, the methodology concerning linguistic diversity is introduced.

3.1 Data Sources and Similarity Measures

Since the thesis aims to assess linguistic diversity on a global scale, the prerequisite for any data source to be deemed potentially useful will first come down to the sheer size. Secondly, they should be selected in such a way that they cover different linguistic aspects. The databases that were thus found to be potentially useful in a global context and their respective linguistic area are:

- Speakers - Ethnologue (Eberhard & Fennig, 2025)
- Phylogenetics - Glottolog (Hammarström, Forkel, Haspelmath, & Bank, 2025)
- Phonology - PHOIBLE (S. Moran & McCloy, 2019)
- Morphosyntax - Grambank (Skirgård, Haynie, Hammarström, et al., 2023)
- Lexical Semantics - ASJP (Wichmann, Holman, & Brown, 2022)

Considering Table 2.1 listing major linguistic databases as identified by Ponti et al. (2019), there is an overlap through the inclusion of PHOIBLE and ASJP, and in addition, Glottolog, and Grambank are considered. While Ethnologue and Glottolog are not primarily sources of typological information, they are of monumental importance for structuring linguistic data. Furthermore, Ponti et al. include WALS, which has been left out of this thesis. Instead, a database that is best regarded as an extension and improvement upon the aforementioned, Grambank, is included. Also, URIEL (Littell et al., 2017) and its successor URIEL+ (Khan et al., 2025) are purposefully left out due to them mainly being supersets of the other databases. While they are extensive, they impute missing feature values, creating a black box nature and making them difficult to analyze. Each of the utilized databases will be described in further detail in the coming sections, together with the similarity measures derived from them.

3.1.1 Ethnologue and Language Data

Description

Ethnologue is a catalog seeking to contain information on all of the world's living languages, describing itself as one of the most authoritative sources on languages, as retrieved on the 6th of March 2025 (Ethnologue, 2025a). The first edition of Ethnologue was published in 1951, motivated by a desire to share information on the linguistic endeavor of bible translation, and grew from including data on 46 languages to identifying 4,493 languages by 1969. In 1971, the data was integrated into a computerized database and is today maintained by SIL International, a faith-based non-profit developing language solutions (SIL, 2025).

Today, the database is published in its twenty-eight edition and contains data on 7,159 living languages (Eberhard & Fennig, 2025). An important aspect of the database is the ISO639-3 standard developed by Ethnologue together with the International Organization for Standardization (ISO) to create a unique, three-letter identifier to denote any language, living or extinct. Ethnologue itself seeks to include all living languages or any language that has gone extinct since the catalog was first published, and as such, it forms the basis for the data about the current linguistic situation in the world for this thesis.

Data

The data used stems from a dump of the Ethnologue database in 2022, thus equivalent to the 24th edition, made available to the University of Vienna. The data is structured using ISO639-3 codes to identify languages and ISO3166-1 to identify countries, with each row in the data forming a country-language pair, endowed with speaker data from Ethnologue combined with speaker data from the Joshua Project (2025). The dataset was originally compiled and recently employed by Zeh (2025). Since the speaker data from Ethnologue is proprietary, it can't be made openly accessible, but the measures derived from it can be found in Table 7.3 in the Appendix. By using the data from Ethnologue with ISO639-3 codes as a linguistic framework, defining what a language is and what makes two varieties separate entities is left up to Ethnologue, with their distinction being based on a combination of mutual intelligibility and ethnolinguistic identity (Ethnologue, 2025b).

Since Ethnologue serves as one of the main authorities on global language statistics, it forms a reliable basis for an analysis of linguistic diversity, where the relative abundance of languages is to be taken into account. Therefore, Ethnologue data with information on country-language pairs with speaker populations are used in this thesis as a baseline for investigating the extensiveness and coverage of databases with typological information and to measure linguistic diversity. A caveat to the data, which should be discussed, is the status of L1 and L2. While one animal usually does not belong to multiple species, one person can speak multiple languages, and as such, all language configurations of a given person would, in the ideal of worlds, be taken into account when measuring linguistic diversity (Greenberg, 1956). However, as argued by Bromham et al. (2022), L2 data is not readily available for most languages, and even L1 data from Ethnologue remains estimates associated with uncertainty. In this thesis, no distinction is made between L1 and L2 due to the conceptual and operational difficulty associated with

the terms, and due to the Joshua Project basing their estimates on ethnic identity. Since ethnicity and language are strongly intertwined (Fought, 2011), any clear distinction between L1 and L2 becomes impossible. Instead, in the context of this work, any speaker number should be viewed as a representation of the linguistic landscape (Carr, 2022) in a given country for the purpose of estimating the diversity of the languages in use. And since the data considers languages in use, it naturally follows that "extinct" languages or languages without speakers are not considered. When referring to this data in the rest of the work, it will be referred to as the *speaker data* or the *baseline*.

3.1.2 Glottolog

Glottolog, similar to Ethnologue, is a "linguistic catalog" seeking to provide resources to describe and classify the languages of the world (Nordhoff & Hammarström, 2011). It was conceived out of the need for open and more comprehensive bibliographies, identifying Ethnologue among others as insufficient in that regard (Hammarström, 2015). Glottolog therefore builds on an idea of collecting references on language, and linking them to a linguistic unit defined as a *languoid* to escape the conceptual problems associated with the dichotomous view of "languages" as opposed to "dialects" (Nordhoff & Hammarström, 2011). In Glottolog, each languoid is given a unique glottocode and corresponds to either a family, language, or dialect, and is a mathematical set that can contain any number of subsets. These subsets can be either other languoids, such that a family consists of languoids equivalent to the languages belonging to that family, or the subset can be doculects, which are any linguistic variety described by a given documentation. As such, a languoid consists of multiple doculects, each describing that particular language (Hammarström, 2015).

Apart from providing a structure to categorize languages and language families, Glottolog also provides geographic information about each languoid by assigning each language a geographical coordinate of longitude and latitude and assigning it to its linguistic area, called a *macroarea*. These areas were designed by Hammarström and Donohue (2014) to address the need of relating linguistic structures with geography and divides the world into 6 areas heavily corresponding to continental divisions: Eurasia, Africa, North America, South America, Papunesia and Austrailia, with at least 250 languages uniquely found in each of them (Glottolog Project, 2024).

Since Glottlog extensively, transparently, and accurately provides structured linguistic data, language families and macroareas were sourced and added to all languages in the speaker data to add a phylogenetic and geographical dimension to asses the coverage of the databases. Furthermore, since Glottolog provides multiple levels of language classification into families, subfamilies, etc., on an entirely global scale, it was utilized to calculate a linguistic similarity measure based on phylogenetics. As mentioned in Chapter 2, cladistic distance is a commonly used measure of linguistic distance. It is frequently operationalized by counting the steps one needs to go from one language to another in a cladistic tree (Heeringa et al., 2023), meaning that the smaller the distance, the closer related the languages are inferred to be (Gurevich, Herman, Toubal, & Yotov, 2025), assuming that language change is constant.

The problem with using cladistic distance is, however, that languages do not change at a

constant pace (Trudgill, 2020), meaning that two languages that diverged later can be more dissimilar than two languages that diverged early, if one of the languages changes quickly or the other two exert influence on each other through contact. Therefore, cladistic distance can only be used as a proxy for actual similarity and resemblance, although such measures have been suggested to correlate strongly with other similarity measures, especially lexical distance (Heeringa et al., 2023). A further conceptual problem with cladistic distance is that it entirely depends on expert classification, and experts are not guaranteed to agree about what constitutes a language family (Dahl, 2008). While glottolog provides a unified framework for classifying languages and subfamilies, there is still a potential risk of research bias involved, since different language families will have received differing amounts of scholarly attention, resulting in more extensive classification, which introduces variability to the measures when applied to different language families. In addition, similarity based on phylogenetics only takes vertical development of languages into account, ignoring horizontal transfer. For example, French and English are phylogenetically fairly dissimilar, belonging to different language families, but share a large amount of lexicon due to language contact, which is disregarded by cladistic-based similarity measures. Another problem lies in the fact that a cladistic distance can only be meaningful within a language family, resulting in two languages from different language families always being modelled as equally dissimilar, even though this might not be the case. For example, the existence of an Altaic language family has been proposed due to observed similarities between e.g., Turkic and Mongolic languages, but the proposal has never been accepted and achieved scholarly consensus. Thus, the Turkic and Mongolic languages will be modelled as equally similar to the Turkic and Indo-European languages, although they aren't. Although all of these concerns should be raised, data on language kinship exists in some form for more or less all languages of the world through Glottolog, making phylogenetic similarity potentially useful on a global scale.

In this thesis, the property of Glottolog being structured as a mathematical set was used to derive a phylogenetic similarity measure. Each language was represented as the set of languoids itself was a subset of, such that the set defining English, for example, would be {Indo-European, Classical Indo-European, Germanic, Northwest Germanic, West Germanic, North Sea Germanic, Anglo-Frisian, Anglic, Later Anglic, Middle-Modern English, Macro-English}. For any two languages, their phylogenetic similarity was then given as the overlap between these sets using Jaccard similarity (Jaccard, 1912), which allows for different set lengths and efficient computation.

To get a formal definition of Jaccard similarity, let the set of intermittent phlyogenetic stages representing language A be a, and the set representing language B be b. Then, the phylogenetic similarity between A and B, $s(A, B)$ is defined as

$$s(A, B) = \frac{|a \cap b|}{|a \cup b|}. \tag{3.1}$$

Programmatically, these pairwise similarity calculations were implemented using the *vegdist* function from the *vegan* package (Oksanen et al., 2025) in R (R Core Team, 2024)[1]. Based on

[1]All code used in the thesis can be found at:
https://github.com/Eszettfors/Global-Linguistic-Diveristy—Similarity

these similarity measures, what here is referred to as *Mean Phylogenetic Similarity* was calculated for every country as

$$\frac{\sum_{i \neq j} s_{ij}}{N(N-1)} \tag{3.2}$$

where *N* is the number of languages in the similarity matrix and s_{ij} is the phylogenetic similarity between any two different languages *i* and *j*.

To exemplify its implementation, consider again the set used to describe the ancestry of English and compare it to the set describing German: {Indo-European, Classical Indo-European, Germanic, Northwest Germanic, West Germanic, High German, Upper German, Middle-Modern High German, Modern High German, Upper Franconian, Global German}. The ancestry is shared up until West Germanic, where a split takes place, meaning the sets share 5 items, giving an intersection of 5. The union of the two sets gives {Indo-European, Classical Indo-European, Germanic, Northwest Germanic, West Germanic, North Sea Germanic, Anglo-Frisian, Anglic, Later Anglic, Middle-Modern English, Macro-English, High German, Upper German, Middle-Modern High German, Modern High German, Upper Franconian, Global German} consisting of 17 items. As such, their similarity is $5/17 = 0.29$. The mathematical properties of the measure imply that if the number of intermittent stages between the root and the split is larger, i.e., the languages share more common ancestry, the nominator increases, and as such also the measured similarity. If more branching takes place after the split, i.e., the languages are perceived to develop and diverge, the denominator increases, ergo the similarity decreases. As such, this becomes a measure of phylogenetic similarity, which, however is not an absolute measure, but relative to the set used. Again, in the case of English and German: if one were to consider the similarity between Germanic languages, it would make sense to drop the 'Indo-European' element and use 'Germanic' as a root. This would result in different values for the measured similarity, since both the union and intersection reduces by 1, meaning that the similarity of English and German would have different values in a Germanic context compared to an Indo-European.

Since this is a novel approach to quantifying phylogenetic similarity prompted by available global data on phylogenetic relationships between languages, its exact alignment with branching distances hasn't been established. Conceptually, however, the measures are aligned and thus also suffer from the same shortcomings described in the previous paragraph. Establishing precise alignment between branching distance and the set-based approach introduced here would warrant its own chapter and would be an appropriate addition to the literature, but this goes outside the scope of the thesis. Instead, it is acknowledged that the measures conceptually are the same, but the set-based approach is intuitively more efficient and appropriate for working with the data available to measure phylogenetic similarity on a global scale through Glottolog.

3.1.3 PHOIBLE

Description

PHOIBLE is a repository for collecting phonological inventory data compiled from source documents, tertiary sources, and databases (S. Moran & McCloy, 2019). The version used in the thesis

is PHOIBLE 2.0, which was released in 2019 and contains data on 3020 inventories in 2186 languages. Each phonological inventory is represented as a set of phonemes as described in the source document for the language entry, and each phoneme is consistently encoded in the database in Unicode IPA according to the 2005 update of the IPA chart (International Phonetic Association, 1999).

PHOIBLE is part of CLLD – Cross-Linguistic Linked Data, which is a project of the Max Plank Society seeking to link together and publish datasets that can be used for large-scale comparison of languages in the area of diversity linguistics (Forkel & Haspelmath, 2023). In the collection of datasets published by CLLD, PHOIBLE is given the description:

> "The world's largest database of phonological inventories." (datasets, published datasets)

motivating its inclusion in this thesis, signifying it as the most extensive source of phonological feature representation available.

Analysis

The data was downloaded from the PHOIBLE-GitHub repository (PHOIBLE, 2025) and read into R (R Core Team, 2024).[2] The dataset was subset to only include languages present in the speaker data, bringing some caveats to it. As explained in Section 3.1.1, linguistic diversity is to be assessed on a country level, limiting the analysis to languages with population data. These are limitations that do not necessarily apply to typological databases in CLLD, which seeks to help

> "...collect the world's language diversity heritage." (Forkel & Haspelmath, 2023)

thus for example including extinct languages. Furthermore, the CLLD framework uses Glottolog and the associated glottocodes to uniquely define and link languages. As a result, what is defined as a language in Ethnologue might not have the same status in Glottolog and, in extension, PHOIBLE and other CLLD databases, resulting in potential data loss. While this is undesirable, it is unavoidable due to the different frameworks being applied by, on the one hand, Ethnologue and the Joshua Project, and on the other hand, Glottolog.

For each country present in the speaker data, the number of languages in the PHOIBLE data was counted and divided by the total number of languages present in that same country, giving a measure of coverage between 0 = no languages covered by PHOIBLE and 1 = all languages covered by PHOIBLE. Similarly, the coverage of speakers in each country was calculated using the sum of speakers in a given country as opposed to the number of languages. On a language level, the coverage was determined by assessing the percentage of languages, language families, and speakers also found in PHOBLE.

Furthermore, the representation of the categories macroarea and speaker number was assessed by comparing the distribution across categories in both databases. Macroarea was, as previously discussed, adapted from Glottolog, and speaker data was derived from Ethnologue

[2] All code used in the thesis can found at

and the Joshua Project. Based on the speaker data, each language was assigned to one of four categories based on the total number of speakers grouped into four quartiles. The ranges of the number of speakers in each category can be seen in Table 3.1.

Category	minimum number of speakers	maximum number of speakers
Low	1	1 400
Lower-Mid	1 400	10 000
Upper-Mid	10 000	71 900
High	71 900	971 424 800

Table 3.1: A table introducing the speaker category extracted based on the number of speakers of any given language. Each category contains 25 % of languages for which speaker data exists. The values denote the ranges of speaker numbers in the respective category.

For each category, the percentage of languages belonging to that category in PHOIBLE was subtracted from the percentage of languages belonging to the same category in the baseline, giving the *Absolute Percent Difference* as a measure of representation. As such, a positive value equals an overrepresentation of the category in PHOIBLE, while a negative value equals an underrepresentation, with the magnitude of the effect indicated by values from 0 to 100. In addition, coverage on a country level was assessed based on the speaker data. For each country, the number of languages and speakers was summed up, and the coverage was given as the percentage of languages and speakers also present in PHOIBLE.

For each country, a similarity matrix was derived based on the phonemic inventory of the languages of that country present in PHOIBLE. Since PHOIBLE maintains multiple inventories per language, e.g., English contains inventories for, among others, both British and American English, a strategy for deciding on which to use when multiple were present had to be decided upon. As a pragmatic solution, it was decided always to use the larger inventory in case of conflicts, assuming that the larger is more likely to capture any language's true phonemic richness. There are most likely better solutions; however, constructing them would go outside the scope of this thesis. Instead, a structured bias towards varieties with larger inventories is introduced, creating a smaller limitation. To measure the similarity between the phoneme inventories of any two given languages *A* and *B*, each language was represented by its set of phonemes derived from PHOIBLE, denoted as *a* and *b*. The phonemic similarity *s* between A and B was then calculated as the Jaccard similarity (Jaccard, 1912), which measures the percentage of overlap between two sets as suggested by S. P. Moran (2013) and implemented on phoneme inventories by Archibald (2022) and Anderson et al. (2021). Formally, this was defined for phylogenetic similarity in equation (3.1), and again, pairwise similarity calculations were programmatically implemented using the *vegdist* function from the *vegan* package (Oksanen et al., 2025). Based on these similarity measures, what here is referred to as *Mean Phonemic Similarity* according to Equation (3.2). This gives a naive measure of phonemic diversity in each country, which in turn can be incorporated with richness and relative abundance data and be utilized in the Leinster-Cobbold framework to measure linguistic diversity.

This measure of phonemic similarity is a conceptually simple measure that tells one how

much of the total phoneme inventory is shared, and making use of set operations through Jaccard similarity allows for efficient calculations through hashing (Elgendy, Younes, Abu-Donia, & Farouk, 2024), which is necessary if considering language on a global scale due to the large number of languages. However, possible limitations exist since phonemes can differ in importance between languages, meaning that although two languages are considered entirely similar with identical inventories, one might make more use of a certain subset of items to a greater extent than the other, adding a dimension of dissimilarity between the languages which isn't captured by the measure. Furthermore, the implementation of this measure is naive in the sense that each phoneme is considered equally dissimilar to all other phonemes, which isn't the case, e.g. /p/ and /b/ share more similarities in articulation compared to /p/ and /n/. This could potentially be addressed by e.g., a weighting of the Jaccard similarity as implemented by Anderson et al. (2021). Still, the measure is straightforward and efficiently calculated with the data provided by PHOIBLE and thus appropriate for this thesis.

3.1.4 Grambank

Description

Grambank is a typological database describing the morphosyntactic features of languages based on available grammars to

> "...investigate the global distribution of features, language universals, functional dependencies, language prehistory, and interactions between language, cognition, culture, and environment." (Skirgård, Haynie, Blasi, et al., 2023, p. 1)

Grambank was released in 2023 and is currently in its first version, v1.0 (Skirgård, Haynie, Hammarström, et al., 2023). Similar to PHOIBLE, it is part of the CLLD (Forkel & Haspelmath, 2023) and available through Zenodo. It currently contains 2467 languages and dialects across 215 language families and 101 isolates, coded for 195 features. Each feature is phrased as a question, such as feature GB020: *Are there definite or specific articles?*. If this is true, according to a grammar sketch, it is binarily coded as 1. If it is false, it is coded as a zero, allowing each language to be represented as a vector of size 195. There are six exceptions to this, allowing for multistate encoding, which can be binarised to give a feature space of size 205 (Grambank, 2024b). The questionnaire giving these 195 features was developed by adapting 118 features from the Sahul questionnaire (Reesink, Singer, & Dunn, 2009), 42 features from the Nijmegen Typological Survey (Skirgård, van der Meer, & Hammarström, 2014), and 4 from Pioneer (Dunn, Levinson, Lindström, Reesink, & Terrill, 2008). The remaining 34 features were developed by partially remodelling features from WALS (Dryer & Haspelmath, 2013), giving concepts which are language independent, thus allowing for cross-linguistic comparison (Grambank, 2024a).

In many ways, Grambank is similar to WALS, and according to Skirgård et al., they are best viewed as "sister-databases". While Grambank carries on the spirit of WALS by modelling structural concepts for languages on a global level, it was, in contrast to WALS, created with computational analysis in mind by avoiding data inconsistency and avoiding feature dependency problems (Grambank, 2024c). Furthermore, Grambank addresses the important aspect of data completeness. In contrast to WALS, which has 84% missing data, Grambank only contains 24%

missing data in total (Skirgård, Haynie, Blasi, et al., 2023). This has considerable implications for cross-linguistic similarity comparison, since a lack of available data for any two languages might result in them accidentally appearing more similar than they are (Dahl, 2008). Still, Grambank is the most comprehensive, complete, and reliable source of morphosyntactic typological data on a large scale, thereby also making it the most appropriate source for this thesis.

Analysis

The analysis of Grambank follows a structure similar to that of PHOIBLE as outlined in Section 3.1.3. v1.0 of the database was downloaded through Zenodo and read into R, after which the percentage of languages, language families, and speakers present in both the speaker and Grambank data was counted and expressed as a percentage. The number of languages was aggregated on macroarea and speaker categories, and over- and underrepresentation was given as the absolute percent difference for each category in Grambank relative to the speaker data. For each country, the percentage of languages and speakers present in Grambank relative to the baseline was calculated, and a similarity matrix was derived based on the languages present. For that purpose, however, a stance needs to be taken on how complete data should be for distance measures to be deemed reliable in light of the discussion on data completeness in Grambank and WALS.

Since one feature can't represent the grammatical functioning of a language as a whole, a lack of data completeness presents a problem. Take English and Japanese, for example: the languages share values for the grambank feature *GB028: Is there a distinction between inclusive and exclusive?*, since neither of the languages mark inclusiveness-exclusiveness, which thus would make them appear identical if only considering this parameter. In reality, however, Japanese and English function very differently; Japanese is typically considered a synthetic language (Majtczak, 2014) while English is viewed as analytic (Yutong, 2024). Concluding that English and Japanese are morphosyntactically similar based on this one observation would therefore result in a type I error.

Dahl (2008) decided in his analysis of WALS that 101 was an appropriate minimal number of features for his investigation of a posteriori sampling related to distance measures, while Wichmann and Holman (2010) found 45 features to be an optimal minimal number of features for the WALS data based on the probability of finding related languages based on feature similarity. Skirgård, Haynie, Blasi, et al. (2023) decided to remove Grambank languages and features with more than 25% missing data for the purpose of feature imputation, and Ploeger et al. (2025) explicitly followed this line of cropping more than 25% missing data. While using a high threshold for data completeness would indeed result in more reliability, it might excessively disregard languages, which would counteract the purpose of measuring global linguistic diversity. To assess the impact of different data completeness thresholds on the linguistic richness of the Grambank dataset, the empirical cumulative distribution was plotted against the completeness as can be seen in Figure 3.1. Data completeness is hereby defined as the percentage of instances where data is present compared to where it could be present (Hassenstein & Vanella, 2022), which can be done on both a parameter and language level. On a language level in Grambank, it can be observed that there is a small uptick beyond 25% completeness, after which the curve becomes steeper. This means that from this point on, an increase in data completeness is asso-

ciated with an increasing loss of linguistic richness. As such, it follows that 25% completeness would be a suitable threshold for maximizing linguistic richness while minimizing missing data. Furthermore, a 25% completeness[3] cut-off threshold applied to the 195 features of Grambank would result in every language considered for similarity measures being defined for at least 49 features, closely aligning with the feature number used by Wichmann and Holman (2010).

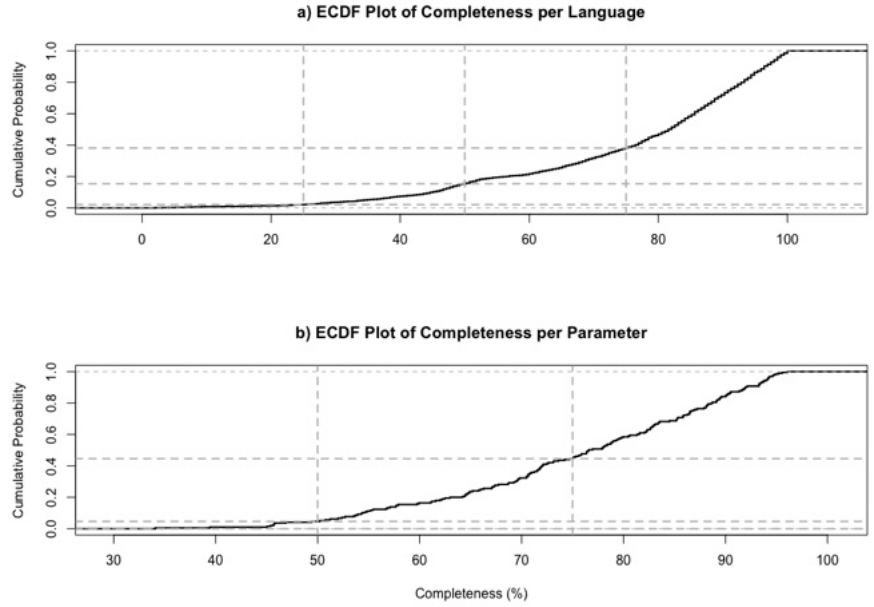

Figure 3.1: Plots showing the empirical cumulative distribution of completeness in Grambank based on a) language and b) parameters. Dashed lines mark 25%, 50% and 100% completeness.

After applying the cut-off threshold to the Grambank data, pairwise structural similarity was calculated for languages in each country in the same way as Dahl (2008); Ploeger et al. (2025); Wichmann and Holman (2010) and formalized as the complement to the Gower coefficient as defined by Hammarström and O'Connor (2013). Let the two languages A and B be represented by their respective feature vectors from Grambank. The structural similarity *s* between A and B is then defined as

$$s(A,B) = \frac{\#_{i \in DEF(A,B)} A[i] = B[i]}{|DEF(A,B)|} \tag{3.3}$$

where $DEF(A,B)$ is the set of features for which both A and B are defined and *i* is the *i*th feature in this set. Although the cropping applied ensured that all languages considered for pairwise distance were defined for at least 49 features, this did not, however, ensure that any two pairs of languages are both defined for the same 49 features, meaning that the measure still could be unreliable. Therefore, a further constraint was applied at the point of implementation

[3] The 25% missing data threshold defined by Skirgård, Haynie, Blasi, et al. (2023) would be equivalent to a 75% completeness threshold

of the morphosyntactic similarity, only calculating pairwise distances if $|DEF(A, B)| \geq 49$. As such, a similarity matrix was derived for each country, based on which *Mean Morphosyntactic Similarity* (MMS) could be derived as a naive measure of morphosyntactic diversity identically to MPS in Equation (3.2). Furthermore, for each similarity matrix, a reliability matrix was derived containing $|DEF(A, B)|$ for all language pairs in a given country, since the more features two languages both are defined for, the more complete the data is and the more reliable the measure. Based on this matrix, the mean number of shared features among the languages in a country was calculated as an indicator of the reliability of derived morphosyntactic similarity for a given country.

3.1.5 The ASJP database

Description

The ASJP database is a database developed by Wichmann et al. (2022) to determine similarity between languages, utilized in particular by the Automated Similarity Judgement Program (ASJP) to classify language groups automatically (Brown, Holman, Wichmann, & Velupillai, 2008). As such, it takes a lexicostatistical approach as devised by Swadesh (1955) by comparing shared cognacy from a set of words corresponding to semantical concepts. According to Swadesh, this would be concepts that make up the 'core-vocabulary', such as *hand or name*, thus being present in most languages. This resulted in a list of 100 concepts (Swadesh, 1971), that ASJP purposefully sought to collect for as many languages as possible, representing each word phonetically using 41 symbols known as *ASJPcode* (Brown et al., 2008). The ambition to collect lists of 100 words was later reduced to 40 words, due to observing similar classificatory results (Holman et al., 2008), and as of version 20 - which is used in this thesis - the database contains wordlists for 5590 ISO639-3 languages (Wichmann et al., 2022) of which 86% are annotated for all 40 concepts according to List et al. (2022). As such, it

> "...is the most extensive wordlist collection in terms of cross-linguistic coverage."
> (List et al., 2022, p. 2)

Analysis

Similar to Grambank and PHOIBLE, the number of speakers, families, and languages present in the ASJP database was counted and expressed as a percentage of the baseline data to measure the coverage. The numbers were aggregated on macroarea and speaker-community-size to estimate the representativeness of the database, and for each country, the percentage of languages and speakers found in the ASJP database was calculated, again using absolute percent difference as a measure of representativeness. Furthermore, a similarity matrix for each country was derived based on the wordlists in the ASJP database to give a measure lexical similarity.

In the ASJP database, multiple ISO-codes might be represented by multiple wordlists, and likewise, one concept might be represented by multiple values, as in the case of e.g., Swedish. In the ASJP database, both Swedish and Swedish 2 exist: Swedish has a list of 40 words and

Swedish 2 has a list of 107 words, with two entries for multiple concepts such as "SKIN", which is represented by the synonyms *hud* and *skin*. To resolve this and make each language and concept uniquely defined, all wordlists were subsetted to the 40 concepts of interest as defined by Holman et al. (2008), and if multiple words for a single concept existed, the first one was selected according to the order in which they appear in the wordlist. Then, the data completeness for each language was calculated as the percentage of the 40 concepts present in the wordlist. If there were two competing wordlists for the same language as defined by their ISO-code, the one with the highest data completeness was selected, as this would correlate with more data and thus more reliable results. If two competing wordlists had the same completeness, the one appearing first in the database was selected.

To calculate lexical similarity between pairs of languages, the normalized Levenshtein distance (LDN) was used as described by Petroni and Serva (2010) and Bakker et al. (2009), being utilized within the Automated Similarity Judgement Program as well. Consider the concept *c* in the pair of languages A and B represented by the two strings a_c and b_c in ASJPcode. Let the distance between a_c and b_c be the Levenshtein distance, equaling the smallest number of insertions, deletions, and substitutions necessary to turn a_c into b_c, denoted $d_l(a_c, b_c)$ (Levenshtein, 1966). LDN is then defined as

$$LDN(a_c, b_c) = \frac{d_l(a_c, b_c)}{max(l(a_c), l(b_c))} \tag{3.4}$$

where *l(a_c)* and *l(b_c)* are the lengths of the strings, thus normalizing the measure by the largest possible Levenshtein distance, such that maximal dissimilarity between the two words equals 1. The lexical similarity between A and B is then given as 1 minus the mean LDN for all c's for which both A and B are defined. Formally, this is defined as

$$s(A, B) = 1 - \frac{1}{|C|} \sum_{c \in C} LDN(a_c, b_c) \tag{3.5}$$

where C is the set of concepts for which both A and B are defined. It follows that if two languages have similar lexicons, the LDN between the representation of concepts will be low, giving a high similarity. Wichmann, Holman, Bakker, and Brown (2010) argue, however, that this measure is susceptible to chance similarity due to, e.g., overlapping of phoneme inventories. They therefore propose a further normalization step, dividing LDN by the mean LDN for every other pair of words and thus giving LDND, which is more complex and requires more computing power. Their implementation of LDND is done in the context of inferring phylogenetic relationships, but for the purpose of assessing pure similarity between a set of words, it can be argued that the random similarity added by similar phonemic and phonotactic structures is a valid and desirable feature, which is why this thesis stays with the implementation of LDN.

Since not all languages had 100% data completeness, a compromise between reliability and extensiveness needed to be struck in the same manner as with Grambank. Considering the empirical cumulative distribution function plots for the completeness of languages and parameters in ASJP in Figure 3.2, the cumulative probability remains very low up to approximately 50% completeness, after which it begins to rise, with a noticeable stepwise increase from around 75% probability. To the extent of my knowledge, an optimal completeness for reliable mea-

sures has not been suggested in the literature, but the Automated Similarity Judgement Program does not calculate distances for languages with wordlists with fewer than 28 words, which would equal a completeness of 70% (Automated Similarity Judgment Program (ASJP), 2024). Since that program is based on the same data with similar methods, setting such a threshold would be appropriate. The cumulative probability at 70% completeness amounts to about 0.13, meaning that such a threshold would exclude 13% of languages, which is a reasonable trade-off and was thus applied before any similarity calculations took place.

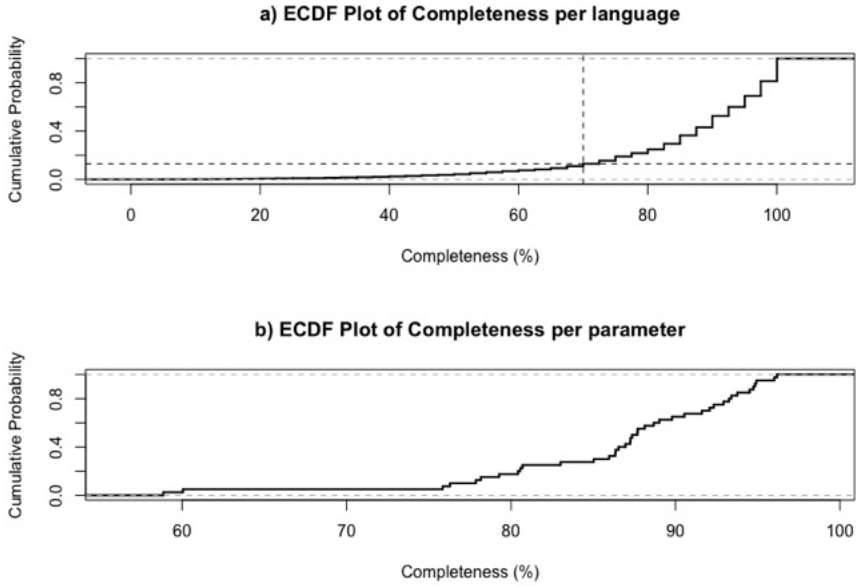

Figure 3.2: ECDF-plots for completeness of a) languages, and b) parameters in the ASJP database, where parameters are defined as the 40-item Swadesh lists introduced by (Bakker et al., 2009).

The analysis was carried out using R (R Core Team, 2024), and the Levenshtein distance was programmatically implemented with the *stringdist* method from the Stringdist-package (van der Loo, 2014), giving a similarity matrix for each country with pairwise lexical similarity between all languages present. Based on this, the mean pairwise similarity in each country was calculated according to Equation (3.2), giving the Mean Lexical Similarty (MLS) in each country.

3.2 Analyzing Similarity Measures

To investigate the behaviour of the similarity measures derived from the typological databases and Glottolog through the equations 3.1, 3.3, 3.5, giving the similarity of phylogeny, phonemic inventory, morphosyntactic features, and core-lexical vocabulary respecitvely, first clustering analysis to verify the measures ability to correctly capture linguistic structure and asses their alignment was employed on a small subset of European languages. Then, the correlation between the measures was analysed based on the largest common subset of languages found in

all databases to understand the associative relationship between the measures and investigate if any measure could be used as a proxy for the other.

3.2.1 Cluster Analysis

To verify that the similarity measures derived indeed capture similarity, clustering methods can be employed and verified by comparing the resulting clusters with known linguistic relationships. One of the most well-explored linguistic areas is Europe, receiving extensive scholarly attention through the 20th century and still to this day (Kortmann & van der Auwera, 2011a). As such, these languages lend themselves well to verifying that similarity measures indeed capture known similarities. The line of reasoning goes that if there are patterns in the measured similarity between languages that are coherent with known patterns of linguistic similarity, then this strongly indicates that the measures capture similarity correctly. This aligns with an approach taken by Gamallo, Pichel, and Alegria (2020), who also used the well-known Indo-European language families to verify that their clusters are resonable, and thus verify that their measures are reliable.

The languages were selected with the prerequisite of being present in all databases, to allow for comparisons of clusters based on the respective similarity measures, and being considered a European language, i.e., being present in Kortmann and van der Auwera (2011a). This gave a list of 36 languages for which similarity measures can be derived according to each respective measure. As such, four complete 36 x 36 similarity matrices with values between 0 and 1 were calculated and turned into dissimilarities by subtracting each value from 1. These were then turned into distance matrices using the *dist* function from the stats package in base R using Euclidean distance, based on which agglomerative hierarchical clustering with complete linkage was performed using the *hclust* function. Agglomerative hierarchical clustering successively joins the datapoints into larger groups according to their proximity, based on which dendrograms can be constructed, creating a hierarchy of similarities, such that two languages present in a subcluster are judged as being more similar to each other than languages outside of the subcluster (Everitt, Landau, Leese, & Stahl, 2011). This allows us to study the hierarchies and substructures visually, with the height where the fusion of branches takes place indicating the similarity of the respective cluster, thus making it useful to validate if the similarity-based groupings correspond to expected groupings based on linguistic knowledge and intuition.

A challenge in using hierarchical clustering is selecting a linkage method to define the distance between subgroupings, of which many exist, all potentially resulting in different clustering results. In ecology, a standard for using an average linking method has tended to be developed, allowing ecologists to handle this variety while achieving satisfactory results (Clarke, Gorley, Somerfield, & Warwick, 2014; Clarke, Somerfield, & Gorley, 2016), such that the average distance between the datapoints in the subgroupings is used. In linguistics, however, signs of a tendency to use Ward's method as a linking method can be seen developing, used recently in e.g., Gamallo et al. (2020); Heeringa et al. (2023). Ward's method analyzes the variance of the clusters instead of the distances directly, and while Gamallo et al. argue that it has proven good performance in many applications, it introduces a level of complexity that reduces its interpretability, directly counteracting its purpose in this thesis. Therefore, the ecological tra-

dition was followed, and average linking was used for all clustering purposes.

To investigate the measures and their reliability, the dendrograms based on the clustering were plotted and visually inspected to find signs of linguistic relationships being captured. In a further step, the clusters based on the respective measure were quantitatively compared with one another to study their alignment. If the different clusterings are aligned, it would indicate that the measures agree about classifications and have a similar ability to identify hierarchical substructures. For that purpose, cophenetic correlation and entanglement were calculated for all pairs of dendrograms using the *dendextend* package (Galili, 2015), a methodology used to compare the results of agglomerative hierarchical clustering, exemplified by Rojo et al. (2024). First, a tanglegram is constructed by placing the two dendrograms opposite each other and drawing a line between all matching leaves. The entanglement is then a measure of the extent of which these lines intersect, such that an entanglement of 1 indicates full entanglement and 0 no entanglement, i.e., the dendrograms are topologically identical. This measure is implemented by considering the dendrograms as two vectors V_1 and V_2, both containing the set of leaves in the orders as given by the respective dendrogram, each given a value from 1 to the size of the tree, such that the same leaf corresponds to the same value in both vectors. The distance between the dendrograms is then defined as

$$d(V_1, V_2) = \sum_{i=1}^{N} |V_{1i} - V_{2i}|^L \tag{3.6}$$

where N is the tree size and L is a parameter indicating the sensitivity to displacement of ranks, and then normalized by the maximal distance, i.e., V_2 is the reverse of V_1, thus giving a measure of entanglement. This is, however, entirely dependent on the layout of the dendrograms, which can be modified by rotating the subgroups, something that isn't given by the clustering in itself. There are various methods for doing this to untangle the dendrograms, and in this thesis, the step2side algorithm was used since it is considered the most effective and is frequently used (Nguyen, Chawshin, Berg, & Varagnolo, 2022). For all measures of entanglement, the default parameter value of L = 1.5 from dendextend was used.

Another approach to measuring the agreement is to compare the cophenetic correlation coefficients. This measure, commonly used in biostatistics, takes the cophenetic similarities, denoting the heights in the dendrogram where any two leaves are joined according to the structure, and correlates them using Pearson's r (Saraçli, Doğan, & Doğan, 2013). While these cophenetic matrices do not say anything about the actual clustering outcome, they can be viewed as containing simplified information about the clustering structure, which in turn is derived from the full information found in the similarity matrices (Everitt et al., 2011). As such, the cophenetic correlation coefficients also indicate alignment between two dendrograms, but focus on the measured proximity between the datapoints and their pairwise distances, instead of on the ordering of the leaves, which is defined by the overall structure of the tree. Therefore, these methods complement each other, and a large r and a low entanglement indicate alignment concerning both pairwise distances and overall structure. To measure the cophenetic correlation, the respective cophenetic matrices for each clustering were derived using the *cophenetic* function in the stats package. The matrices were flattened to only contain the upper triangle, and Pearson's test of correlation was performed using the cor.test function in the stats package.

As such, estimates of Pearson's r with 95% confidence intervals were obtained.

3.2.2 Correlation Analysis

Since the cluster analysis mainly served to investigate the behavior of the similarity measures in a controlled environment of a select few languages, not much can be inferred from those results generally. For that purpose, a correlation analysis was carried out with a much larger sample. To maximize the sample size while retaining comparability across measures, the languages for which all similarity measures could reliably be calculated were selected by intersecting all data sources except Ethnologue, since the relative abundance of languages isn't relevant for assessing similarity. This resulted in 1012 languages, for which pairwise similarities could be calculated, generating four 1012 x 1012 matrices, one for each similarity measure (phylogenetic, lexical, morphosyntactic, and phonemic inventory). Each similarity matrix was flattened, and the upper triangle was extracted, giving four vectors, each containing all unique pairwise similarities according to the respective linguistic domain. As such, the index of each vector refers to a pair of languages.

The distributions of the pairwise similarities were then explored using histograms, and for each type of measure, a normal Quantile-Quantile plot was generated (Wilk & Gnanadesikan, 1968). Due to all measures except morphosyntactic similarity heavily deviating from a normal distribution, non-parametric correlation tests in the form of Spearman's ρ were utilized (Spearman, 1904). As such, the ranks of the language pairs according to the respective measure were compared and correlated as

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.7}$$

where d_i is the difference in rank between any language pair for the two measures correlated and n the number of observations. As such, a value of 1 indicates complete agreement between the ranks and -1 a complete disagreement, with increasing absolute values indicating a larger magnitude of the effect. This was implemented using the cor.test function from the stats package, and each statistic was reported together with its p-value. All 6 combinations of the measures were plotted as scatter plots with a linear smoothing, and the nature of the relationship between the measures was assessed visually.

3.3 Analyzing Diversity Measures

To analyze diversity, the distribution of mean pairwise similarities on a country level derived from the respective databases were investigated using first violin plots and then maps, coloring each country according to the mean similarity for a given dimension, to find spatial patterns in structural diversity. Similar to the correlation analysis introduced in the previous section, the association between the mean similarities was estimated using Spearman's ρ as defined in Equation (3.7). Since the number of datapoints where restricted to the number of countries for which similarity measures could be derived, only 217 datapoints were available for each measure. Therefore, it was deemed necessary to quantify the uncertainty of the estimate, which

was done using bootstrapping with 1000 replicates to construct 95% confidence intervals.

Then, the global diversity was assessed by implementing the Leinster-Cobbold framework as outlined in Chapter 2. Since Greenberg suggested the use of lexical similarity, and the analysis of the databases showed that the ASJP database had the most extensive language and speaker coverage, lexical similarity implemented as normalized Levenshtein distance between core vocabulary in Swadesh lists from the ASJP database was used. For each country, the non-naive diversity measure was calculated for q = 0, 1, and 2 according to Equation (2.5) with an M x M similarity matrix Z where M is the total number of languages found in the country according to both the baseline data from Ethnologue and the Joshua project as described in Section 3.1.1 and in the ASJP database, such that

$$ {}^{q}D^{Z}(p) = (\sum_{i}^{M} p_i (Zp)_i^{q-1})^{\frac{1}{1-q}} $$

where

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

For each country, the relative abundances p_i were calculated such that the relative abundance of language i, with the number of speakers s_i given as

$$ p_i = \frac{s_i}{\sum_i^M s_i}. \tag{3.8} $$

These are only estimates of relative abundance of language, since we lack data on multilingualism, thus presenting a substantial limitation. The real linguistic situation is most likely radically different, but with the given data, only this way of modelling relative abundance is possible. The non-naive diversity was calculated identically to the above, with the exception that an N x N identity matrix I was used instead, where N is the number of languages found in the given country according to the baseline data.

These measures were plotted on maps to analyze global patterns, and the change in diversity of single countries of interest for varying q was investigated using diversity profiles, as suggested by (Leinster & Cobbold, 2012). These are simply line plots with ${}^{q}D^{Z}(p)$ plotted as a function of q for a range of values of q to show how the diversity structure in a given setting behaves as the sensitivity to dominant languages varies.

Since the number of data points on which diversity is calculated varies between the non-naive and naive case due to the limited coverage of the ASJP database, the measured non-naive diversity for a given country will be biased by an error $\epsilon_0 = N - M$, expressing missing coverage in the ASJP. While this bias is unavoidable for assessing perceived global diversity since this is restricted by the available data, the problem can be circumvented when studying the impact of incorporating similarity. By using an M x M identity and similarity matrix to determine both the naive and non-naive diversity, all measures will utilize the same relative abundance measures, removing the bias when measuring the effect of setting $I = Z$. To quantify the effect that accounting for similarity has, each country was ranked according to its diversity in both the naive and non-naive case for q = 2, again to stay close to Greenberg's envisioned implementa-

tion, and then the ranks were correlated using Spearman's ρ. If Spearman's ρ = 1, then both the naive and the non-naive assign countries the same rank, and the impact is minimized. If Spearman's ρ = -1, then the measures assign the countries the inverse ranks, and the impact is maximized. Therefore, this becomes a measure of effect sizes bounded between 1 and -1, where 1 equals no effect, and -1 the largest possible effect. Furthermore, the impact was investigated qualitatively by plotting the diversity ranks against each other in a scatter plot, together with the diagonal y = x representing the expected relationship between the ranks if an effect on the ranking was non-existent. That way, countries deviating from this line would signal either a large gain or loss as a result of including similarity in the measure, thus identifying the countries most heavily impacted by accounting for similarity.

4 Cross-Linguistic Databases

In this chapter, the analysis of the databases as described in the previous chapter is carried out. First, the speaker data is presented to give an idea of how linguistic data is distributed globally. Then, the speaker data is used as a baseline to asses the coverage and representativeness of the databases, one at a time.

4.1 Language and speaker data

To measure linguistic diversity, the most critical aspect is the number of languages present, the richness. If there is no notion of how many languages are spoken in a setting, little can be said about its diversity. Similarly, the extent to which a language is used is of almost equal importance; a language that is used only by a small fraction is not as defining as a language used by a relative majority. As such, assessing the data available on languages and speaker populations first is vital for understanding linguistic diversity.

4.1.1 Languages

Firstly, it should be acknowledged that the data presented here as a baseline for assessing the linguistic situation globally is not a perfect representation, since estimating language data is notoriously difficult and associated with conceptual problems such as how to define two languages for example, as discussed in Section 3.1.1. While Ethologue states that 7159 languages are living today (Eberhard & Fennig, 2025) and Glottolog asserts that there are 8223 languages in use (Glottolog, 2024), the baseline data in this thesis accounts for 6757 languages in 239 countries and territories[1]. It thus naturally follows that no result derived in this thesis will be a perfect reflection of the true global linguistic diversity today, but it is the best estimate given the available data.

These 6757 languages are spread unevenly across the world, with the most languages found in Africa (31.43 %), Papunesia (30.46 %) and Eurasia (22.83%), with substantially fewer languages in Australia (2.05%) and North and South America (7.47 & 5.76%), as can be seen in Figure 4.1. This is roughly similar to the estimate for 6409 languages by Hammarström (2016), placing 28.8% in Africa, 28.0% in Papunesia, 22.2% in Eurasia, 8.7% in North America, 7.6% in South America, and 4.5% in Australia. It is important to keep in mind that languages can be used across larger areas than their macroarea, which does not categorize where the language "lives" but rather to where it originates, giving 6 geographical areas of concentrated linguistic richness (Hammarström & Donohue, 2014).

[1] For a full list of countries/territories, see Table 7.1 in the appendix

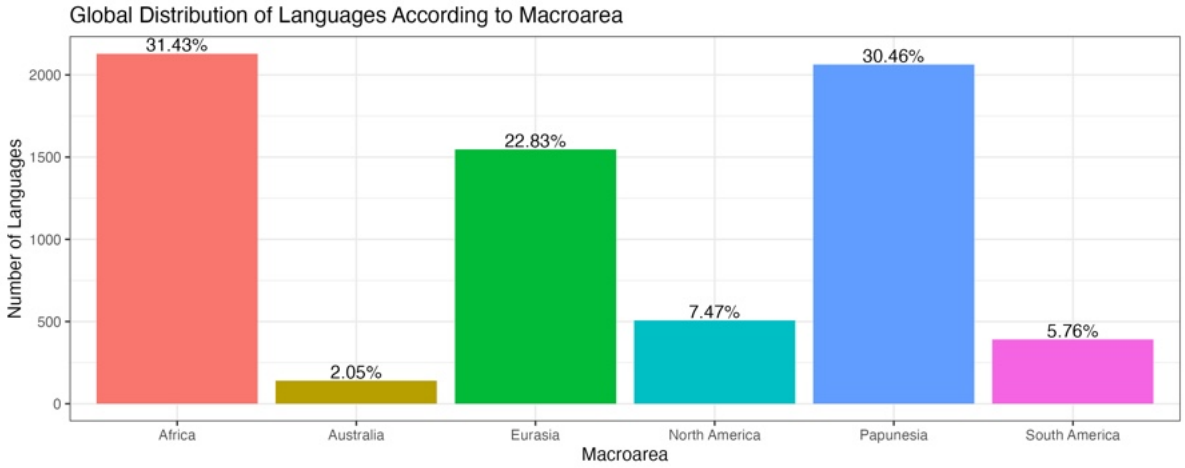

Figure 4.1: Barplot showing the global distribution of language numbers across Macroarea.

4.1.2 Language Families

Classification Category	Number of Languages	Speakers
Artificial Language	2	101500
Bookkeeping	86	3 574 205
Isolate	107	1924381
Mixed Language	3	21400
Pidgin	5	151446
Sign Language	144	28 105 317
Speech Register	6	49 550
Unattested	16	230 195
Unclassifiable	5	83 350
Total	374	34 241 344

Table 4.1: Table showing the number of languages and speakers belonging to non-genealogical classification categories from Glottolog.

In the speaker data, 214 distinct language families according to the classification of Glottolog are found as well as 107 isolates seen in Table 4.1. To get an idea of how much phylogenetic diversity can be captured by this data, one can compare it to the 247 families and 183 isolates identified by Glottolog (2024) in total. Furthermore, the use of ISO-codes to define languages creates some mismatch, as seen by the 71 languages categorized as *bookkeeping* by Glottolog, which means that they do not classify them as languages anymore. The reason for such a reclassification mainly comes down to the definition of language employed by Ethnologue and Glottolog, as varieties mutually unintelligible from all other varieties. An example is given by Glottolog pertaining to the language Yarsun, that was claimed by Ethnologue to be a distinct language and thus classified as such in Glottolog. In the light of new research however, the evidence of Yarsun being distinct from all other languages was deemed insufficient,

and the Glottocode was classified as bookkeeping. That way, Glottocodes are never retired, which might be the case of ISO-codes (Glottolog, 2024). There are also 14 languages considered unattested, meaning that the data attesting to them has vanished, which is in contradiction to the speaker data claiming there are speakers of them. All in all, these categories are non-genealogical, and sum up to 374 languages and approximately 34 000 000 speakers. As a direct consequence, phylogenetic similarity can't be assessed for this part of the dataset. Important to notice are also the 144 sign languages, covering more than 28 million speakers, which, due to their inherently different nature as a signed as opposed to spoken language, are difficult to conceptualize from a similarity perspective. For example, sign languages have concepts analogous to phonemes with parts of gestures forming minimal pairs (Goldin-Meadow & Brentari, 2017), but how could these concepts be compared to the phoneme inventories of spoken languages? Similarly, lexical similarity between signed languages has been operationalized by Börstell, Crasborn, and Whynot (2020), but going a step further and combining the spoken and signed domains would be necessary to analyze communities where they exist alongside each other, which goes beyond the scope of this thesis.

Considering the distribution of language families across macroareas as seen in Figure 4.2, the largest number of language families is found in Papunesia, being the home to 33.19% of all language families excluding isolates. This is almost twice that of the second most family-dense macroarea, South America. The fewest language families are found in Australia with 7.52% of all language families, followed by Eurasia with 10.62%.

Looking at the distribution of families within each macroarea as seen in Figure 4.2, Papunesia is dominated by the Austronesian language family, with more than half of all languages belonging to it. A similar pattern is observed in Australia, which is dominated by the Pama-Nyungan family, and in Africa with the Atlantic-Congo family. On the other hand, the regions of North and South America and Eurasia are more split, with the Sino-Tibetan and Indo-European language families together dominating Eurasia.

4.1.3 Speakers

Finally, the data covering the speakers is to be considered. Viewing Table 4.2, the number of speakers is tallied up to approximately 7.63 billion. Ideally, this should approximately equal the world population to best capture linguistic diversity. In 2022, the year from which the Ethnologue data stems, the world population amounted to 8.021 billion according to Worldometer (2024), and thus this estimate is reasonable, especially considering that language data is collected asynchronously, which with an increasing world population would result in underestimations of total speaker numbers. At the same time, L1 speakers are not explicitly included in the data, making the estimate even more imprecise, but still, the observed number is close enough to be considered an accurate representation of global linguistic usage. To put the extensiveness of the data into comparison: Bromham et al. (2022) in their investigation of global linguistic diversity included 6511 languages, of which 6171 had L1 speaker numbers not equal to zero, summing up to 3 091 724 926 L1 speakers, in addition to 9 world languages which were coded as separate variables[2].

[2] these numbers were calculated based on the supplementary material provided by Bromham et al. (2022)

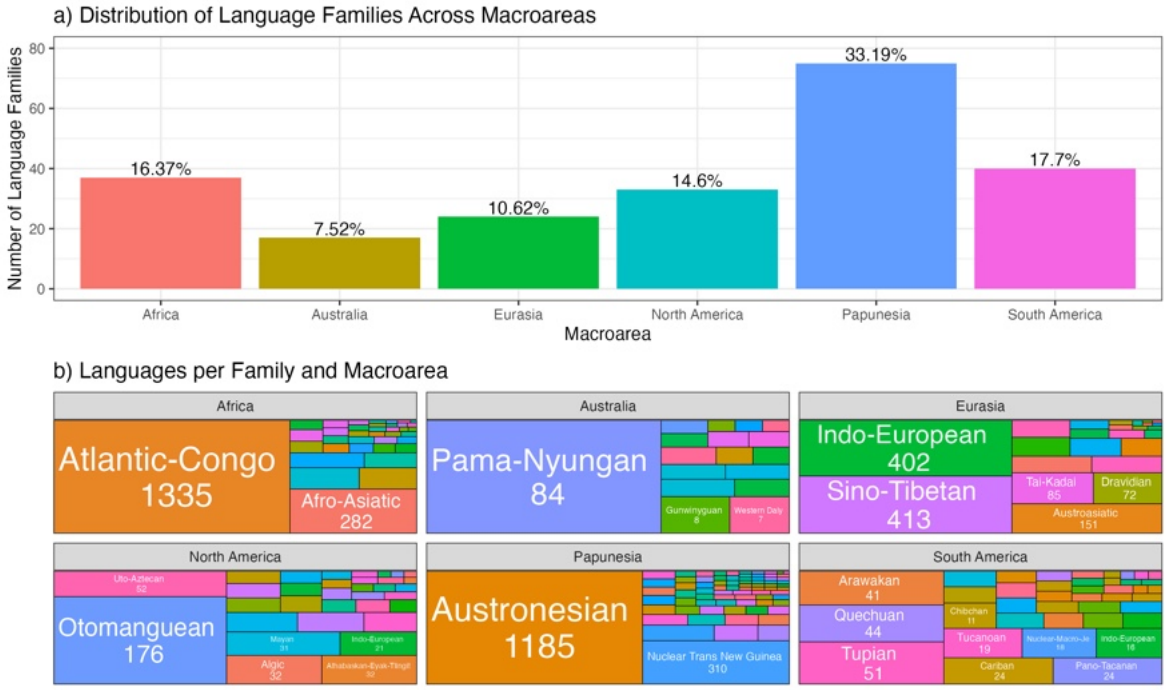

Figure 4.2: a) Barplots showing the distribution of language families across macroareas and b) a treemap depicting the distribution of languages and language families across macroareas. Each rectangle represents a language family. The area of the rectangle is proportional to the number of languages belonging to that language family in that macroarea. The number in each rectangle indicates the number of languages belonging to that language family.

The distribution of speakers is uneven, with approximately 78 % originating in Eurasia, which can be presumed to partially correlate with the large populations of e.g. India and China, which are located in Asia and to a large extent speak languages from the region. Also, many languages originating in Europe are dominant in former colonies, e.g., Australia and Brazil, which explains why the number of speakers of languages belonging to these macroareas is low. Especially interesting are the speaker numbers in Australia; although demonstrating a high richness of languages, the relative abundance is abysmal with only 64 163 estimated speakers of languages belonging to the macroarea.

Considering the distribution of speakers within each macroarea according to the categories in Table 3.1 with 25% of all languages in the respective category as seen in Figure 4.3, two different profiles emerge. In Eurasia and Africa, most languages have many speakers (71 900 - 971 424 800), with diminishing proportions as the number of speakers decreases. In Australia, Papunesia, and the Americas, the opposite relationship is observed. Most languages belong to the smallest speaker category (1 - 1400) with diminishing proportions as the number of speakers increases. This is most prominent in Australia, with 90.65% of languages belonging to the category with the fewest speakers, and no language belonging to the category with the most speakers. In Papunesia and the Americas, it is, on the other hand, not as extreme, with about

Macroarea	Languages	Speakers	Speakers (%)	Median Speakers per Language
Africa	2126	1 348 242 826	17.66	42 750
Australia	139	64 163	0.00084	46
Eurasia	1535	5 932 359 233	77.71	30 000
North America	505	29 389 183	0.38	3690
Papunesia	2063	298 676 980	3.91	2400
South America	389	25 328 804	0.33	1800
Global	6757	7 634 061 189	100	10000

Table 4.2: Table showing the number of languages per macroarea as well as the number and percentage of speakers.

8-10% of the languages in all three regions belonging to the category with the most speakers and a somewhat more even distribution across the fewer speaker categories.

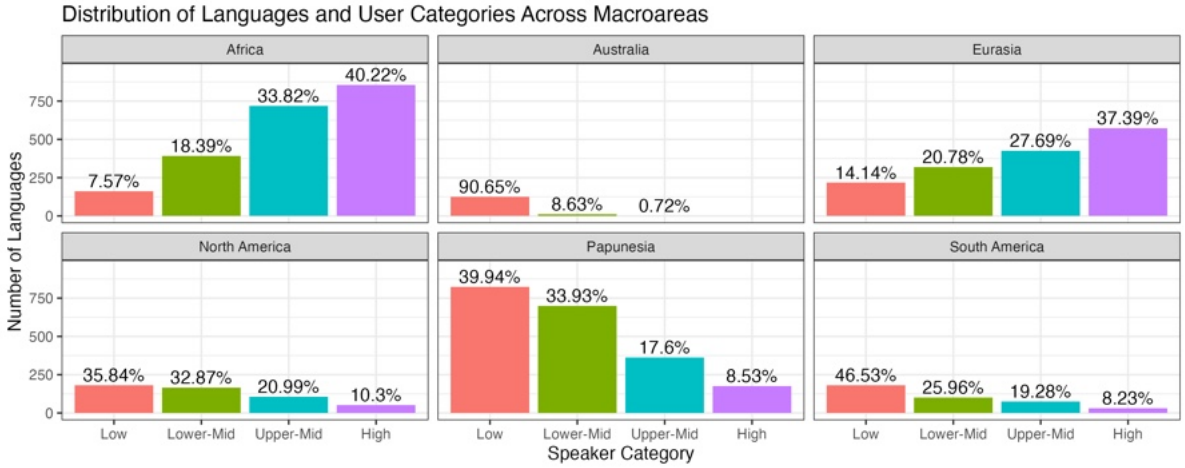

Figure 4.3: Barplot showing the distribution of languages according to speaker across macroarea, with 25% of all languages belonging to each category.

If one considers the number of speakers belonging to different language families as seen in Figure 4.4, it is interesting to note that Indo-European languages are more prominent than they were when only the number of languages was considered. In North America and Eurasia, the relative majority uses Indo-European languages, while in South America and Australia, much larger proportions are Indo-European, if one compares with Figure 4.2. The only macroarea where a substantial proportion of speakers do not speak Indo-European languages is Papunesia, which instead is dominated by the Austronesian language family.

Countries

It is also prudent to consider the distribution of languages on a country level. The median number of languages present in a country is 23 (mean = 51.30, IQR = 46, SD = 96.59), with the

Figure 4.4: Treemaps depicting the distribution of speakers and language families across a) the entire world and b) macroareas. Each rectangle represents a language family, with the area of each rectangle being proportional to the percentage of speakers of languages belonging to that family.

countries with the most languages being Papua New Guinea (848), Indonesia (712), and Nigeria (522). Among all countries, only 8 countries house a single language according to the data, namely San Marino, Saint Helena, Falkland Islands, Niue, Tokelau, the Vatican, Jersey, and Pitcairn.

While the number of languages and the distribution of speakers are elaborated further as measures of diversity in Section 5.2, it is important to acknowledge that languages have different presence on a global scale. While the macroregion division uniquely assigns each language to a specific region for typological purposes, some languages are used globally, being present in multiple countries. The same way that a language with many speakers is more important for assessing diversity within a country than a small language, a language present in multiple countries will be more important for assessing diversity globally compared to a language only present in one country. A table with the 20 most important languages in that regard, measured by the number of countries in which these languages are present, can be seen in Table 4.3. Here, the classic "world languages" (Ammon, 2010) are found, such as English, present in 173 countries, followed by French in 118 and Chinese in 104. However, not all languages are what would be categorized as world languages, such as Turkish and Greek, but are still the eighth and ninth most extensively spread languages according to the data, which could be due to large diasporas due to mass migration, as in the case of Turkish migration to Europe (Açıkalın, 2024). Since these languages are present in many different countries, the coverage of these in cross-linguistic databases is paramount, since they will have a larger impact on the global linguistic diversity relative to less interconnected languages.

	Language	Family	Macroarea	No. speakers	No. Countries
1	English	Indo-European	Eurasia	366538940	173
2	French	Indo-European	Eurasia	70940000	118
3	Chinese, Mandarin	Sino-Tibetan	Eurasia	971424800	104
4	Arabic, Levantine	Afro-Asiatic	Eurasia	46967800	88
5	German, Standard	Indo-European	Eurasia	66610700	76
6	Spanish	Indo-European	Eurasia	451044090	73
7	Russian	Indo-European	Eurasia	138431120	71
8	Turkish	Turkic	Eurasia	76112100	71
9	Greek	Indo-European	Eurasia	11268800	68
10	Hindi	Indo-European	Eurasia	609305600	68
11	Italian	Indo-European	Eurasia	40440200	60
12	Portuguese	Indo-European	Eurasia	224495870	56
13	Tagalog	Austronesian	Papunesia	44629680	56
14	Armenian, Western	Indo-European	Eurasia	3671480	52
15	Korean	Koreanic	Eurasia	81698500	50
16	Ukrainian	Indo-European	Eurasia	31580800	50
17	Japanese	Japonic	Eurasia	120905700	48
18	Polish	Indo-European	Eurasia	42957300	46
19	Romanian	Indo-European	Eurasia	24928500	45
20	Bulgarian	Indo-European	Eurasia	7468000	41

Table 4.3: Table showing the 20 languages found in the most countries together with their language families, macroareas, and speaker numbers.

4.1.4 Summary

To conclude this section, the baseline data have been analyzed along the dimensions of languages, speakers, language families, and macroareas to give an idea of the global distribution of languages. It has been shown that languages are unevenly distributed across macroareas, especially concerning speaker numbers. A small number of language families dominate, especially the Indo-European, when one considers relative abundance. Generally, the macro areas of Africa and Eurasia are thriving, with many languages and speakers, while Papunesia shows an incredible richness of both languages and language families, but smaller proportions of languages with many speakers.

4.2 Database Coverage

Having shown the global composition of languages and speakers in data from Ethnologue and the Joshua project in the previous section, the extent to which this linguistic data is represented in the selected cross-linguistic databases will now follow. The exact coverage per country in the databases can be found in Table 7.1 in the Appendix.

4.2.1 PHOIBLE

As mentioned in section 3.1.3, PHOIBLE contains data on 2186 languages. As seen in Table 4.4, of these 2186 languages, 1820 languages also have data on speakers, resulting in a data loss of 366 languages (17%). Following the discussion in Chapter 3, a data loss is expected due to only considering living languages with speaker data and a possible mismatch in language classifications between Ethnologue and Glottolog. Using the 6757 languages from Ethnologue as a baseline gives a coverage of 26.93%, thus Phoible can be considered limited in its capability of capturing linguistic richness. On the other hand, 160 language families and 51 isolates[3] are present, giving a coverage of 74.76% not counting the isolates, and an even larger proportion of speakers are covered by PHOIBLE, namely 82.36%. As such, PHOIBLE manages to capture relative abundance very well.

Macroarea	Languages	Language Families	Speakers
Africa	703	35	937 084 445
Australia	117	16	38 021
Eurasia	417	23	5 243 890 027
North America	113	29	10 343 941
Papunesia	182	28	232 337 691
South America	288	37	16 890 968
Global	1820	160	6 440 585 093
Coverage (%)	26.93	74.76	84.36

Table 4.4: Table showing the statistics of languages present in the PHOIBLE data. The coverage is given as the percentage of instances found in PHOIBLE compared to the baseline.

Macroarea

If one considers the distribution of languages in PHOIBLE across macroarea as seen in Figure 4.5, it turns out that PHOIBLE is underrepresenting Papunesian languages, which only make up 10% of languages in the PHOIBLE data, but 30% in the baseline, giving an absolute percent difference of about 20%. On the other hand, South American languages are overrepresented, amounting to an absolute percent difference of about 10%, making up a larger portion of the PHOIBLE data compared to the baseline. Further overrepresentation is observed for Africa and

[3] In addition, PHOIBLE contains one language classified as pidgin and one language which is kept by Glottolog for bookkeeping, indeed indicating a small mismatch in language classification.

Australia (7% and 4% respectively) while only minimal differences in representation are found for Eurasia and North America.

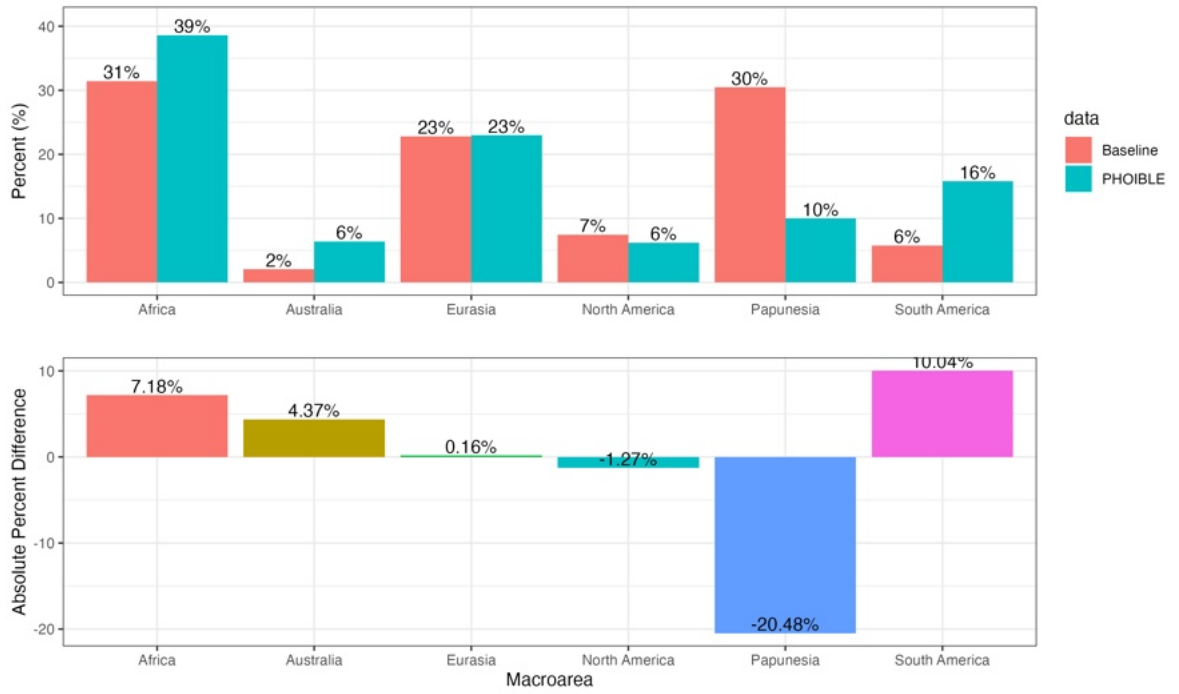

Figure 4.5: Barplot comparing the distribution of languages across macroareas in the PHOIBLE and the baseline data. Absolute percent difference shows the size of the effect, with a negative value indicating underrepresentation in PHOIBLE.

Speaker Numbers

Analyzing the representation of languages with varying numbers of speakers according to the categorical speaker category variable introduced in Table 3.1, it seems that PHOIBLE has an overrepresentation of languages with many speakers. Languages with more than 71 900 speakers make up 25 percent of the languages in the baseline, but when considering the languages included in PHOIBLE, they make up 41% of languages, giving an absolute percent difference of 16.32% as seen in Figure 4.6. On the other hand, languages with very few speakers (less than 1125) are somewhat better represented with only 3.31 percentage points underrepresentation, while the largest underrepresentation (8.84%) is seen for Lower-Mid languages (1400 - 10 000). Due to the overrepresentation of languages with many speakers, it can be suggested that PHOIBLE is better aligned with measures taking relative importance into account in contrast to measures only considering richness, while noting that the representation of speakers of different size-groups in themselves is not particularly good in PHOIBLE.

Looking further into which languages aren't included in PHOIBLE and contributing the most to loss in speaker coverage, the treemap in Figure 4.7 gives an idea of which languages these are. Since Indo-European and Sino-Tibetan are the most prolific languages generally, it is no

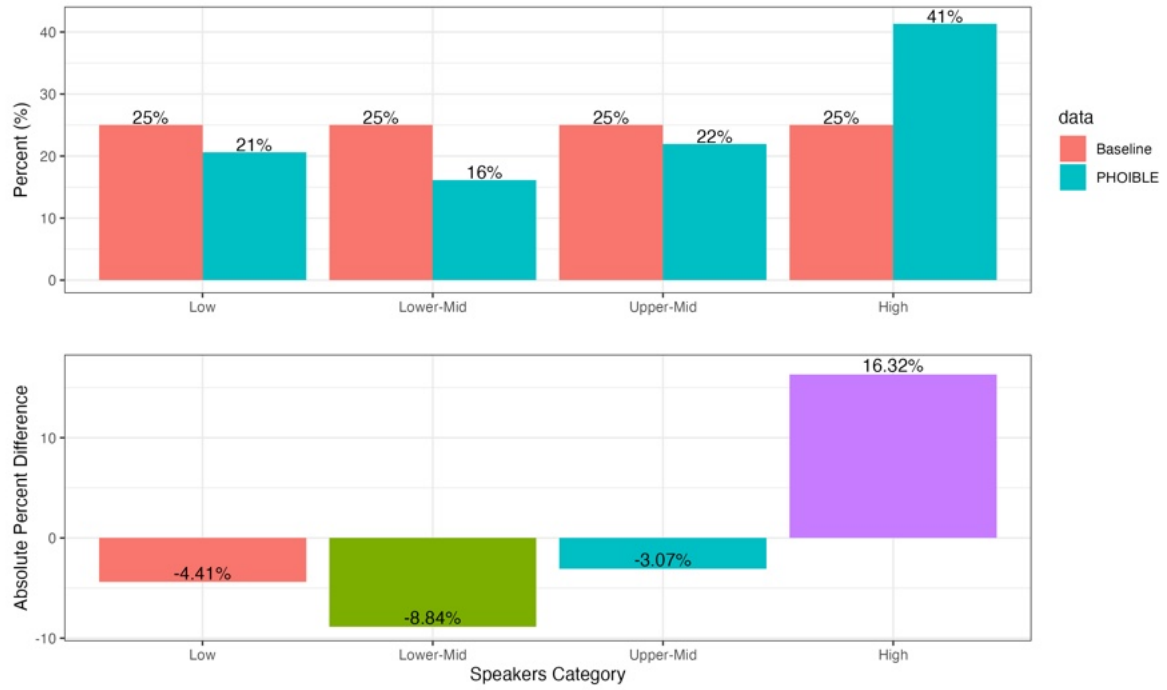

Figure 4.6: Barplot showing how the distribution of languages categorized into quartiles based on speaker numbers in the PHOIBLE data. A negative absolute percent difference indicates underrepresentation in PHOIBLE.

surprise that the speakers missing belong to languages of these families. For example, many languages belonging to the Chinese branch of the Sino-Tibetan languages are found, such as Xiang and Jinyu. From Asia, languages such as Punjabi, Western and Deccan are missing. Only a few languages from the European part of Eurasia are missing here to a large extent, e.g., Bavarian and Sicilian; languages commonly considered dialects roofed by the respective standard varieties of German and Italian (Giacalone, 2016; Rowley, 2011), but also standardized languages such as Serbo-Croatian and Albanian. Furthermore, many Arabic varieties roofed by standard Arabic not being covered by PHOIBLE contribute heavily to the missing number of speakers, such as Algerian Arabic and Levantine Arabic.

Countries

Finally, the coverage of languages in PHOIBLE per country is to be considered. It naturally follows that if a large proportion of the languages or speakers in a country is covered, then any diversity measures based on phonemic similarity from PHOIBLE will be a better estimate of the true diversity of that specific country. As a result, the lower the coverage, the more unreliable and less representative the estimate will be.

If one considers the coverage of number of languages in PHOIBLE as seen in Figure 4.8, the median coverage in the 239 countries included in the data is 66.67% (mean = 62.74, IQR = 28.19, SD = 23.19), meaning that on average, 66.67% of languages in a country for which language

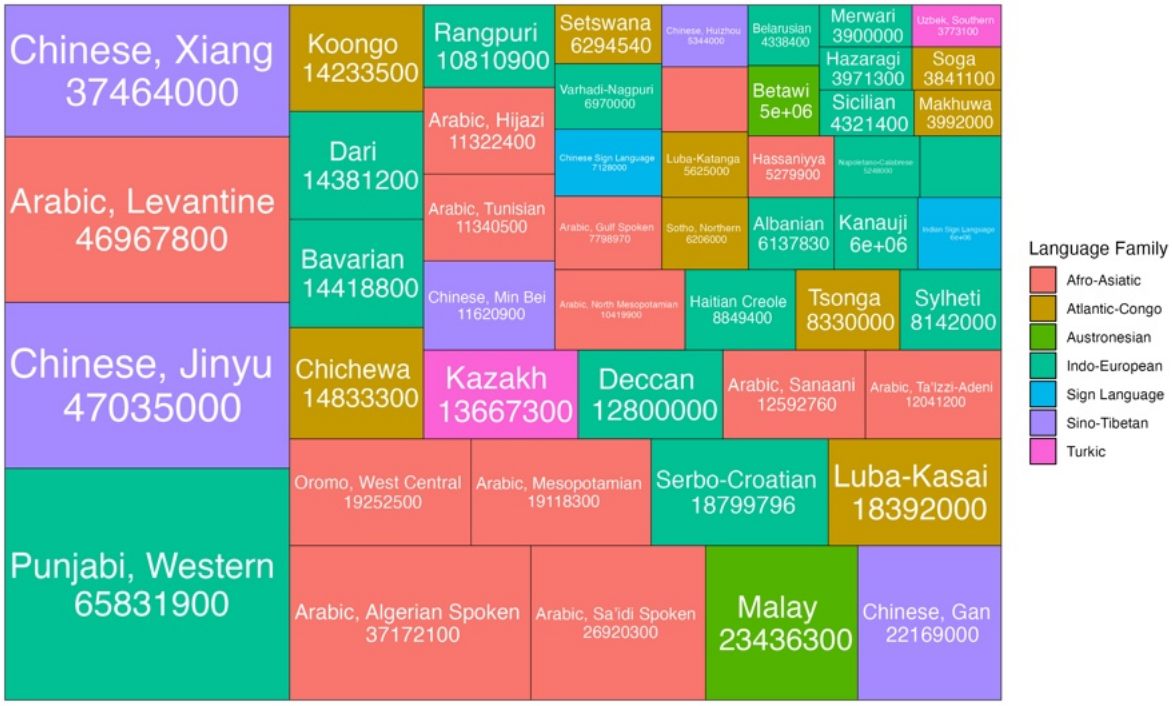

Figure 4.7: Treemap showing the 50 languages with the most speakers missing in PHOIBLE colored according to language family. The area of the rectangles is proportional to the number of speakers.

speaker data exist, there also are data on the phonemic inventory. The single smallest coverage value is 0%, which applies to only 4 countries (Niue, Pitcair, Tokelau and Tuvalu), where thus the phonemic diversity can't be assessed. Meanwhile, however, 21 countries have 100% of languages covered, 5 of which only house a single language due to which any diversity measure per default will be zero. The coverage of the number of speakers is as previously concluded better in PHOIBLE overall, which also holds up on a country level as seen in Figure 4.8. The distribution has two peaks: one at almost full coverage, and one very small at almost no coverage. Since 4 countries has no languages covered in PHOIBLE, it follows that they also lack speaker coverage but there are also a few countries with coverage below 25%. The median coverage of speakers is 89.10% with a mean of 70.61% (IQR = 52.57, SD = 34.62), a difference which can be attributed to the larger number of countries without any coverage pulling the mean down.

Considering the geographical distribution of the coverage by plotting a world map with countries colored according to the coverage in PHOIBLE as seen in Figure 4.9, one can see that the countries of South America enjoy good coverage in PHOIBLE, while North America is somewhat neglected, especially Mexico. This is in agreement with the previous conclusion that South American languages are overrepresented in PHOIBLE. Russia and many Central Asian countries tend to have somewhat better coverage together with Australia and New Zealand. South and East Asia and the Middle East have a very low coverage, together with central Africa. Europe, on the other hand, has very good coverage.

Even more interesting is to consider the speaker coverage as seen in Figure 4.9 b. As pre-

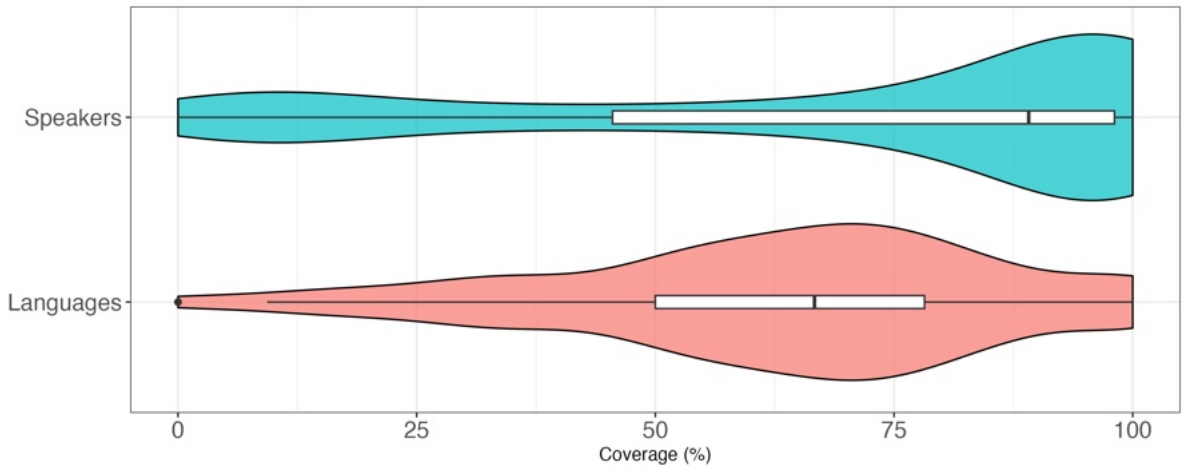

Figure 4.8: A violinplot showing the distribution of language and speaker coverage in the PHOIBLE data on a country level using kernel density estimation and a boxplot.

viously established, PHOIBLE covers approximately 84% of language speakers worldwide as compared to roughly 27% of languages. It is thus no surprise that the coverage on a country level is higher regarding speakers. The increased coverage is most prominent in North America and Europe, where most countries have almost full coverage apart from a few exceptions such as Suriname, Austria, and Belarus, where the coverage is almost nonexistent. In South East Asia, the coverage also improves considerably, and to some extent in Africa as well where countries such as Niger and Somalia enjoy almost full speaker coverage. However, large swaths of North Africa and the Middle East have spots of very low speaker coverage, which is likely due to many spoken varieties of Arabic not being included in the database. Furthermore, it is clear that the underrepresentation of Papunesia has a large impact on the coverage in Oceania, where Papua New Guinea and many island nations have no speaker coverage at all.

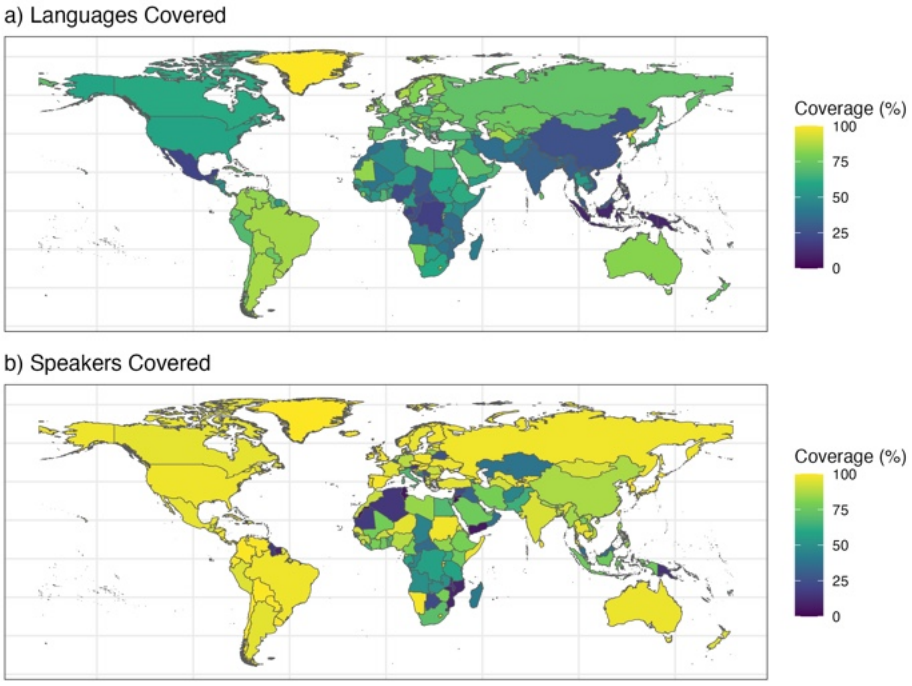

Figure 4.9: World map showing the coverage in PHOBILE of a) languages and b) speaker populations. The coverage is calculated as the percentage of languages and speakers present in PHOIBLE compared to the baseline.

Summary

To conclude the section, the coverage of PHOIBLE is generally more extensive concerning speakers than language numbers, suggesting that similarity measures derived from PHOIBLE can reliably be combined with diversity measures taking relative abundance into account as opposed to only richness. The database also shows some geographical bias with the macroarea Papunesia being heavily underrepresented and South America overrepresented with regard to language richness. On a country level, the Americas, Europe, and Central Asia are where most countries with high coverage can be found, while countries in Africa, the Middle East, and Southeast Asia to a larger extent lack data. For these regions, estimates employing similarity measures based on PHOIBLE data will therefore be more unreliable.

4.2.2 Grambank

Compared to PHOIBLE, Grambank has a somewhat higher coverage of languages and language families as seen in Table 4.5. Grambank includes 2149 languages with speaker data, thus covering 31.76% of languages as compared to PHOIBLE's 26.91%, while also extending its coverage to more language families 196 (88.15%) compared to PHOIBLE (73.93%). Together with 195 language families, 72 isolates are represented as well as 2 sign languages, namely Finnish and Japanese Sign Language. However, Grambank seems to have a relatively low coverage of

speakers. Grambank only covers 64.60% of available speaker data, whereas PHOIBLE covers 84.36% of speakers. Considering that Grambank covers more languages, it can be suggested that Grambank primarily includes languages with smaller speaker communities, thus making it less suitable for linguistic diversity measurements taking relative abundance into account. If one also takes into account the 25% cut-off threshold for data completeness to calculate similarities, which was discussed in Chapter 3, the coverage isn't affected tremendously. 141 languages, 1 language family, and approximately 19 million speakers are disregarded. Therefore, and since similarity measures require sufficient data completeness, the following analysis of the database coverage is based on the 2108 languages with at least 25 % data completeness and their associated speaker numbers.

Macroarea	Languages	Language Families	Speakers	Completeness (%)
Africa	536	35	705 633 477	71.48
Australia	77	17	36 292	62.12
Eurasia	482	23	4 088 397 418	83.70
North America	180	31	19 605 658	83.67
Papunesia	688	63	257 234 516	74.00
South America	186	37	13 705 421	72.01
Global	2149	196	5 084 612 782	75.80
Coverage (%)	31.80	91.59	66.60	-
> 25% Completeness	2108	195	5 065 185 564	76.9
Coverage (%)	31.12	91.12	66.34	-

Table 4.5: Table showing the statistics of languages present in Grambank. The coverage is given as the percentage of instances in Grambank relative to the baseline.

Macroareas

Looking into the representation of Macroareas in Grambank as seen in Figure 4.10, a fairly balanced representation is found. While PHOIBLE heavily underrepresents languages from Papunesia, Grambank contains an almost equal proportion of languages from the macroarea compared to the baseline. A small tendency to overrepresentation is found for South America with an absolute percent difference of 2.78%, with the only clear underrepresentation being Africa, observing an absolute percent difference of -6.65%. As such, it follows that Grambank provides a great representation of languages from the linguistic macroareas of the world and counts as very representative in that regard, especially for the language of Eurasia.

Speaker numbers

If one looks at the groupings of languages according to the size of their speaker communities, Grambank has a quite even distribution, representing all groups fairly well. There is some overrepresentation of languages with many speakers and an underrepresentation of languages with very few speakers, but for the categories in between, lower- and upper-mid number of

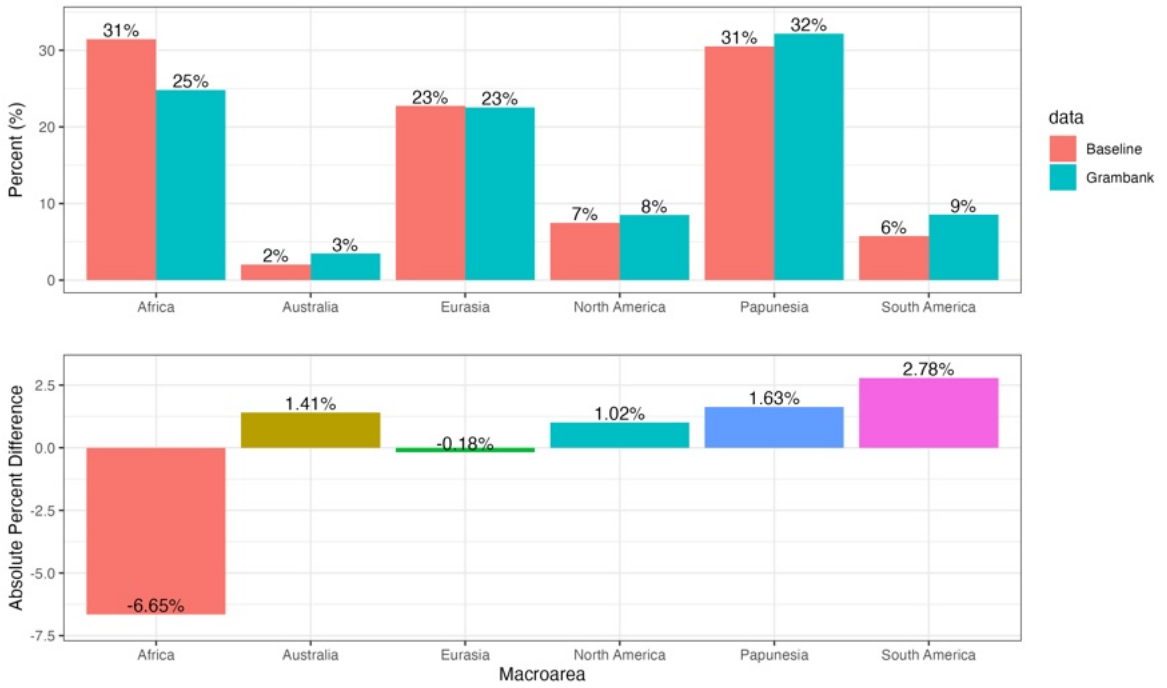

Figure 4.10: Barplot showing the difference in the distribution of languages across macroareas grambank relative to the baseline. A negative value indicates underrepresentation in Grambank.

speakers, the proportion of speakers is almost identical to the baseline. This can be contrasted to PHOIBLE, which was heavily overrepresenting languages with many speakers. Again, this shows that Grambank works very well as a large and balanced sample of the languages of the world if looking at richness, but since the human population isn't evenly distributed across the world's languages, many speakers are disregarded, which is problematic in the context of measuring linguistic diversity. Arguing that relative abundance should, like similarity, be taken into account to accurately describe the diversity of a setting, it would even be positive if languages with more speakers are overrepresented, since they are given a heavier weighting in a diversity context.

Having established that Grambank has a relatively low speaker coverage, it is relevant to consider which languages aren't included in Grambank. In Figure 4.12, the 50 most spoken languages according to the baseline that aren't included in the Grambank data can be seen. Based on this, it seems that predominantly Indo-European languages are missing, especially from Asia. Bengali, which constitutes an important language on the Indian subcontinent, is missing, as well as languages such as Urdu and Gujarati. In addition, two prominent European languages, Standard German and Spanish, are missing, which could potentially limit the usefulness of Grambank in some regards. Since Spanish is a world language with many speakers in multiple countries, it is part of many language communities as a lingua franca. Not only does its absence thus hamper the possibility of incorporating structural diversity with models of lingusitic diversty, but also limits the exploration of how e.g. the grammatical structure of Spanish

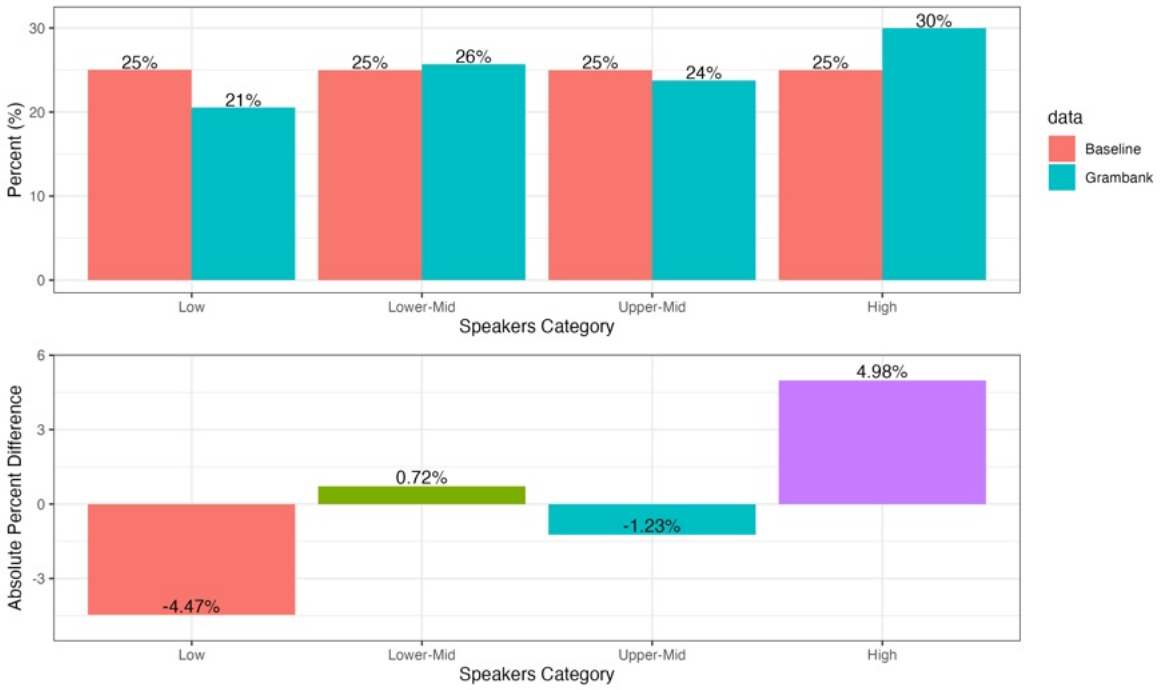

Figure 4.11: Barplot showing the difference in the distribution of languages in the baseline and Grambank data across categories based on the number of speakers.

might interact or has interacted with other languages in communities where it acts or has acted as lingua franca and the impaction this could have on linguistic diversity.

Furthermore, Grambank, similar to PHOIBLE, is missing some Chinese languages and spoken varieties of Arabic from the Afro-Asiatic family, which have many speakers. In addition, a few prolific African languages from the Atlantic-Congo family are missing, such as Yoruba and Fulah, as well as Turkic languages such as Kazakh and South Azerbajani.

Countries

On a country level, Grambank has a distribution of speakers and languages that is somewhat similar to that of PHOIBLE. Considering the violin plot in Figure 4.13, the speaker coverage follows an hourglass shape, peaking both at top and the bottom, meaning that there are mainly two groups of countries: those with almost no speaker coverage and those with near full coverage. The median speaker coverage is 67.68 % (IQR = 77.21, SD = 37.68) with a mean of 57.48%, so on average, the speaker coverage is lower compared PHOIBLE (median = 89.10%, mean = 70.61%). The language coverage has a prominent peak at about 60%, together with a second peak at full coverage. The median language coverage is 56.25% (IQR = 22.22, SD = 21.21) with a mean of 57.42%, which is somewhat lower than in PHOIBLE (median = 66.67%, mean = 62.74%), so even though grambank covers more languages, both the language and speaker coverage tends to be lower on a global country based scale. This is very interesting because it implies that the languages not included in Grambank are of higher importance globally, contributing to

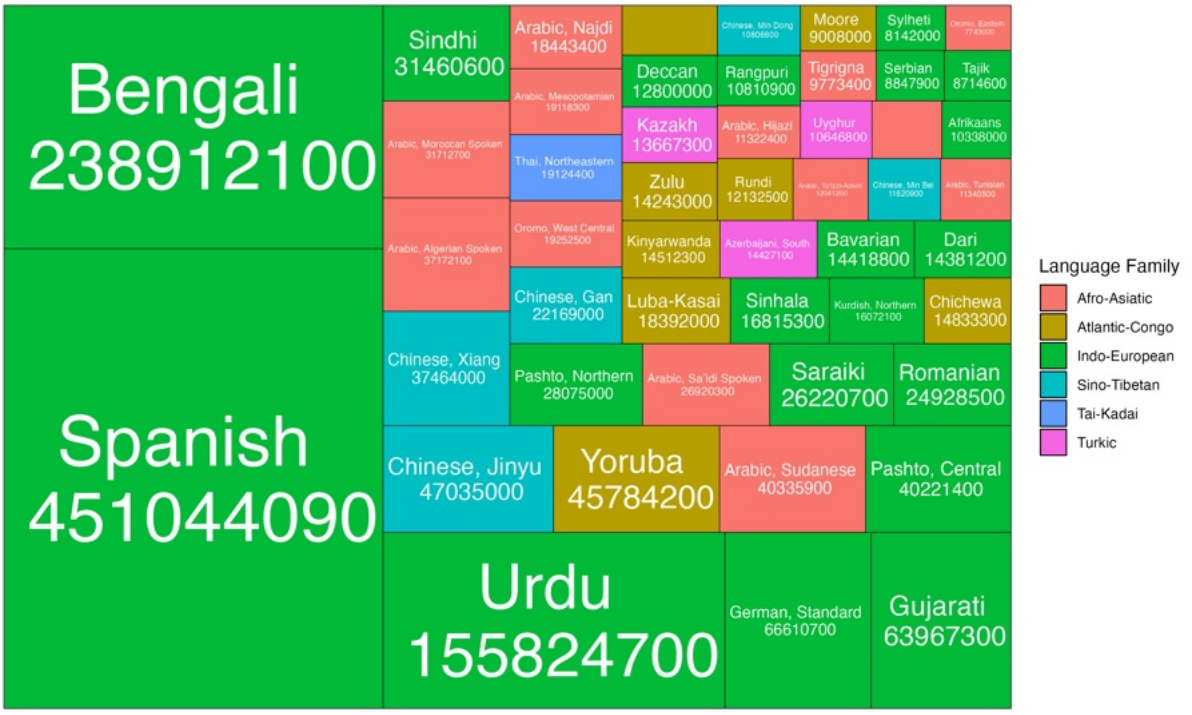

Figure 4.12: Treemap showing the 50 languages with the most speakers missing in Grambank colored according to language family. The area of the rectangles is proportional to the number of speakers.

linguistic diversity in multiple countries, as established for e.g. Spanish and Standard German. All in all, only 3 countries lack any language and speaker data (Jersey, Pitcairn, and São Tomé & Príncipe), while 24 countries have full data coverage.

The geographical distribution of the coverage is made visible through the maps in Figure 4.14. Considering language coverage, the Americas, apart from Mexico, Eastern Africa, and North Eurasia are well covered, while there are a few blind spots in the South of Asia and Central Africa, such as Pakistan and Nigeria, with very low coverage. Considering the speaker coverage, the impact of Spanish not being included in Grambank becomes evident, as most of Central and South America has a very low coverage together with Spain in Europe, meaning that for this part of the world, syntactical similarity derived from Grambank would be an unsuitable measure. The Maghreb and Iraq also have low speaker coverage, which must be the result of spoken Arabic varieties missing, as concluded previously, and also Central Asia has low speaker coverage, correlating with the missing Turkish languages. In Europe, interestingly, most countries seem to either have a very high or very low speaker coverage; many countries, such as Poland, Ukraine, and Finland, have near full speaker coverage, while countries such as Norway, Germany, and Romania have almost zero speaker coverage.

Furthermore, taking completeness into account as seen in Figure 4.14c, it seems that although the number of languages included in Grambank is relatively evenly spread over the world, the completeness of the data appears to be much higher the further away from the equator one goes. In Europe, North America, Central Asia, and East Asia, the languages in Grambank

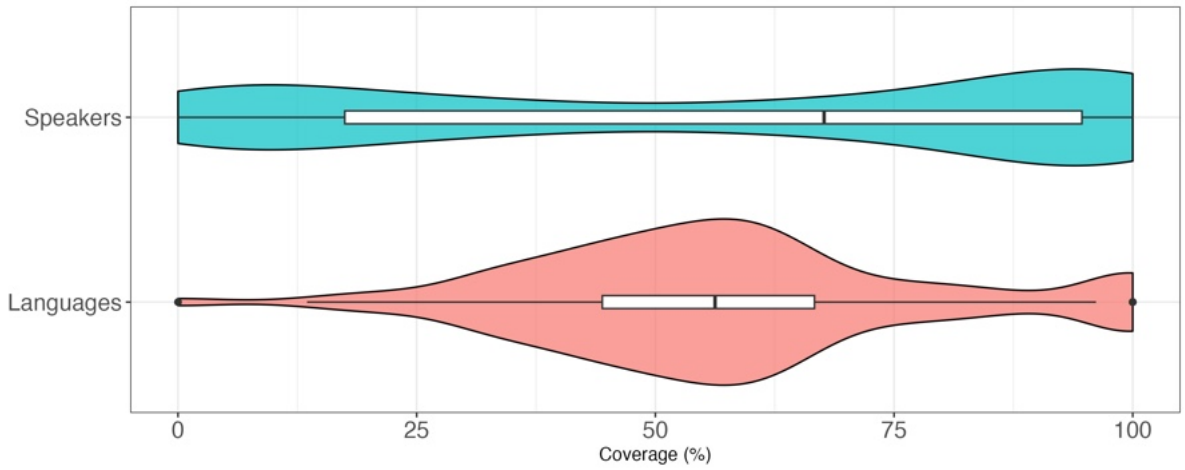

Figure 4.13: A Violin plot displaying the distribution of language of speaker coverage in Grambank on a country level using kernel density estimation and a boxplot.

are on average commonly defined for more features, as compared to the countries around the equator and it almost appears as if there is a gradient where the further away from the equator a country is, the more features the languages on average are commonly defined for. A test of correlation using Pearson's r ("Pearson's Correlation Coefficient", 2008) with the average number of features commonly defined in Grambank and absolute latitude of the capital in each country confirms the suggested relationship, revealing a significant and moderate positive effect.($r = 0.54$, 95% CI [0.44, 0.63], $p < 0.001$).

Summary

To conclude the descriptive analysis of Grambank, it can be stated that the database is extensive in terms of the number of languages and language families covered. Its inclusion of languages from different macro regions and of different speaker group sizes is representative of the suggested underlying distribution, although some underrepresentation of Africa is found. The database has a relatively low coverage of speakers compared to the vast number of languages amd most importantly, it is missing a few extensively spoken languages such as Spanish and Standard German which crucially limits the possibility of combining the structural similarity derived from Grambank with relative abundance weighted diversity measures for large parts of the world, most notably South America. It is also suggested that such measures will be more reliable for countries farther away from the equator due to more extensive data completeness for the languages found there, and that country-level coverage of languages and speakers is noticeably low. The combination of the above-listed characteristics suggests that Grambank functions well as a representative sample of the diversity of the languages of the world, but is very limited regarding the operationalization of similarity-weighted linguistic diversity measures on a global scale.

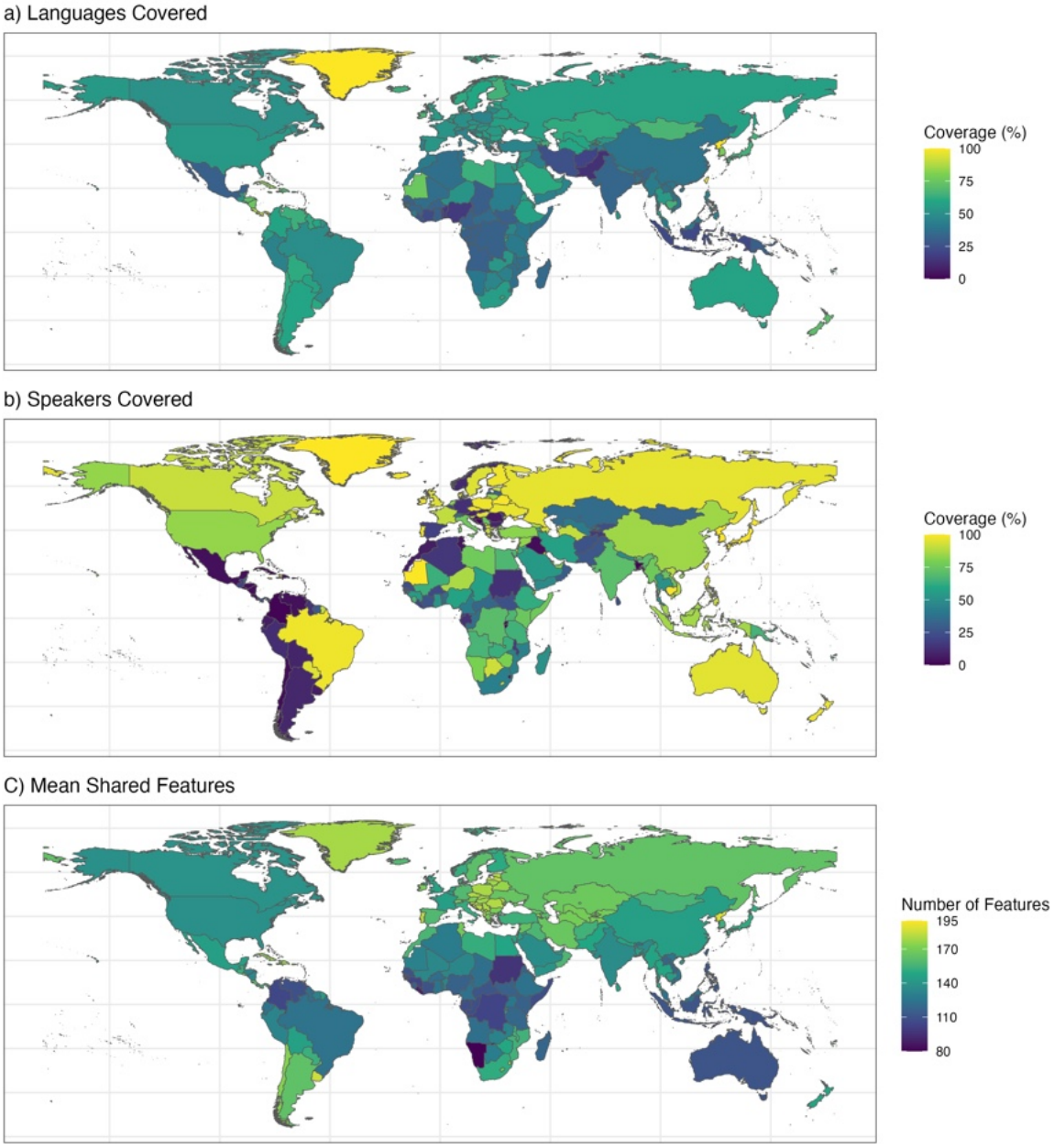

Figure 4.14: World maps showing on a country level a) the percentage of languages covered in Grambank, b) the percentage of speakers covered in Grambank, and c) the mean number of features in Grambank for which pairs of languages are defined

4.2.3 The ASJP Database

The final database to consider in this analysis is ASJP, which is notably extensive in its coverage (List et al., 2022). Considering Table 4.6, ASJP covers a little more than 5000 languages, giving a coverage of 75%. At the same time, it covers all 214 language families present in the baseline and 96.44 % of speakers. As such, ASJP covers most of the data that constitutes the foundation for assessing linguistic diversity on a global scale. Using a threshold of at least 70% data completeness for the purpose of calculating similarity, the extensiveness is reduced somewhat to 67.60% of languages covered, but the speaker coverage remains high with 94.77% coverage.

Macroarea	Languages	Language Families	Speakers	Completeness (%)
Africa	1722	37	1 271 102 655	85.52
Australia	109	17	51 691	84.08
Eurasia	918	23	5 764 778 088	87.51
North America	359	33	25 609 754	90.88
Papunesia	1633	76	281 156 456	86.75
South America	322	39	19 692 687	91.23
Global	5063	214	7 362 391 331	87.00
Coverage (%)	74.93	100	96.44	-
$\geq$ 70% Completeness	4567	214	7 235 194 838	90.9
Coverage (%)	67.60	100	94.77	-

Table 4.6: Statistics showing the number of languages, language families, and data completeness in the ASJP database relative to the baseline.

Macroarea

Concerning macroareas, ASJP follows the underlying distribution very well, which could be expected given the extent of the coverage, but there seems to be some bias against languages from Eurasia, which can be seen in Figure 4.15. While ASJP overrepresents languages from Africa and Papunesia with an absolute percent difference of approximately 2 to 3%, Eurasia is underrepresented by 5.4 percentage points. This is somewhat analogous to Grambank, where, instead, Africa was strongly underrepresented. Overall, however, ASJP is suggested to represent the global linguistic setting from a geographical point of view well.

Speakers

With regard to speaker numbers as seen in Figure 4.16, ASJP is very balanced, only showing an overrepresentation of languages with high speaker numbers, giving a percent absolute difference of 3.07% compared to the baseline. Lower-mid and Upper-mid speaker number sizes see a small underrepresentation on the other hand. At the same time, the same proportion represents the languages with the smallest speaker communities as in the baseline. As such,

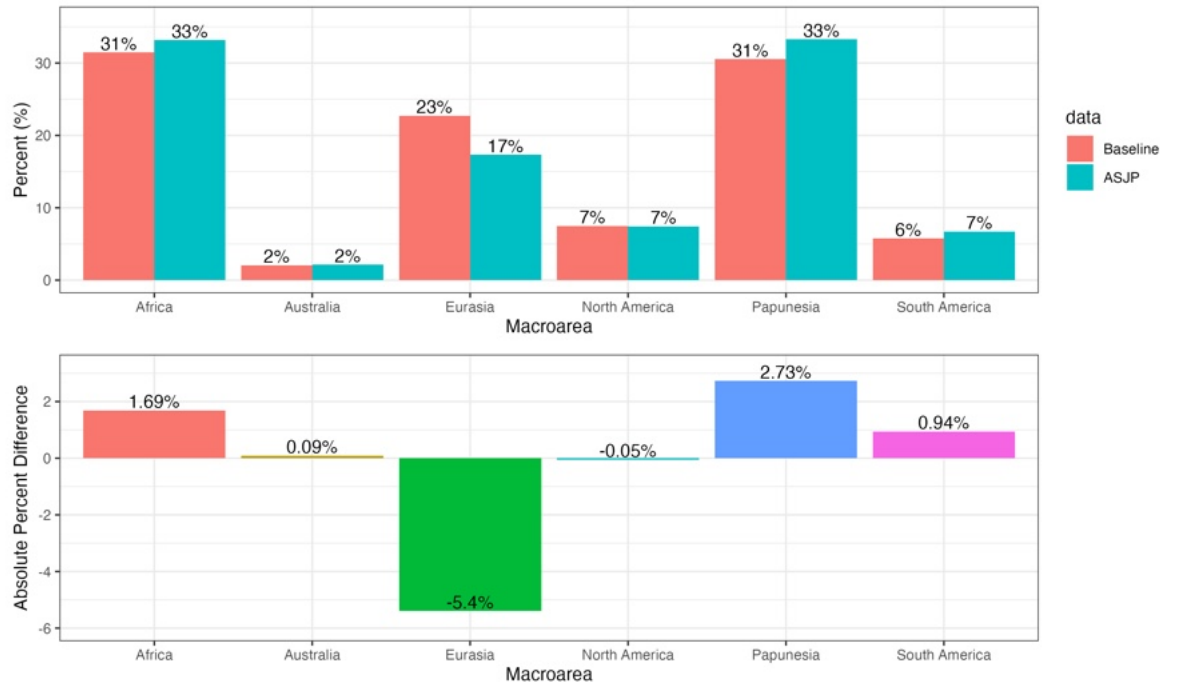

Figure 4.15: Barplot showing the difference in the distribution of languages in the baseline and ASJP data across macroareas. A positive absolute percent difference indicates overrepresentation in ASJP.

ASJP provides an excellent representation of languages of all speaker-community sizes, similar to that of Grambank.

Looking into the speaker numbers of the languages not covered by the ASJP data in Figure 4.17, the underrepresentation of Eurasian languages is reflected, as seen by the many Indo-European languages that are not covered. Sindhi and Deccan are two of many languages found in India that are not present, but one can also find languages from Europe, such as Serbian and Walloon. One can also notice the absence of two sign languages, Indian and Chinese sign languages, but as previously discussed, comparing the similarity between sign language and spoken language is outside the scope of what the thesis intends to achieve so it simply can be acknowledged that sign languages make up such a large number of speakers that their absense in the data can be noticed, while acknowledging the implicated conceptual difficulties in considering the similarity of languages. From the Afroasiatic family, again, some varieties of Arabic are left out, which thus is confirmed as a trend across all typological databases included, but the extent of this neglect in ASJP is notably smaller than in both PHOIBLE and grambank.

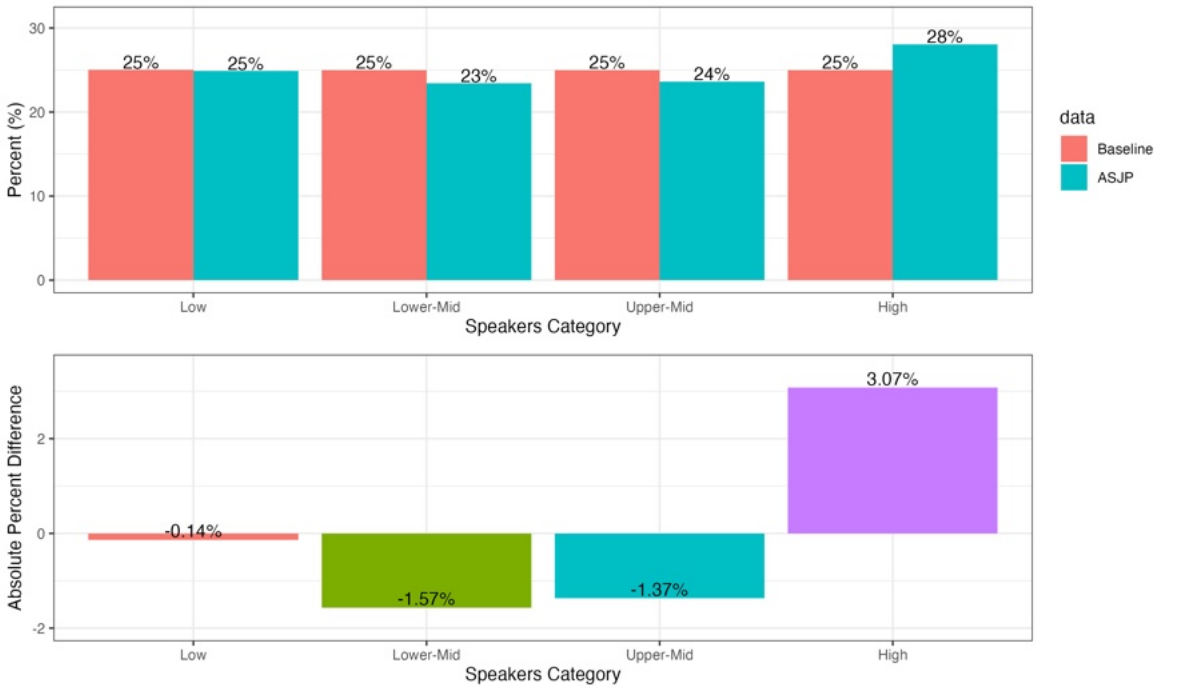

Figure 4.16: Barplot showing the difference in the distribution of languages in the baseline and ASJP data across categories based on the number of speakers. A positive absolute percent difference indicates overrepresentation in ASJP.

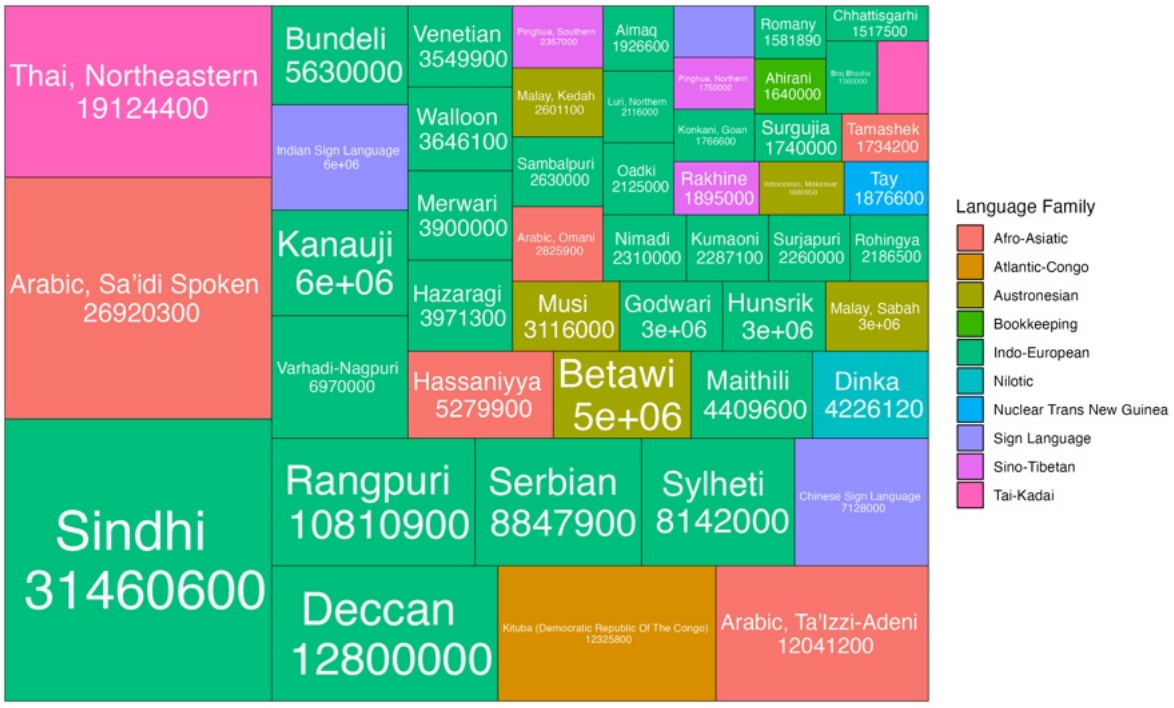

Figure 4.17: Treemap showing the 50 largest languages by speaker number that aren't present in the ASJP database colored according to language family.

Country

On a country level, the extensiveness of ASJP becomes especially apparent as seen 4.18. The violin plot reveals that the speaker coverage is central around full coverage with a clear peak at almost 100% coverage, which is summarized with a median speaker coverage of 98.81% on a country level with a small interquartile range of only 6.70. As can be seen, there are a small number of countries where the coverage in ASJP is notably low, which is why the mean coverage is reduced to 89.92% (SD = 22.47), but this can be contrasted with both PHOIBLE and Grambank, which both had fairly prominent peaks around almost no coverage. Furthermore, only two countries, Jersey and Pitcarin, have no speaker coverage at all and subsequently no languages covered either. However, apart from these two countries, the country with the third lowest coverage still has 45.94 % of languages covered, which signals that the coverage of ASJP is tremendous on a country level compared to both PHOIBLE and Grambank. The median language coverage is somewhat lower than the speaker coverage at 87.5% coverage (mean = 83.75, IQR = 14.33, SD = 15.35), but is still to be considered very high.

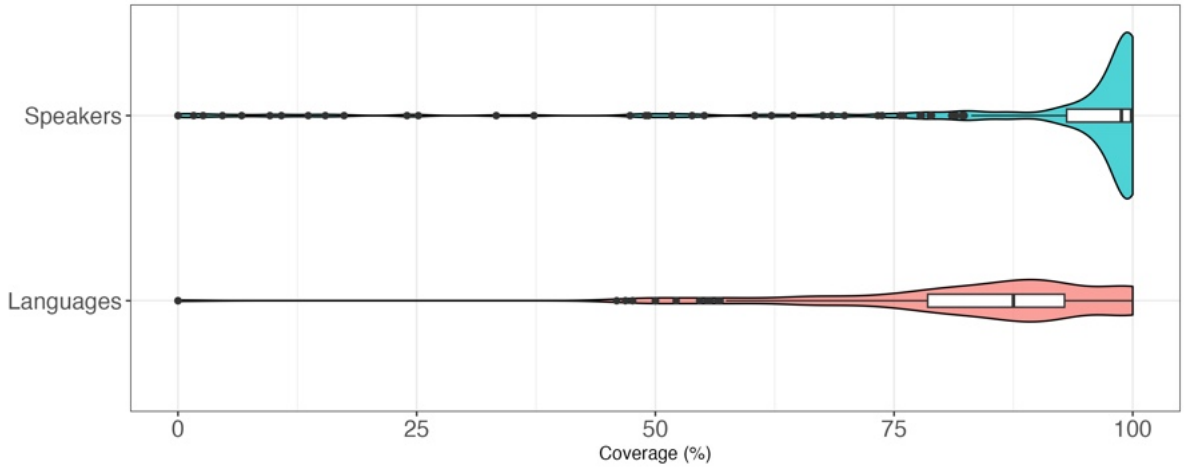

Figure 4.18: Violin plot showing the distribution of speaker and language coverage in the ASJP data.

Further consideration of the geographical distribution of this coverage can be taken by looking at the maps in Figure 4.19. Across the board, the language coverage is very high in both Americas, where only the language coverage in Mexico, as in PHOIBLE and Grambank, is somewhat lower. Much of Europe, Central Asia, and the Middle East have a high language coverage, which also goes for much of Africa and Oceania. The blind spot in ASJP observed on a language level is exerted in Southeast Asia, with e.g., Pakistan (55.41%) , India (46.86%), and China (56.78%) having a noticeably lower language cover than other regions of the world. This lack of language coverage is somewhat mended when considering speaker coverage, which is noticeably higher, especially for China (98.16%). On a speaker level, lack of coverage is found spotted over the world, not forming any clear groupings, with moderate coverage for e.g., Belgium (48.98%), Surinam (47.43%), and Oman (55.12%). Especially low coverage is noticeably found for Mauretania (17.40%) and Guyana (15.43%). As such, it can be concluded that, apart from

a few exceptions, the coverage of the ASJP database is almost all-encompassing on a country level when considering the relative abundance of speakers, making it appropriate for deriving similarity measures to improve models of global linguistic diversity.

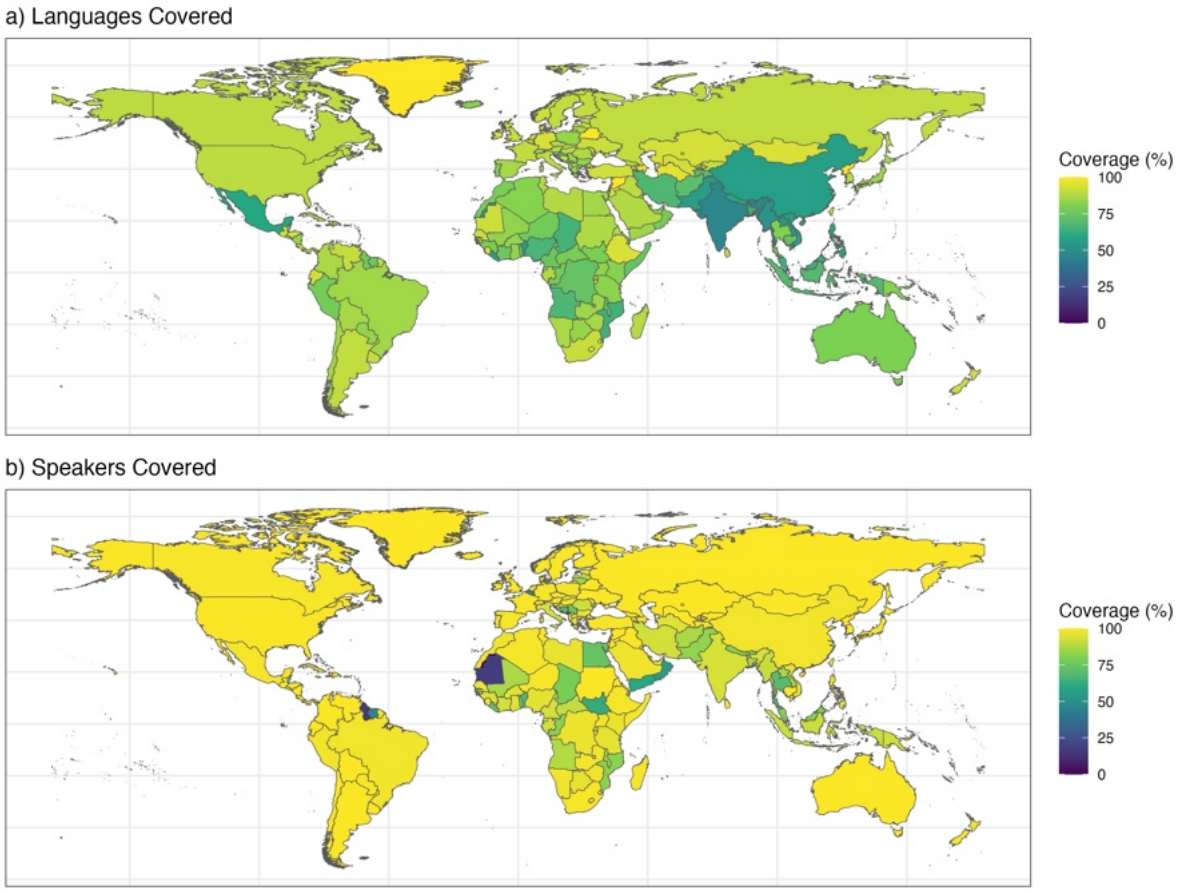

Figure 4.19: Maps showing the geographical distribution of a) languages and b) speakers covered on a country level in the ASJP database.

4.2.4 Summary

In this section, the global linguistic coverage of the three cross-linguistic databases, PHOIBLE, Grambank and the ASJP database have been compared to assess the extent they are representative of the global linguistic setting and could be used to derive linguistic similarity measures for the purpose of assessing linguistic diversity on a country level. The most important summary statistics derived from the analysis to convey this have been summarized in Table 4.6.

	Phoible	Grambank	ASJP
Language Coverage	1820 (26.93%)	2108 (31.12%)	4567 (67.60%)
Family Coverage	160 (74.76%)	195 (91.12%)	214 (100%)
Speaker Coverage	$6.446 \cdot 10^9$ (84.36%)	$5.065 \cdot 10^9$ (64.65%)	$7.235 \cdot 10^9$ (92.67%)
Median Country Language Coverage	66.67%	56.25%	87.50%
Median Country Speaker Coverage	89.10%	67.68%	98.81%

Table 4.7: Table summarizing the language, speaker, and family coverage of the typological databases PHOIBLE, Grambank, ASJP based on the data which can be used to reliably derive similarity measures. The ASJP data assumes 70% data-completeness and the grambank data 25%.

It has been shown that the most extensive as well as representative database is the ASJP database, which contains data for deriving similarity measures covering 92.67% of the speaker data. At the same time, it covers all language families and two-thirds of all languages found in the baseline, and as such, the ASJP database is, by considering coverage, unquestionably aligned the best with the endeavor of assessing global linguistic diversity. PHOIBLE, with its large speaker coverage, could potentially be used reliably as well, but lacks extensive language richness, and is heavily biased against most prominently the languages of Papunesia and multiple Arabic varieties as seen in Figure 4.20, resulting in low speaker coverage throughout North Africa and the Middle East. Grambank on the other hand covers more languages than PHOIBLE and is a more representative sample of the global linguistic setting, but has a remarkably low speaker coverage, and above all, excludes languages that are present in multiple countries such as Spanish and Standard German, rendering measures of similarity effectively meaningless for assessing diversity in much of South America. While the curators of Grambank have as an ambition to cover as many languages as there exist grammars, the database can not be used for operationalization of morphosyntactic similarity on a truly global scale in the current version.

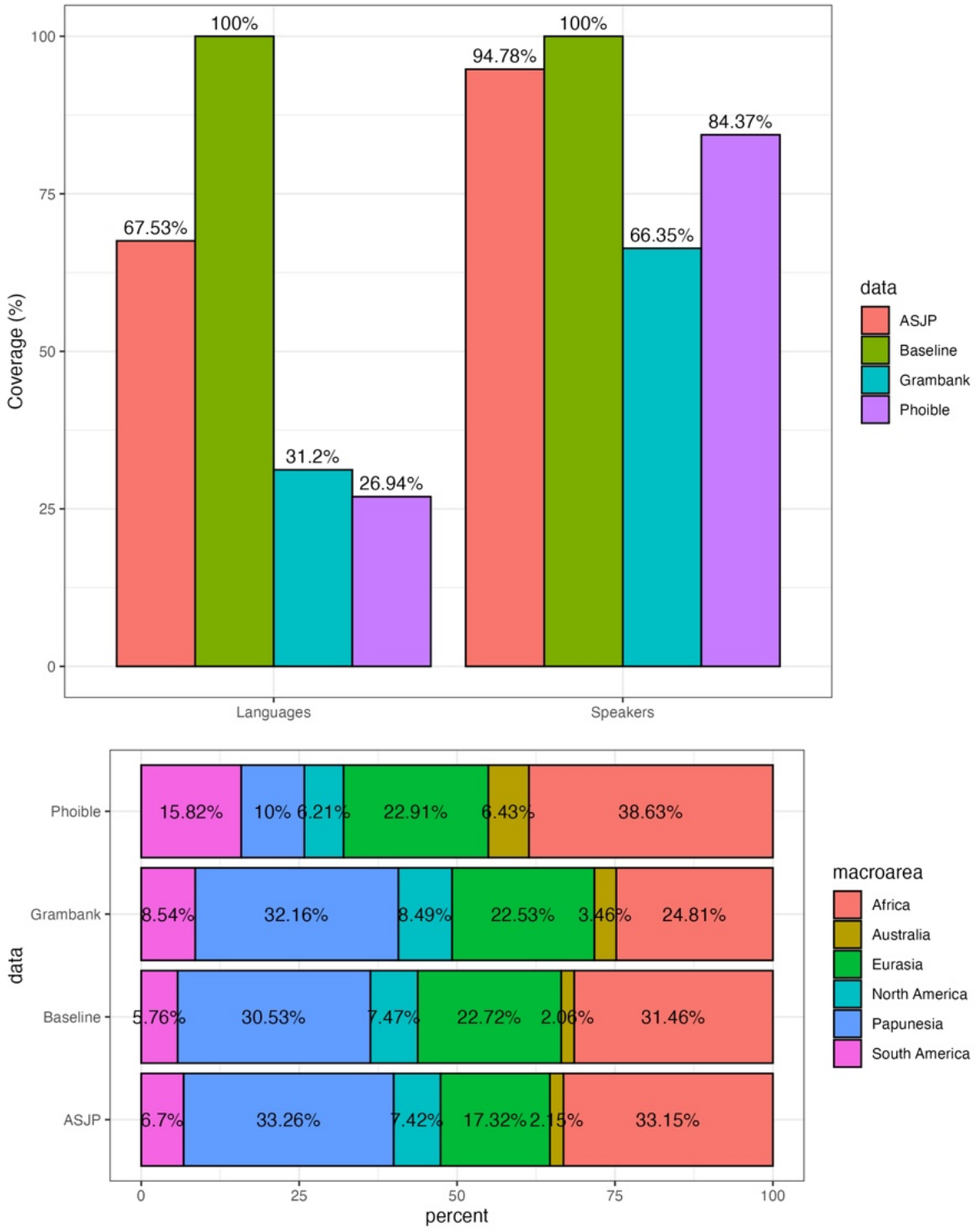

Figure 4.20: Barcharts displaying the coverage of the databases PHOIBLE, Grambank, ASJP compared to the baseline, concerning languages, speakers and macroareas

5 Distance and Diversity

Having considered the coverage of cross-linguistics databases on a global scale, this section devotes itself to first analyzing the similarity measures derived from the typological databases. This is followed by the implementation of a similarity-weighted diversity index and subsequent evaluation of its impact on how diversity is perceived.

5.1 Linguistic Similarity Measures

5.1.1 Cluster Analysis

In this part, the ability of the similarity measures ability to capture known linguistic relationships is verified using agglomerative hierarchichal clustering with average linkage. Assuming that the phylogenetic distance measure correctly captures phylogenetic relationships, nonrelated languages should appear as isolated simplicifolious, with language families forming subclusters. Looking at the dendrogram based on phylogenetic similarity in Figure 5.1, one can see that Maltese (Semitic), Georgian (Kartvelian), and Basque (isolated) are correctly clustered as isolated branches, unrelated to any other languages in the subset. Three clusters equivalent to language families are identified: the Turkish languages, the Finno-Ugric languages, and the Indo-European languages. Estonian and Finnish are correctly found on a common branch, as well as Skolt and Northern Sami, with Hungarian joining close to the first split; this does imply that Hungarian would be closer related to the Sami languages than the Finnic, which isn't the case according to Glottolog (Hammarström, 2015), identifying them as three distinct subgroups. Instead, this small relative similarity between Saami and Hungaric languages is a measurement error resulting from the longer branches of the Finnic subfamily.

Within the Indo-European cluster, the Romance, Balto-Slavic, and Germanic languages can be seen as correctly modelled, with Greek joining as isolated. Armenian and the Albanian languages are modelled as a subbranch, likely due to a similar branching structure, as was the case with Hungarian and the Saami languages. Furthermore, the Romance languages appear as a more distinct cluster compared to the other Indo-European subfamilies, which indeed can be a sign of research and knowledge bias. Since the Romance subfamily and its development from Latin through multiple stages, the lineage of the languages has potentially been more precisely described, resulting in larger distances to other subfamilies; this does however not mean that they are in reality more dissimilar or distant from e.g. the Balto-Slavic languages compared to the Germanic languages; they have simply been studied more. A possible remedy to the measurement error induced by this bias would be to use a weighted version of the Jaccard similarity, giving smaller weights to shared ancestry closer to the protolanguage, assigning successively larger weights to shared ancestry further away, thus reducing the impact

of overbranching on dissimilarity that would result from research bias.

On the whole, it appears that the measure indeed captures phylogenetic relationships well, allowing for the correct identification of language families and subfamilies, but the similarity values observed are not meaningful themselves, while also being quite susceptible to bias.

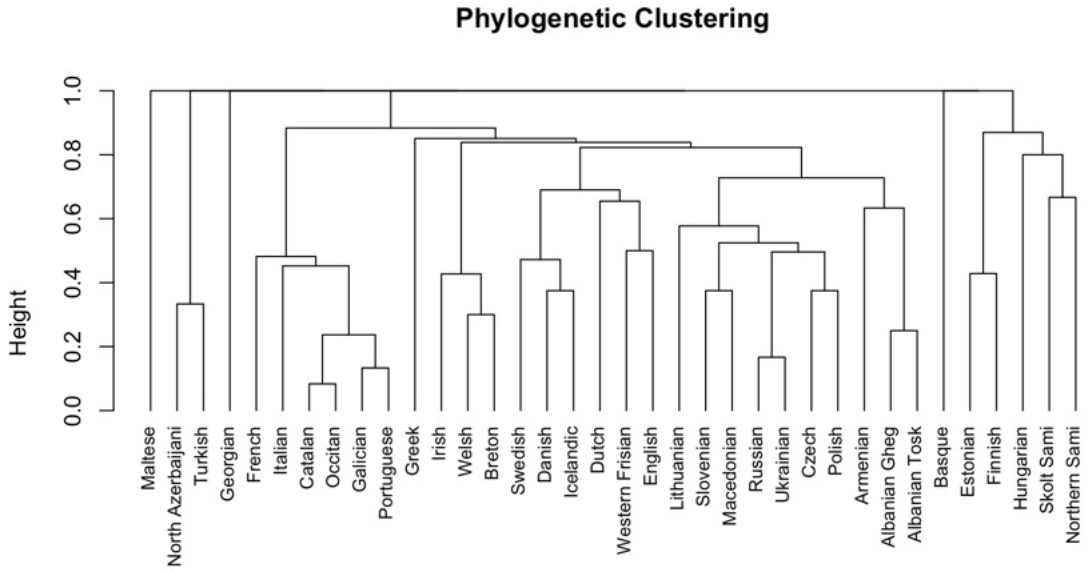

Figure 5.1: Dendrogram resulting from agglomerative hierarchical clustering with average linkage based on phylogenetic similarity.

Lexical Similarity Clustering

If one instead looks at the lexical clustering in Figure 5.2, it too clearly allows the identification of language families, capturing relatedness of languages as expressed through similarity in core vocabulary. The language families are all more or less splitting at the same height, with the Turkish family joining with Georgian very late, which could suggest small lexical similarity, possibly because of their geographic proximity. Maltese and Basque are identified as isolated, joining with the Indo-European and Finno-Ugric families very late, respectively. In the Finno-Ugric branch, the Finnic and Saami languages join before Hungarian, forming a clear cluster. The Indo-European family sees Armenian and Greek as isolates, with Armenian joining together with the Albanian languages, and Greek with the Romance languages

The Indo-European family can be identified in the middle, with all branches fairly equally distant, considering the height of the splits. The Balto-Slavic is found with the Slavic languages forming a tight cluster, with Lithuania joining later. Interestingly, Irish joins the other Celtic languages very late, suggesting a high degree of lexical dissimilarity even though they belong to

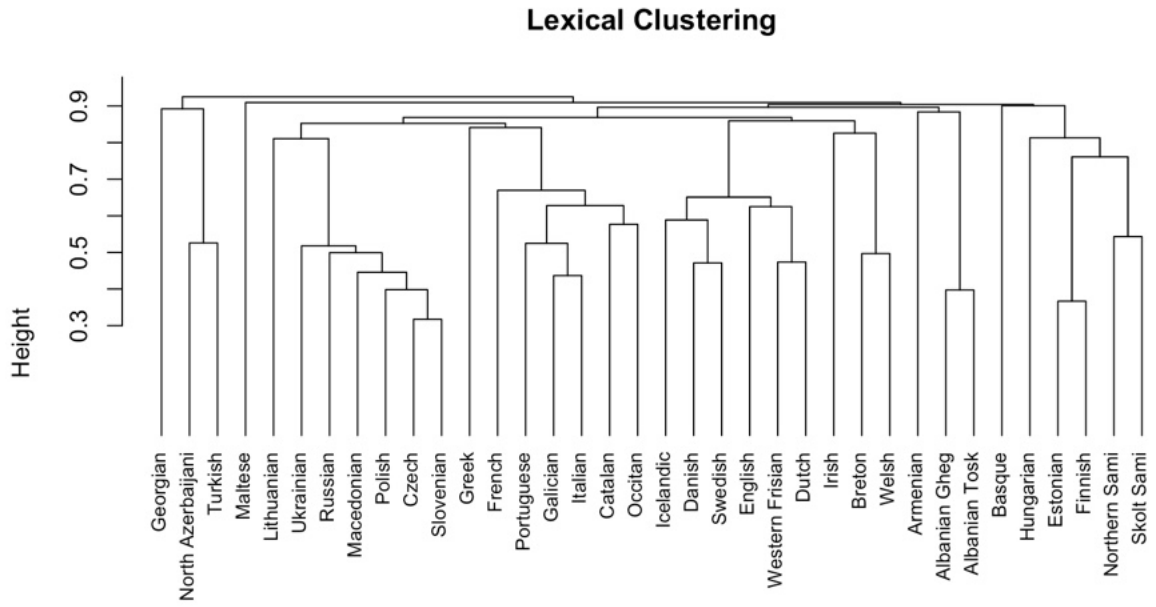

Figure 5.2: Dendrogram resulting from agglomerative hierarchical clustering with average linkage based on lexical similarity.

the same family, albeit different branches. Within the larger subfamilies, the clustering also seems to align well with known relationships between languages, apart from a few deviations. E.g., Czech and Slovenian are clustered closer together than Polish and Macedonian, while both Slovenian and Macedonian are South Slavic languages, and Polish and Czech are West Slavic. Similarly, one would expect Ukrainian and Russian to be clustered as a bifolious clade. Especially interesting to note is that the lexical clustering places Danish and Swedish closer together than Icelandic, which wasn't the case for the Phylogenetic measure. Reasonably, Danish and Swedish should be modelled as more similar languages, since they are generally considered mutually intelligible and have developed faster in parallel compared to the conservative Icelandic (Trudgill, 2020). The fact that this is reflected in the clustering speaks for lexical similarity being able to capture similarities between languages in a meaningful way, and still detect and maintain high-level phylogenetic relationships to the same degree as the phylogenetic similarity measure.

Morphosyntactical Similarity Clustering

Consideration of Figure 5.3 suggests that phylogenetic relationships aren't as clearly captured by morphological similarity as by lexical similarity, but some clear division is found. Whereas the dendrogram based on lexical similarity early splits into multiple subclusters corresponding to different language families, the morphosyntactic dendrogram primarily consists of two large subclusters. To the left, a cluster of only Indo-European languages, which successively split into subclusters largely corresponding to the major Indo-European subfamilies. On the other side, a

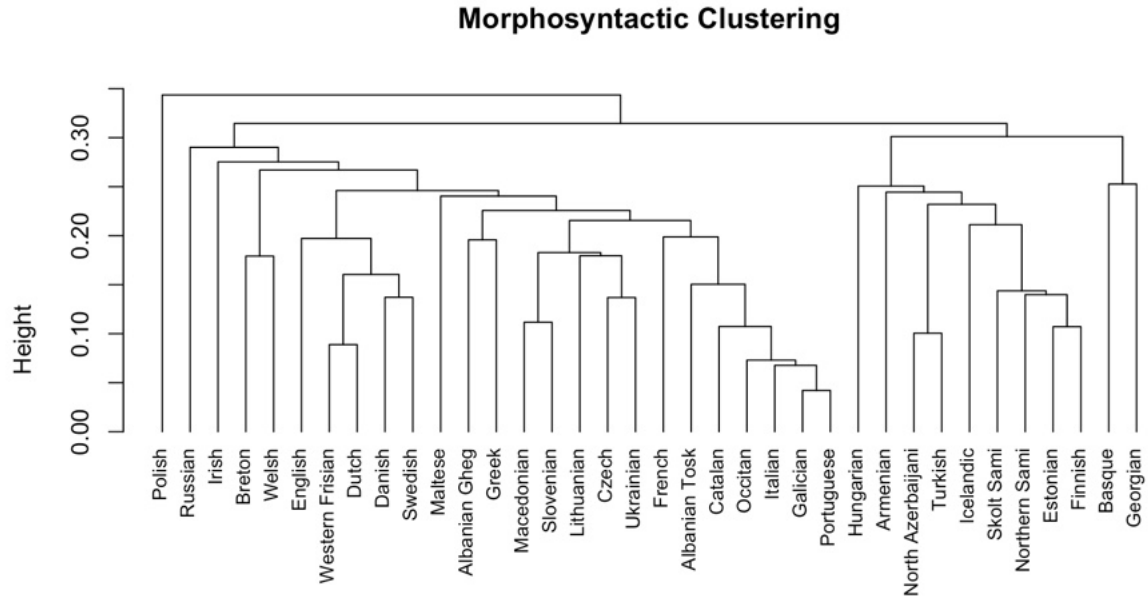

Figure 5.3: Dendrogram resulting from agglomerative hierarchical clustering with average linkage based on morphosyntactic similarity.

large cluster with non-Indo-European languages is found, among which Georgian and Basque form a bifolious clade, an observation that is very interesting to consider. Basque and Georgian are both ergative languages and share some superficial similarities, which has resulted in suggestions of them being related, the so-called Caucasian-Basque connection (Pereltsvaig, 2020; Trask, 1997). Although the idea of the languages being related has not found scholarly acceptance, the fact that they are known to show syntactic similarities and this being captured by measure is a strong argument for the measuring meaningfully capturing similarities between languages. Equally so, it is interesting to note that the other languages that are part of this large cluster are agglutinative, as both Basque and Georgian also are. For example, the Finno-Ugric family, apart from Hungary, which joins later, forms a tight-knit cluster that then joins with the Turkic cluster. Both these language groups are known for being highly agglutinative, while the Indo-European languages generally tend to be fusional and analytical. It is inviting to suggest that the morphosyntactic similarity measures capture this fundamental difference between, on the one hand, agglutinative languages and, on the other hand, fusional languages. This is further supported by Maltese being found among the Indo-European languages, being highly fusional due to influence from Italian (Hoberman, 2007; Kortmann & van der Auwera, 2011b), while Armenian, which phylogenetically is a Indo-European language and tended to fusional inflection at earlier stages, today rather has an agglutinative syntax attributed to influence from Turkish to the point that

> "...modern Armenian is simply Armenian phonology and morphology with Turkish syntax" (Vaux & Sayeed, 2017, 1158).

The only thing that can be observed to be off from the clustering based on morphosyntactic

is the high similarity between Icelandic and the Finnic and Saami languages. This relationship might be spurious at best, but it is interesting to note that all four Finnic and Saamic languages place themselves among the eight languages most morphosyntactically similar to Iceland, and that Icelandic generally scores so high in similarity to the other agglutinative languages, compared to the much more closely related Germanic languages. To explain this observed similarity further, one would have to look into the specific values from Grambank for Icelandic and compare, but for this thesis, we leave it at the fact that the morphosyntactic similarity measure indeed captures meaningful similarities in syntax and morphology.

Phonemic Similarity Clustering

Finally, the dendrograms derived from clustering based on phonemic inventory in Figure 5.4 can be considered. Here, the structure is not as distinctive as in the other dendrograms, but some substructures that are to be expected are found. For example, all Germanic languages form a cluster together with the Celtic language Welsh, except for Danish, which is a simplicifolious. This is interesting because it could reflect the fact that Danish has gone through numerous quick changes to its phonetic segments that in its spoken form is rendered partly unintelligible to Swedish speakers (Gooskens, van Heuven, van Bezooijen, & Pacilly, 2010). Portuguese is a similar case, being placed as its own branch and being known for its peculiar phonology, often referred to as sounding Eastern European (Deacon, 2016). Apart from this, the Turkic languages are again found together, and the Finno-Ugric languages Finnish, Northern Sami, and Hungarian are clustered fairly close, together with Maltese. Interestingly, Estonian is not included in the cluster, although the languages most similar to it according to the measure are indeed Finnish, Hungarian, Northern Sami, and also Maltese.

Other signs that the measure captures similarities in phoneme inventory meaningfully are the clustering of Armenian and Georgian, which is not improbable given their historical language contact Vaux and Sayeed (2017), and a cluster of south and west Slavic languages is observed together with the Albanian languages can be seen as well. Finally, it is very intriguing to notice that Breton and French are clustered together, since although they are only distantly related, French has exerted influence on Breton for centuries, with every current speaker also being fluent in French, thus influencing all linguistic dimensions, which could be why they are measured as similar (Kennard, 2021). Furthermore, it can generally be observed that splits in the dendrogram take place very high up, signaling that the subclusters identified are quite dissimilar. As such, while certainly there seems to be a much higher degree of arbitrariness to the measure of phonemic inventory, it does seem to correspond to known similarities and, in particular, manages to capture areal similarity as opposed to phylogenetic.

Alignment

Much of the qualitative analysis of the dendrogram just carried out can be supported by quantitative measures as seen in Table 5.1, where the entanglement and cophenetic correlation values signal how well the respective clusters are in agreement. Indeed, the dendrograms based on lexical similarity and phylogenetic similarity show the highest degree of similarity with a

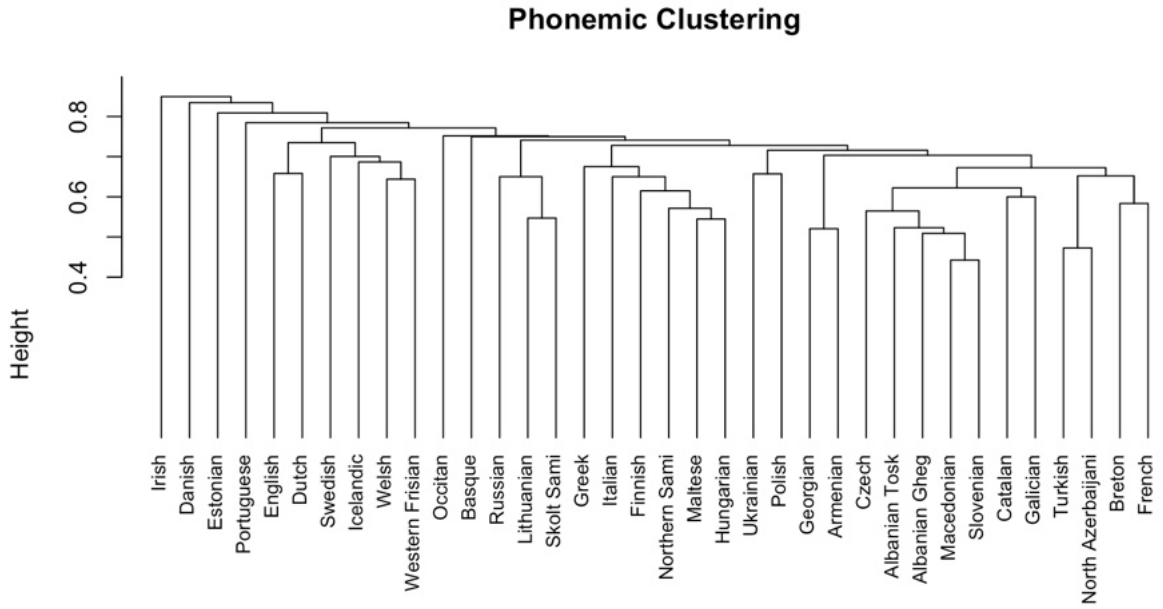

Figure 5.4: Dendrogram resulting from agglomerative hierarchical clustering with average linkage based on Phonemic similarity.

large correlation coefficient ($r = 0.83$, $p < 0.05$), indicating that there is a large association between the heights the respective language pairs join for the first time. The entanglement, an ilustration of which can be seen in Figure 5.5, also indicates extensive agreement between the clusterings, signaling that the general structure of the clusters allows the leaves to be aligned very well. Cases where misalignment can be observed concern Hungarian, which, as previously discussed, was placed closer to the Sami languages according to the phylogenetic measure. On the whole, the dendrograms are almost perfectly aligned with an entanglement of 0.04. Tanglegrams between the other dendrograms can be found in the Appendix.

	Phylogenetic	Lexical	Morphosyntactic	Phonemic
Phylogenetic	0	0.83[0.80, 0.85] | 0.04	0.55[0.49, 0.69] | 0.10	0.11 [0.04, 0.20] | 0.19
Lexical		0	0.51[0.44, 0.56] | 0.06	0.14[0.06, 0.21] | 0.20
Morphosyntactic			0	0.12[0.05, 0.20] | 0.17
Phonemic				0

Table 5.1: Table showing agreement of the clustering based on phylogenetic, lexical, morphosyntactic, and phonemic similarity. In each cell, the left-hand value indicates Pearson's correlation coefficient between the respective dendrograms' cophenetic distances with 95% confidence intervals, such that -1 indicates complete disagreement and 1 complete agreement. The right-hand value denotes the entanglement between the respective dendrograms, such that 0 equals the best possible alignment, and 1 the worst possible alignment.

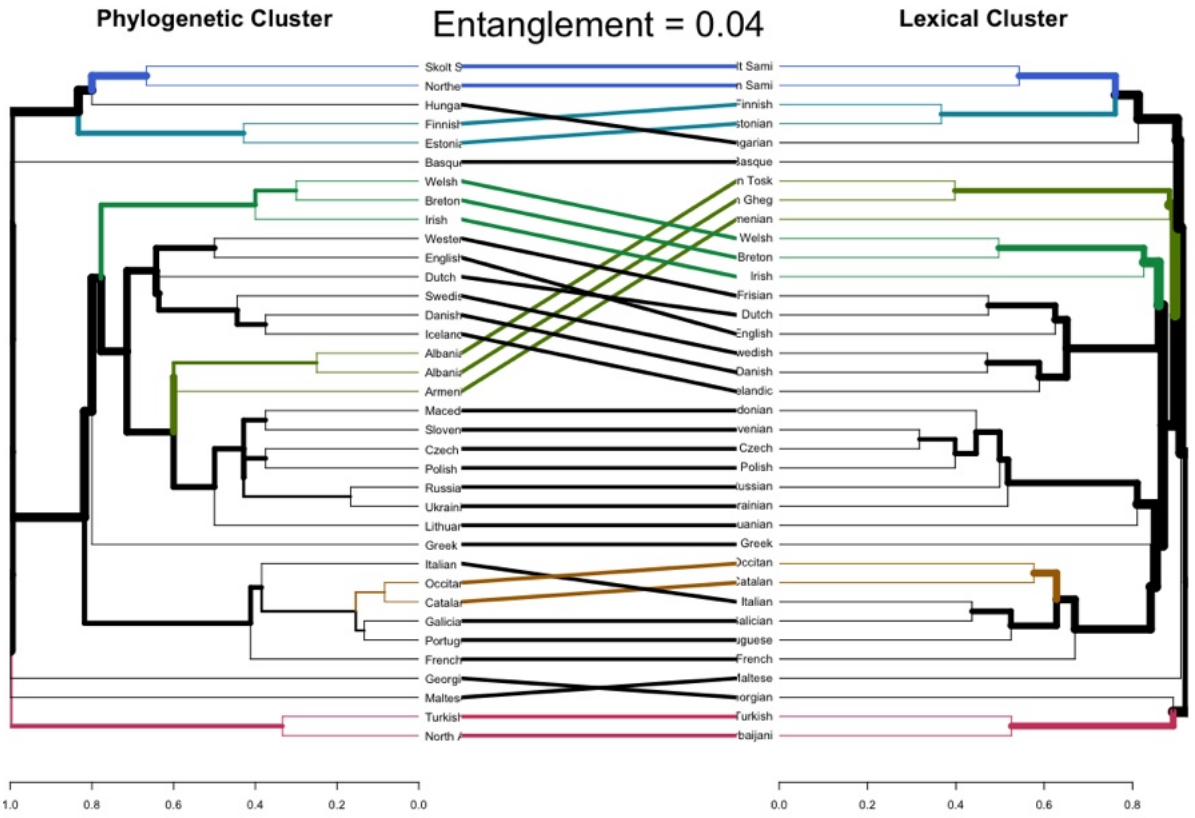

Figure 5.5: A tanglegram showcasing the structural alignment between the dendrograms resulting from hierarchical clustering based on phylogenetic and lexical similarities. Colored branches indicate identical substructures in the respective dendrograms. The dendrograms were untangled using the step2side algorithm.

The phylogenetic clustering furthermore seems to be decently aligned with Morphosyntactic clustering, with a moderate cophenetic correlation ($r = 0.54$, $p < 0.05$) but somewhat more

entangled with an entanglement of 0.10, signaling that the pairs of leaves are well aligned, but the overall clustering structure is more discordant; the measures are not entirely agreeing on higher level clusters. Lexical and Morphosyntactic similarity, on the other hand, are very well aligned with an entanglement of only 0.06, similar to that of Lexical and Morphosyntactic similarity, but with only a moderate cophenetic correlation ($r = 0.51$, $p < 0.05$). Clustering based on phonemic inventory similarity does not align itself particularly well with any other dendrograms, as the cophenetic correlations are all very weak, and the entanglement is noticeably higher, with the best agreement found with the Morphosyntactic similarity at 0.17. As such, these results support the observation that Lexical and Morphosyntactic similarity truly capture meaningful similarities, with Lexical similarity being more well aligned with phylogenetic classifications than Morphosyntactic similarity. Phonemic similarity, on the other hand, as seen by looking at the dendrograms, only shows some signs of capturing meaningful similarity, disaligning itself from classifications based on phylogenetic relationship. Still, there is some signal in the measure, as indicated by the significant, albeit weak, cophenetic correlations.

Summary

To summarize this section, the four linguistic similarity measures have been compared through hierarchical clustering on a subset of well-known European languages. It has emerged that the morphosyntactic and lexical similarity measures manage to meaningfully quantify similarities between the sample of languages according to their respective linguistic dimension, whereby lexical similarity more strongly aligns itself with Phylogenetic similarity, reflecting the genealogical relationship between languages. There are also some findings suggesting that phonemic inventory similarity captures meaningful similarities, but it is very misaligned with the established genealogical taxonomy. Phylogenetic similarity, on the other hand, captures the classification of languages into their respective subfamilies correctly, but the measured value in itself is concluded not to be meaningful due to values being fairly arbitrary and not reflecting actual similarities.

5.1.2 Correlation Analysis

Having considered the behavior of the measures for a handful of languages with well-established relationships and properties, the knowledge that the measures indeed capture similarities allows a further step to be taken by correlating the measures for as many languages as possible. For the four databases, excluding the baseline used in this thesis, there are 1012 languages in common for which similarity measures can be reliably calculated. Calculating all unique pairwise similarities between different languages and removing any missing values due to missing data in Grambank, one arrives at 495 017 instances with one similarity value for each instance, which can be correlated to deduce their relationship on a global level. It is important to note that since this data is the intersection of all databases, the bias existing in those databases will be reflected in this sample, e.g., language from Papunesia being underrepresented due to a lack of coverage in PHOIBLE, languages from Africa lacking coverage in Grambank, and languages from India missing in the ASJP database. The exact number of languages per macroarea can be seen in Table 5.1.2.

Macroarea	No. Languages	Percent (%)
Africa	290	28.7
Australia	99	9.78
Eurasia	223	22.0
North America	79	7.81
Papunesia	137	13.5
South America	184	18.2
Sum	1012	100

Table 5.2: Table showing the distribution of languages across macroareas in the intersection of languages present in PHOIBLE, Grambank, the ASJP database, and Glottolog.

It thus appears that Papunesia indeed is heavily underrepresented, with 13.5% of languages, while South America and Australia are overrepresented, with 18.2% and 9.78% respectively. This quite clearly resembles the distribution of languages found in PHOIBLE, as was seen in Figure 4.5. It therefore follows that this isn't a perfect representation of the entire global linguistic setting, but it is the largest sample possible given the available data, giving robustness and precise estimates.

First, consider the distribution of the pairwise similarity measures as displayed by the histograms in Figure 5.6. All four measures are distributed in their own characteristic way, of which the lexical similarity distribution has the sharpest peak with a median at 0.095 (mean = 0.10, IQR = 0.044, sd = 0.042), meaning that languages on average are very dissimilar, sharing about 10% lexicon as measured by the normalized Levenshtein distance in between form-meaning pairs in the core vocabulary.

The Morphosyntactic similarities follow a more or less normal distribution with a comparably high median similarity of 0.64 (mean = 0.64, IQR = 0.092, sd = 0.072), which is confirmed by the normal Q-Q plot seen in Figure 5.7. This suggests that Morphosyntactical similarity, as measured by the overlap in features based on the Grambank questionnaire, stems from a normally distributed population, which is a quite important observation as it would potentially allow for parametric statistical analysis related to syntactical similarity measured this way. Considering the size of this sample and its global coverage, such a regularized distribution of similarities could also imply a structured evolution of language structures with little randomness, constraining the extent to which languages can both diverge and converge, which is very much coherent with the findings and conclusions of (Skirgård, Haynie, Blasi, et al., 2023) based on the dimensional reduction of the dataset. To exemplify: if there were no limits to the development of morphosyntactic structures, one would expect to find completely dissimilar languages, ergo similarity measures of or close to zero. In the data, the two most dissimilar languages found are Ega of the Atlantic-Congo family in Africa and the Gunwinyguan language Ngandi in Australia. The similarity between these languages measures 0.27, meaning that although they are the two most dissimilar languages found in the data, they still share almost 30% of grammatical structure. Furthermore, the fact that the median is comparatively high at 0.64 and approximately 97% of similarity values are larger than 50% suggests that the languages of the world generally are more similar than dissimilar.

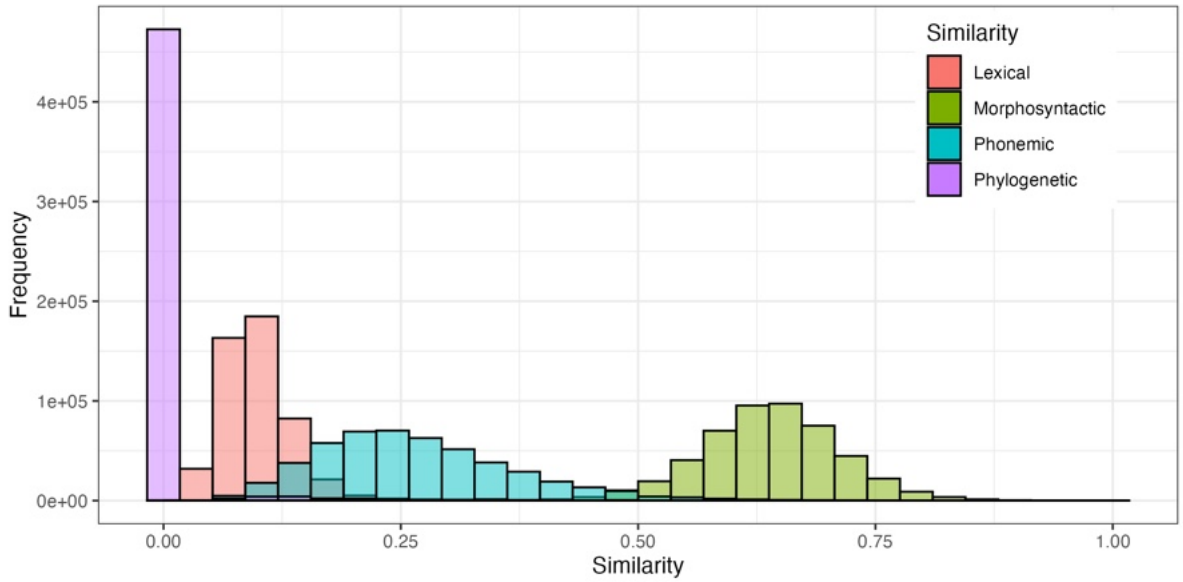

Figure 5.6: Histograms displaying the distribution of pairwise similarities between all languages found in Glottolog, Grambank, PHOIBLE and the ASJP database according to the linguistic domains of lexicon, morphosyntax, phoneme inventory, and phylogeny.

The distribution of phonemic inventory similarities is somewhat more flattened with fat tails and a right skew, but still unimodal around the median similarity of 0.254 (mean = 0.268, IQR = 0.135, sd = 0.106). As such, on average 25% of phoneme inventories are shared by the languages in the data. The most curious distribution is that of phylogenetic similarity, which consists of a sizable bar at 0, and a minuscule peak at around 0.10. As previously concluded, similarities between nonrelated languages can't be modelled by the measure, and since most languages of the world belong to different families, their similarity will always be zero, which, for this data, results in 472,751 language pairs with a similarity of zero, which equals 95.5% of all data points. Again, this highlights the problems of using taxonomy to operationalize linguistic similarity on a global scale, since the distributional property of the measure breaks the assumption of many commonly used statistical tests and techniques used to analyze such data. Consider, e.g, the Pearson's product-moment correlation (Pearson, 1896): it assumes a normal distribution of the variables and thus, for the data at hand, strictly speaking, could only be applied to the morphosyntactic similarity measure. The obvious nonparametric alternative is Spearman's rank correlation (Spearman, 1904), which assesses monotonic association by ordering the datapoints by rank, thus not assuming any certain distributional properties. However, if datapoints are tied, those tied datapoints are assigned the average of their ranks. For example, if the top 5 ranks are tied, they are all assigned the rank *3*. This introduces a reduction in variability, which reduces statistical power. If it occurs for 95% of datapoints, as in the case with phylogenetic similarity, it presents a serious challenge.

Considering the relationship between the variables as seen through the scatterplots with fitted linear trendlines in Figure 5.8, it is interesting to note that no particularly strong mono-

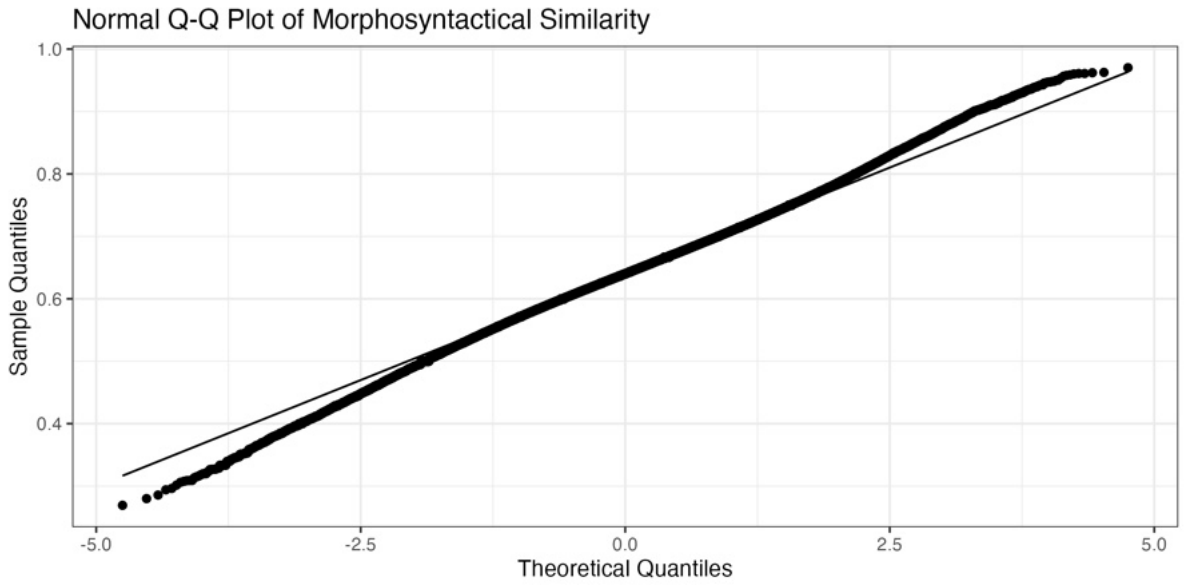

Figure 5.7: A normal Q-Q plot showing the sample quantiles of pairwise measurements of morphosyntactic similarity compared to the theoretical quantiles of an equivalent normal distribution. The close alignment between the quantiles suggest that the measure approximately follows a normal distribution.

tonically increasing relationships can be found as measured by Spearman's ρ, but across the board, all coefficients are statistically significant, even for very small values of ρ. The notably largest effect size is found for all correlations with Phylogenetic Similarity at around 0.2, indicating a weak association, which lends some support for the intuitive idea that it can be used as a proxy for overall language similarity. However, as concluded in the previous paragraph, the overwhelming number of zeros makes the measure unreliable. It is still interesting to consider the relationship and possible value two measures can take on. Viewing lexical and phylogenetic similarity in Figure 5.8a, there are no instances where the lexical similarity is large while the phylogenetic similarity is low; this means that there are no cases where two languages are grouped into different language families while having a particularly large lexical similarity. The case of two non-genealogically related languages having the largest lexical similarity found in the data is between two languages found in Australia, namely the Garrwan language Garrwa, and the extinct Limilngan of the Limilngan-Wulna family. They measure a lexical similarity of 0.359, which is comparable to the lexical similarity between the closely related Germanic languages of English and Danish, which have a lexical similarity of 0.344. Looking into the core vocabulary of these two languages as found in the ASJP database, there certainly are similarities to be found. Compare the concepts I, YOU, WE, PERSON, and FISH in Garrwa and Limilngan in ASJP code as seen in Table 5.3.

Concept	Garrwa	Limilngan
I	Nayu	Naykgi
YOU	ni5ji	Ni5i
WE	nuri	Nuyi
PERSON	muLkgiTi	muLgiTi
FISH	iwan	iwan

Table 5.3: A table showing the Garrwan and Limilngan words for the concepts I, YOU, WE, PERSON, and FISH in the ASJP database.

The similarity between the words for these concepts in the two languages is astonishing. Still, expert classification has concluded that these similarities in the core vocabulary are not evidence of shared genealogy, but rather the result of an intensive borrowing that has taken place in the Australian languages (Koch, 2014). Further indication of this particularity of horizontal transfer captured by measuring lexical similarity, while the phylogenetic measure is insensitive to it, is given by the observation that of the 10 most lexically similar language pairs with a lexical similarity > 0.30 while being unrelated, 9 are found in Australia, and the tenth most lexically similar unrelated languages are two languages found in South America (Huallaga Huánuco Quechua and Jaqaro). Conversely, one can observe that there are many datapoints where the languages have the same classification, giving a phylogenetic similarity of 1, while being lexically dissimilar. The most extreme example found in the data are the South American languages Karajá and Ofayé, belonging to the Macro-Je family and having the same genealogic classification and thus a phylogenetic similarity of 1, but a lexical similarity of 0.0845. Since it has been shown that the languages are related, we lack knowledge of how those languages evolved to create a family tree that meaningfully captures their relatedness in a way that is coherent with reality, which again serves as a prime example of how prone this phylogenetic measures are to distortion through bias in the available knowledge and scholarly attention. Still, it is evident from simply viewing the scatterplot in Figure 5.8 that there is a positive monotonic association between the lexical and phylogenetic relationship on a global scale, but it is far from clear-cut.

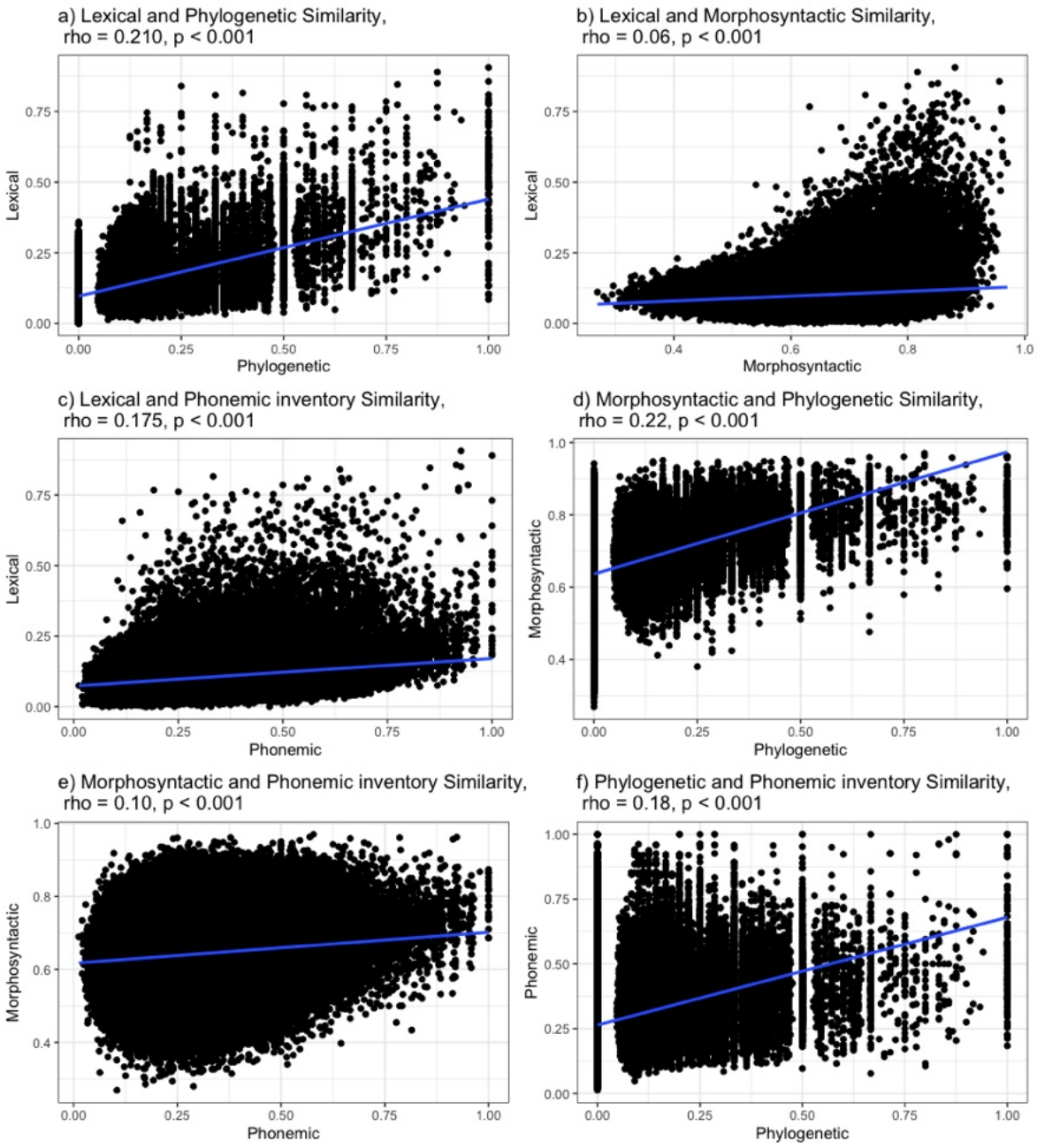

Figure 5.8: Scatterplots showing all combinations of lexical, phylogenetic morphosyntactic and phonemic similarity based on the intersection of languages in all cross-linguistic databases. Each scatterplot is supplemented by linear smoothing to indicate linear relationships between the variables.

Similar arguments can be made for the relationship between phylogenetic similarity and morphosyntactic/phonemic similarity; there is certainly a relationship, but it is not strong enough to claim that phylogenetic similarity functions as a proxy. Interestingly, in the case of morphosyntactic similarity, the scatterplot in Figure 5.8d) shows a relationship to phylogenetic

similarity which is the reverse to the relationship between lexical and phylogenetic: while low phylogenetic similarity is suggested to imply low lexical similarity, this limitation does not exist for morphosyntactic similarity, with unrelated languages existing on the entire scale of possible values, supporting the idea that syntactic features aren't restricted to any group of languages, but rather consists of universal configurations that can exist in any language, ignoring their relatedness. As relatedness increases, however, the possible variation in morphosyntactic distinction seems to decrease to the point that all languages with at least 0.75 phylogenetic similarity have a morphosyntactic similarity of 0.596, i.e., they share at least 60 % of the Grambank features for which both are defined. This creates a lower bound for Morphosyntactic similarity of two languages that can be expressed as a function of their phylogenetic similarity, such that a high phylogenetic similarity implies high morphosyntactic similarity, but a high morphosyntactic similarity does not imply a high phylogenetic similarity. A similar relationship can be observed between morphosyntactic similarity and phonemic inventory, where phonemic inventory similarity sets a lower bound for the morphosyntactic similarity, such that two languages that have a very similar phoneme inventory are implied to be morphosyntactically similar, but not the inverse. The same relationship is observed for lexical similarity and phonemic inventory similarity, but much weaker; lower bounding on lexical similarity by phonemic inventory similarity begins to apply for phonemic inventory similarity > 0.5, but bounding is very weak.

Finally, another obvious bounding relationship is found between lexical and morphosyntactic similarity. While their positive monotonic association is minimal as measured by Spearman's $\rho = 0.06$, p < 0.001, the existence of a relationship is clear from the exponentially increasing upper bound, which is set on lexical similarity by morphosyntactic similarity. If morphosyntactic similarity is low, lexical similarity is implied to be low, but as morphosyntactic similarity increases, the bound on lexical similarity is increased, allowing it to vary. As signaled by the low correlation coefficient, the lexical similarity relatively rarely increases together with morphosyntactic similarity, but the bounding suggests a relationship between the linguistic domains.

All in all, it can be concluded from the correlation analysis that on a global scale, the linguistic similarity measures are weakly positively correlated to the point that no measure can be used as a proxy for any other. The relationship between the variables is complex and suggests the existence of functional constraints, as signified by the bounded restrictions applied by the variables onto each other, with a large amount of variation inside the bounds. This means that each dimension of linguistic functionality explored here contributes unique information independently, and linguistic similarity should thus be evaluated multivariately. To unify the notion of similarity in a linguistic diversity context, one would ideally create a composite similarity index from these variables to represent the functionality of language.

In the next section, the impact of weighting measures of linguistic diversity with similarity will be explored, and as such, the use of a composite index would be ideal, but the lack of coverage in especially PHOIBLE and Grambank, as explored in Chapter 4, would severely limit the possible implementation of such an index of similarity. That leaves lexical similarity based on the ASJP data, which has been concluded to have the largest language coverage and speaker coverage generally, and close to full coverage on a speaker level, while being the most representative at the same time. Another possibility to combat the problem of lack of coverage would be

to use the phylogenetic similarity, but as has been extensively concluded through the analysis in this section, the measure suffers from not measuring any actual function of the languages, skewing the measures in cases where languages are similar but unrelated, which lexical similarity does not. Furthermore, as seen in the literature, lexical similarity is the best predictor of mutual intelligibility (Gooskens & Heuven, 2020), and if degrees of mutual intelligibility are used as the main criterion for differentiating different varieties into languages, then lexical similarity would reasonably contribute the most to that purpose. For that reason, it can be concluded that, by the extent of available cross-linguistic data and the behavior and properties of linguistic similarity measures, lexical similarity derived from the ASJP database is the most suitable operationalization of linguistic similarity on a global scale for the purpose of measuring linguistic diversity.

Summary

In this section, the pairwise similarity measures between all languages common to the ASJP database, Grambank, PHOIBLE and Glottolog have been compared and analyzed. It has been shown that morphosyntactic similarity follows a normal distribution, while phylogenetic similarity takes the value of zero for an overwhelming majority of datapoints, making it unsuitable for analysis. Rank correlation of the measures has shown that they are weakly associated and somewhat constrained by one another, meaning that no measure can be used as a proxy for the others.

5.2 Linguistic Diversity

In this last section, the actual linguistic diversity on a global scale will be investigated. Each country will be considered a setting, and based on this, the average similarity according to the respective linguistic dimension in each country will be assessed. Then, measures of linguistic diversity will be implemented, and finally, the impact of adding lexical similarity on how diversity is perceived will be considered.

5.2.1 Mean Pairwise Similarties

As outlined in Section 3.1, the pairwise similarities in each country were extracted from the respective similarity matrix and averaged to give a single measure of the diversity of the languages in that country. This is not a true measure of diversity, as it does not take neither richness nor relative abundance into account. Instead, it gives an idea of how similar the languages on average are, while minimizing the impact from lacking language and speaker coverage. The exact values calculated can be found in Table 7.2 in the appendix.

The distribution of these mean pairwise similarities across all countries of the world can be seen in Figure 5.9. One would suspect finding results similar to those in Section 5.1.2 Figure 5.6, given that they too explored pairwise similarity measures, with the key difference that in that case, the setting was global. Considering each country as a setting means imposing a geographical restriction such that geographically closer languages are more likely to occur together

and be eligible for similarity calculations. Indeed, considering central tendency, there is quite a lot of similarity, but the most notable difference is the more symmetric distribution of mean phylogenetic similarity, compared to the global setting. It certainly still has a large defining right skew, but the median similarity isn't 0, but instead 0.11 (mean = 0.15, IQR = 0.12, sd = 0.13) and as a matter of fact, there are only 3 countries where the mean phylogenetic similarity is 0, indicating that all languages belong to the different language families. These are Cocos Islands, British Indian Ocean Territory, and Samoa, with 2 respectively 3 languages present. Not only small island nations have such low phylogenetic similarity however; the United States, with its 402 languages has the fifth lowest mean phylogenetic similarity of 0.017. The intuitive explanation for this is that languages spoken in the same country are geographically closer, making it likely that they share a common ancestor, given that languages evolve through fractionalization along a dialect continuum. As such, the restrictions imposed by limiting the setting to countries means the similarity between related languages is considered to a larger extent, making the measure more useful in such a context.

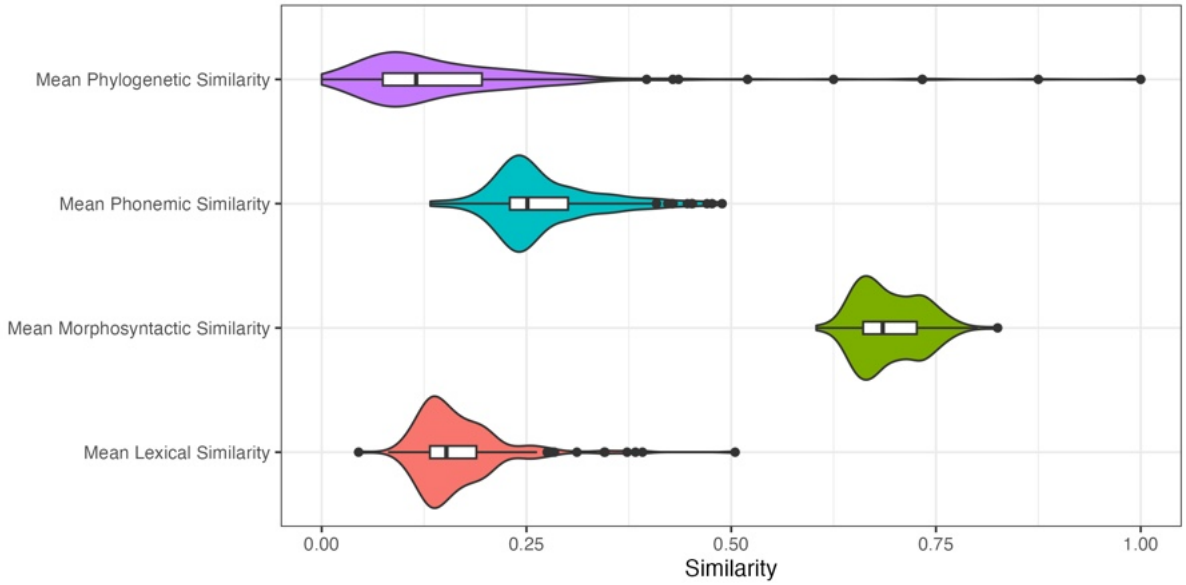

Figure 5.9: A violin plot showing the distribution of the measured mean pariwise similarity along the dimensions of Phlyogeny, Phonemic Inventory, Morphosyntax and Lexicon. Each datapoint corresponds to the value measured for a given country.

Furthermore, one can notice that both the Mean Lexical Similarity and the Mean Morphosyntactic Similarity show signs of bimodality, which for the Mean Morphosyntactic Similarity is especially notorious. This suggests that there are multiple underlying distributions at present: one group of countries where the languages tend to be more similar, and another group of countries where the languages tend to be less similar.

Looking at the maps displaying the mean similarity per country, as seen in Figure 5.10a, some quite distinct trends can be observed. Concerning Mean Phonemic Similarity, there is a quite clear pattern to observe where the similarity is notoriously high in countries around the equator, with an especially high concentration of countries with high similarity in the phonemic in-

ventory found in West Africa. The three countries with the highest Mean Phonemic Similarity are found here, namely Benin, Togo, and Burkina Faso (0.49, 0.48, 0.47), with languages sharing on average 50% of their phoneme inventories. As a matter of fact, countries in Sub-Saharan Africa generally show a remarkably high degree of phonemic similarity with the 17 most phonemically similar, and consequently, phonemically least diverse countries, being found here. This data is based on PHOIBLE, which has been concluded not to have the best of coverage, and considering language coverage on a country level as seen in 4.9, especially countries in Central Africa, Southern Asia, and Oceania have low coverage, reducing the reliability of these results. Still, the coverage of South America is impeccable in PHOIBLE, and here too, a clear concentration of phonemically similar languages can be found closer to the equator. As such, based on these measures of phonemic inventory similarity for the data at hand, it can be suggested that the phonemic diversity is the lowest around the equator. This is not the first time an association between temperature and humidity - which is implied to be regulated by distance to the equator - and the abundance of phonemes has been suggested: For example, Everett (2017) has shown that languages in colder climates use fewer vowels, while Wang, Wichmann, Xia, and Ran (2023) confirm a positive correlation between temperature and the use of sonorous speech sounds. This does not, however, imply that temperature causes the use of certain sounds, as shown by Hartmann, Roberts, Valdes, and Grollemund (2024). Independent of the discussion of cause and effect, a known observational pattern in the phonological domain is here reaffirmed, thus strengthening the suggestion that measuring phoneme inventory using Jaccard similarity from PHOIBLE meaningfully captures linguistic similarities on a sound level. In the context of linguistic diversity, it follows that taking phoneme similarity into account would reduce the perceived similarity more in countries closer to the equator.

Considering the Mean Morphosyntactic Similarity seen in Figure 5.10b, primarily two clusters of countries with noticeably high similarity can be discerned. One is found in South East Africa, with the highest concentration of morphosyntactic similar languages in Tanzania (0.762) and the Democratic Republic of Congo (0.755), meaning that on average, about 76% of grammatical features are shared between the languages in these countries. Interesting to note is that as one goes further away from these countries, the Mean Morphosyntactic Similarity gradually gets lower, denoting a core area of high similarity in South East Africa. Furthermore, this region with notably high similarity quite clearly corresponds to the areas where Bantu languages are extensively spoken, which are known for their broad typological similarities, characterized by having a large number of noun classes, and being agglutinative with a complex verbal inflection (Gibson, Guérois, Mapunda, & Marten, 2024). Furthermore, if one looks at the mean phylogenetic similarity in d, it is evident that there is a high degree of phylogenetic similarity in the countries of Southeast Asia, most notably Tanzania, indicating indeed that there is a large presence of related languages found there. All in all, this could explain why the morphosyntactic measure detects such a high level of similarity in precisely this geographical region.

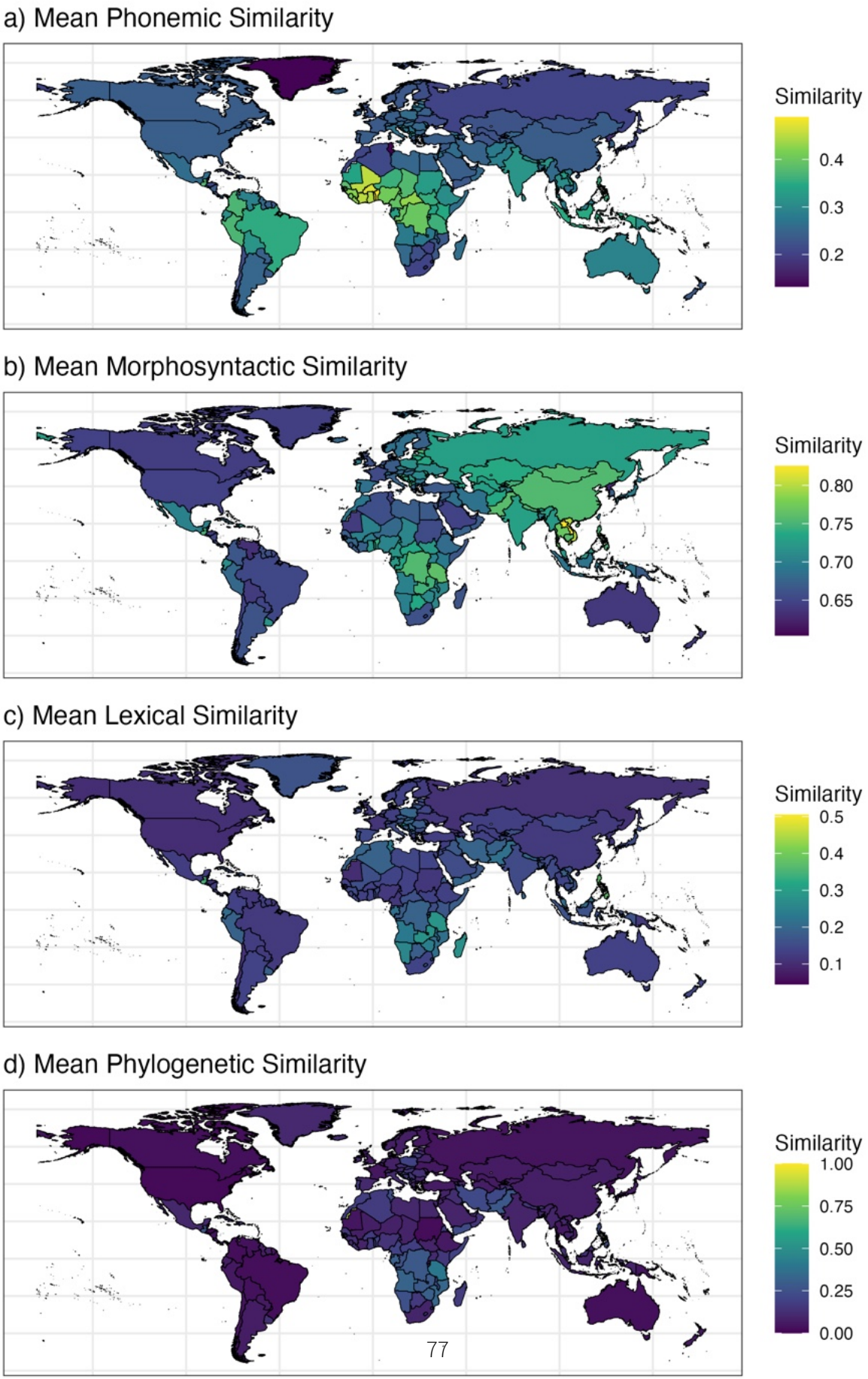

Figure 5.10: Maps showing the global distribution of similarities on the country level as the mean pairwise similarity between languages along the dimensions of a) Phonemic Inventory, b) Morphosyntax, c) Lexicon, and d) Phylogeny.

The second geographic area with a very high concentration of countries with high Morphosyntactic Similarity is Southeast Asia, where the two countries with the highest similarity in this regard are found, namely Laos (0.825) and Vietnam (0.801). From here, the similarity measures then gradually decrease across East- and Central Asia, into Western Europe and the Middle East, which are noticeably more diverse in this regard. The same goes in the other direction towards Australia; the similarity gets lower across Malaysia, Indonesia, and Papua New Guinea until reaching its lowest in Australia and New Zealand. Finding such an epicenter of high morphosyntactic similarity among the countries in Southeast Asia of all places is particularly exciting, since mainland Southeast Asia is the scene of a well-known sprachbund, where languages from multiple different families show numerous common characteristics resulting from language contact, the most important of which might be that the languages are isolating, lacking inflectional morphology (Enfield & Comrie, 2015). Now, looking back at the results from the clustering analysis in Section 5.1.1, it was suggested based on the morphosyntactic similarity measured between the European languages, that languages were to some degree clustered according to the extent the languages have an inflectional morphology. We now find two clear geographical concentrations of morphosyntactic similarity, on the one hand, the African area with many Bantu languages, which are known for being highly inflectional, and on the other hand, the established sprachbund in Southeast Asia known for languages lacking inflection. This strongly supports the idea that the measure meaningfully captures this fundamental similarity/difference in how languages are structured, which is a very important part of how languages function and should be reflected in how diverse an area is considered. We know that the languages in mainland Southeast Asia to a large extent are similar through this established sprachbund, which therefore should contribute to the setting being viewed as less diverse. This would not be captured if one only looked at, e.g., the phylogenetic relatedness of the languages, since the Mean Phylogenetic Similarity is observed to be very low in this region. Also, looking at the Mean Lexical Similarity, which potentially could take similarity through language contact into account, too, does not identify the languages in this area as being particularly similar. Thus, taking morphosyntactic similarity into account as it is operationalized in this thesis would indeed result in this area being viewed as less diverse, which is an accurate reflection of reality.

Unfortunately, the coverage of Grambank is fairly low in its current version, and as shown in Section 4.2.2, languages of Africa are noticeably underrepresented, and it was concluded that the number of features covered by Grambank gets lower the closer to the equator that one goes, making the measure less reliable for precisely these regions. Therefore, it would not be advisable in a global context to utilize the measure given the current state of data completeness, but as we have concluded that it indeed is an important ingredient in describing the functional diversity of countries, especially accounting for similarity arising through contact, contributing to adding more languages to Grambank to increase the coverage would be of paramount importance for improving measures of true linguistic diversity.

Further consideration of the global distribution of pairwise similarities along the dimensions of Lexicon and Phylogeny as seen in Figure 5.10 c and d, hints towards a more clear association between the measures than was found in Section 5.1.2, where the setting was global without any restrictions to which languages could be considered for similarity. Considering Spearman's rank correlations between the mean pairwise similarity of all 217 countries for which all mea-

sures where defined, as seen in Table 5.4, there is a very strong association between lexical and phylogenetic similarity with Spearman's $\rho = 0.75$ with quite narrow confidence intervals (95%CI = [0.67, 0.83]). So, in that regard, a strong positive monotonic association between these measures can be established: if the Mean Phylogenetic Similarity in a country increases, the Mean Lexical Similarity will tend to increase as well. Both Mean Phylogenetic Similarity and Mean Lexical Similarity are suggested to be moderately correlated with Mean Morphosyntactic Similarity (Spearman's ρ = 0.41, 95%CI[0.27, 0.53], and ρ = 0.42, 95%CI[0.30, 0.53]). At the same time, only for Mean Morphosyntactic Similarity is a correlation with Mean Phonemic Similarity significantly different from zero found (Spearman's ρ = 0.20[0.08, 0.34]. As such, it seems that lexical similarity is best aligned with phylogenetic similarity, similar to what the Cluster analysis on a European level concluded, while

	Lexical	Morphosyntactic	Phonemic	Phylogenetic
Lexical	0	0.42[0.30, 0.53]	0.03[-0.10, 0.17]	0.75[0.67, 0.83]
Morphosyntactic		0	0.20[0.08, 0.34]	0.41[0.27, 0.53]
Phonemic			0	0.04[-0.12, 0.16]
Phylogenetic				0

Table 5.4: Table showing the Spearman's rank correlation between mean pairwise similarity measures on a country level, n = 217. 95% confidence intervals estimated using bootstrapping with 1000 replicates are given in brackets to the right of the statistic.

While the global setting was a better condition for considering the association between the measures fundamentally, in practice, the geographical constraint of considering the measures on a country level radically impacts the association of the measures, which could be relevant for considering the linguistic diversity within that setting. However, by only looking at the mean pairwise similarity, the internal structure of the distribution of similarities within each country is not reflected. To understand this better, digging deeper into how the measures behave within and across single countries and regions would be an exciting avenue for further research.

5.2.2 Naive Diversity

Now, we shall take a step away from considering diversity according to how different languages are and function, and instead naively determine global diversity, assuming that all languages are equally dissimilar. This is how linguistic diversity is commonly assessed, meaning that any variety of speech by virtue of being defined as a language contributes equally to diversity, whether that is because it is unintelligible to all other varieties or for political reasons, thus making the definition of language as opposed to a variety or dialect a non-trivial question.

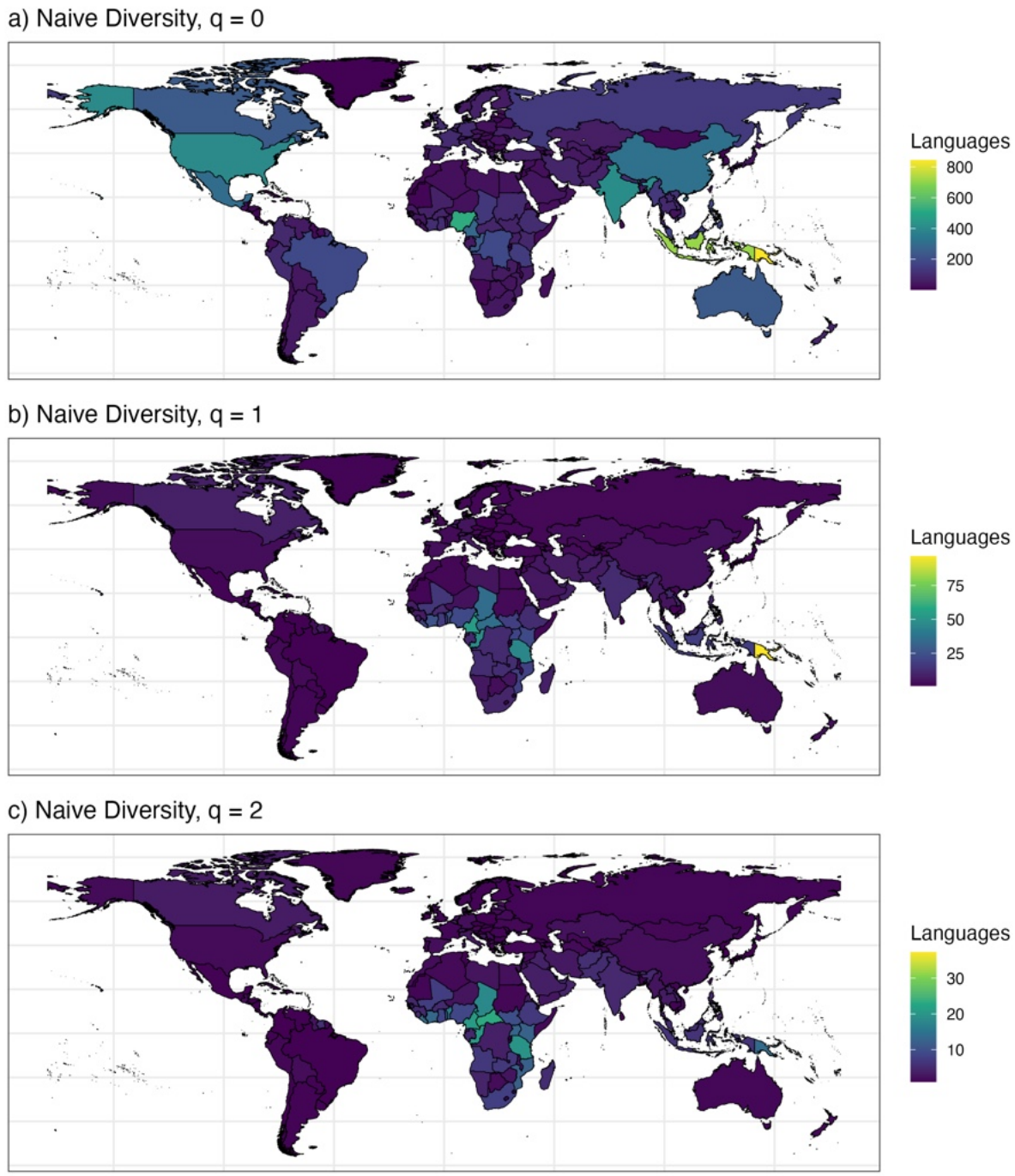

Figure 5.11: Maps depicting naive global linguistic diversity according to three degrees of sensitivity to dominating languages, equivalent to a) Richness, b) Exponent Shannon, c) Inverse Simpson.

Start by disregarding the evenness of the distribution of speakers and simply consider the total number of languages found in each country, which can be seen in Figure 5.11a. According to the framework of Leinster and Cobbold, the domination sensitivity parameter q equals 0,

which is equivalent to richness, or simply counting the languages found. From such a perspective, the most diverse country, as seen in Table 5.2.2 is Papua New Guinea, which is home to 845 languages, followed by its neighbor Indonesia with 712 languages. Behind them, there is a jump down to 522 languages in Nigeria, followed by 414 languages in India and 402 in the USA. Exact values for all countries can be found in Table 7.3 in the appendix, but generally, one can see that across all continents, there is some country that appears fairly diverse, whether Russia in Europe (142) or Brazil (184) in South America. The diversity is, so to say, spotted across the globe.

Rank	Naive Diversity, q = 0	Naive Diveristy, q = 1	Naive Diversity, q = 2
1	Papua New Guinea (845)	Papua New Guinea (96.6)	Vanuatu (37.2)
2	Indonesia(712)	Vanuatu (54.3)	Solomon Islands (29.9)
3	Nigeria (522)	Cameroon (53.6)	Central African Republic (23.2)
4	India (414)	Congo (47.6)	Cameroon (21.0)
5	United States (402)	Tanzania (45.3)	Congo (20.9)
6	China (317)	Solomon Island (42.3)	Tanzania (19.8)
7	Mexico (302)	Central African Republic (36.0)	Chad (18.1)
8	Cameroon (289)	Chad (35.2)	Benin (15.9)
9	Australia (242)	Benin (26.1)	Liberia (14.0)
10	Canada (242)	Côte D'Ivoire (25.4)	Papua New Guinea (13.8)

Table 5.5: A table depicting the 10 most diverse countries according to naive diversity for q = 0, 1, and 2 in the Leinster-Cobbold Framework. The effective number of languages in each country is written in parentheses.

When the sensitivity parameter is then increased to q = 1, as seen in 5.11b, the number of speakers is taken into account by penalizing uneven distributions of relative abundance. This gives the exponentiated Shannon's diversity index, commonly used in ecology, which indicates the number of equally distributed languages that would produce the same level of diversity. Thus, given the distribution of speakers across the 845 languages in Papua New Guinea, the level of diversity corresponds to 96.6 equally distributed languages, which makes Papua New Guinea perceived as the most diverse by a wide margin. The second most diverse country is Vanuatu, with a diversity equal to 54.3 equally distributed languages, followed by Cameroon (53.6), Republic of the Congo (47.6), and Tanzania (45.3). What can be noticed is that many of the countries which, according to richness, were perceived as very diverse now have fallen into obscurity. Indonesia measures a diversity of 17.8, and India too maintains some degree of diversity (11.6), while the USA has an equivalent of 3.5 languages. The reason for this is the dominance of English, with a relative abundance of 74.2%, followed by Spanish at 13.1%. Indonesia lacks an as clearly dominating language, with the most widely spoken language of Javanese accounting for 35.3% of the speakers. In Papua New Guinea, the dominant language Tok Pisin is even smaller, making up a mere quarter (25.4%) of the total speaker number. The geographical trend that can be observed here is that in South America, Europe, and Asia, the diversity is perceived as much lower, while in Sub-Saharan Africa, many countries are perceived as more diverse. This is not surprising, given that the Americas are so dominated by the world

languages of Spanish and English as a colonial legacy, making it appear less diverse under this measure. In Europe and Parts of Asia, countries as nation states were formed along ethnic - and to a large degree linguistic - lines, resulting in one very dominating language, an extreme example found in the data being Poland, where Polish makes up 97.% of the linguistic landscape. In Sub-Saharan Africa, on the other hand, national borders were drawn fairly arbitrarily, largely disregarding cultural and ethnic lines, resulting in a much more even distribution of languages. Thus, it can be theorized that the lack of dominating national languages in Sub-Saharan Africa makes these countries be perceived as more diverse for larger qs compared to other parts of the world as a direct result of their colonial heritage.

If the sensitivity parameter is cranked further to q = 2, the measure becomes equal to the inverse Simpson index, as it is commonly referred to in ecology, which is equivalent to the reciprocal of 1 minus Greenberg's Linguistic Diversity Index. This index was developed by Greenberg and has a standing tradition in linguistic research. What can be observed after increasing the sensitivity to dominant languages is that Papua New Guinea stops being perceived as the most diverse country, falling to rank 10 with a diversity equalling 13.8 equally distributed languages. Instead, Vanuatu is perceived as the most diverse country, only being perceived as slightly less diverse (37.2), followed by Solomon Islands (29.9), the Central African Republic (23.2) and Cameroon (21.0). The reason for this fall in perceived diversity of Papua New Guinea is that there is only one dominating language, Tok Pisin with 25% of speakers, followed by Enga, spoken by 5.54 % of the population. So while Papua New Guinea has an incredible richness which is fairly evenly distributed, making it perceived as very diverse for small q's, but as the importance of richness decreases for the measure, the dominating presence of Tok Pisin takes over and heavily penalizes the diversity. In Vanuatu, on the other hand, there is no dominating language according to relative abundance, with Lenakel spoken by 5.75% of the population, followed by 5% speaking Apma and another 5% speaking Bislama. Vanuatu is, however, with its 115 languages not as linguistically rich as Papua New Guinea, which is home to 845 languages. This raises the question: which country is to be considered more diverse, Papua New Guinea or Vanuatu? The answer depends on what the researcher investigating the diversity is interested in. If one considers richness to be the most important factor for analyzing diversity, Papua New Guinea will be the most diverse, and setting q to 0 will make the most sense. If one is interested in which countries' languages are the most evenly distributed, Vanuatu should be considered more diverse and setting q to 2 for the measure is advised. If one considers both richness and relative abundance to be of approximately equal importance, q should be set to 1, resulting in Papua New Guinea being perceived as the most diverse.

The framework of Leinster and Cobbold allows us to understand this tradeoff between dominance and richness by plotting the diversity for increasing values of q, as can be seen in Figure 5.12. Papua New Guinea (PG) starts as the most diverse country for q = 0, with the USA (US) being perceived as fairly diverse with its high richness, and Vanuatu as not especially diverse due to its low richness. As q increases, the measured diversity of the USA falls sharply due to the dominance of English, and at a q of approximately 0.3, the even distribution of languages in Vanuatu makes it perceived as more diverse than the USA. Papua New Guinea also falls sharply for increasing q, but not as quickly as the USA, since Tok Pisin isn't as dominant as English. First at a q of approximately 1.25, the diversity of Vanuatu is perceived as more diverse than Papua New Guinea and also the most diverse country in the world. As such, how we look at linguistic

Diversity fundamentally depends on how we value the importance of relative abundance, and its addition can radically impact how it is perceived.

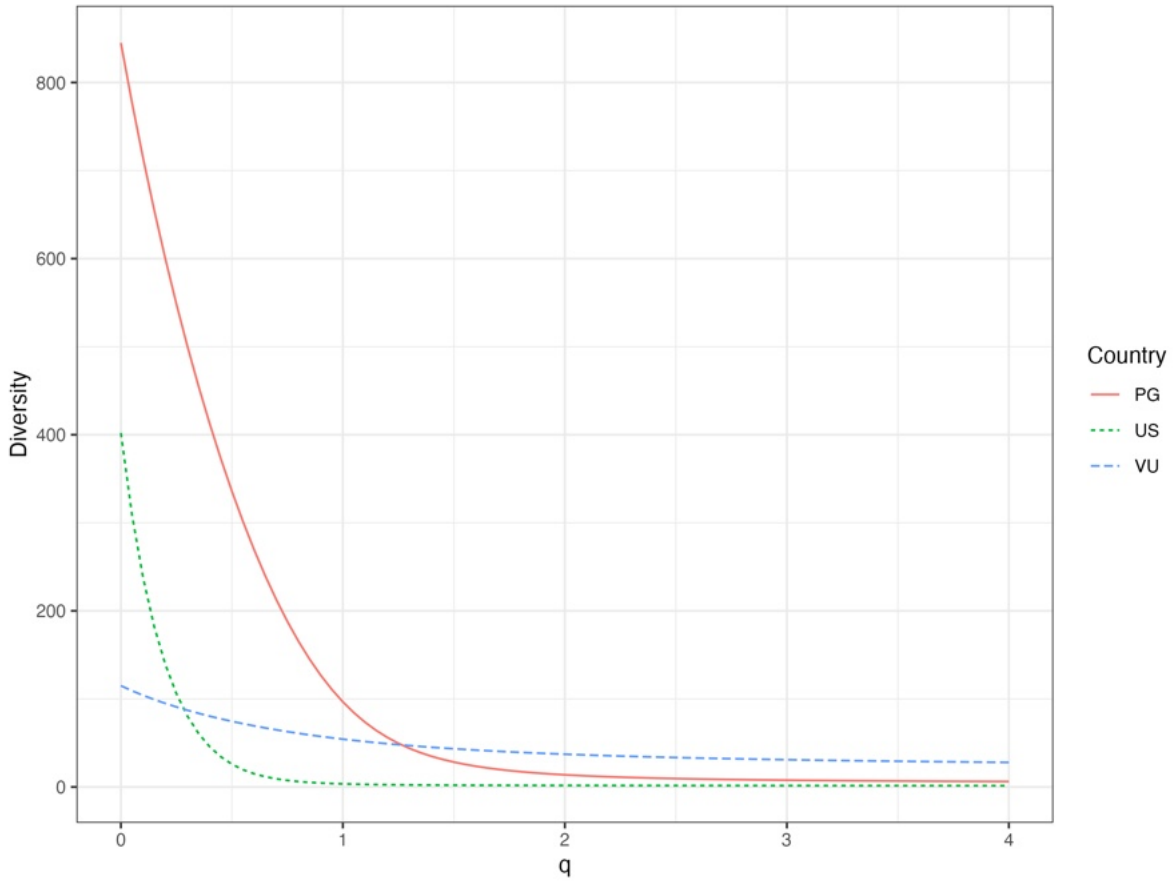

Figure 5.12: A plot showcasing the niave diversity profiles of the three countries Papua New Guinea (PG), the USA (US), and Vanuatu (VU)

5.2.3 Non-Naive Diversity

As seen in the previous section, the relative abundance of speakers is of crucial importance for how linguistic diversity is perceived. It was, however, assumed that all languages were equally dissimilar, which in reality isn't the case, and this has real implications for how we look at diversity. To understand why, consider the two countries of the Czech Republic and Slovakia. According to the data, the Czech Republic has a richness of 33, and the most widely spoken language is Czech, accounting for 95.4% of the population, followed by Slovak, spoken by 1.56%. The dominance of Czech gives a very low naive diversity for q = 1, at only 1.33. In Slovakia, the richness is 27, while the most spoken language is Slovak, spoken by 86% and Hungarian, spoken by 8.35%, giving it a naive diversity with q = 1 of 1.87. Now, let's imagine that these countries joined together into the former country of Czechoslovakia. This country will have a richness of 39, so the naive diversity for q = 0 will not increase much, but this country is now split on two major languages, Czech, accounting for 61.5% of speakers, and Slovak, accounting for 31.7% of speakers. As such, the naive diversity at q = 1, equals 2.696, which is a doubling of the diversity in the Czech Republic alone, and close to a 50% increase to Slovakia, so the new state will be seen as substantially more diverse. The problem is, however, that Slovak and Czech exist on a dialect continuum and the two standardized varieties exhibit a very high degree of mutual intelligibility (Gooskens et al., 2018). Considering the lexical similarity between the languages as measured in this thesis, the languages have a similarity of 0.69. So if the mutual intelligibility criterion is used to differentiate between two languages, Slovak and Czech is best analyzed as one language, meaning that the hypothetical Czechoslovakia is dominated by one language, spoken by 93.5%, and as such much less diverse than the naive measure would indicate.

To address problems arising from similar languages being differentiated on political grounds, the non-naive similarity from the Leinster-Cobbold framework is applicable, weighting the measure for different q's with the pairwise similarities, which would resolve the hypothetical scenario with the joining of the Czech Republic and Slovakia. As such, if two languages have a similarity of one, they contribute to diversity as a single language; if they have a similarity of 0, they contribute to diversity to the same extent they would in the naive case. Using lexical similarity derived from the ASJP database, which had the best coverage, the global linguistic diversity for q = 0, 1, and 2 can be seen in Figure 5.13. Here, it is important to notice that q = 0 is not the same as richness, since the languages still are weighted according to pairwise similarities, so the measured value will be the effective number of completely dissimilar languages, which stays true for increasing values of q, with the addition that the presence of dominant languages is taken into account. According to these non-naive diversity measures, the same general patterns as for the naive measure can be observed; the most diversity can be found in Sub-Saharan Africa and Oceania, and as the sensitivity to dominant species increases, the perceived diversity goes down in large parts of the world. In the naive case, the perceived diversity was, however, very sensitive to increases in q, and we saw Papua New Guinea go from the most diverse country to the 10th most diverse country by increasing the sensitivity to dominant languages. In the non-naive case, however, the most diverse countries in the world are almost exclusively found in Africa for all three values of q, and Cameroon is consistently perceived as the most diverse country, with Papua New Guinea as the second most diverse, as seen in Table 5.2.3. As such, when considering all aspects of diversity: richness, relative abundance, and sim-

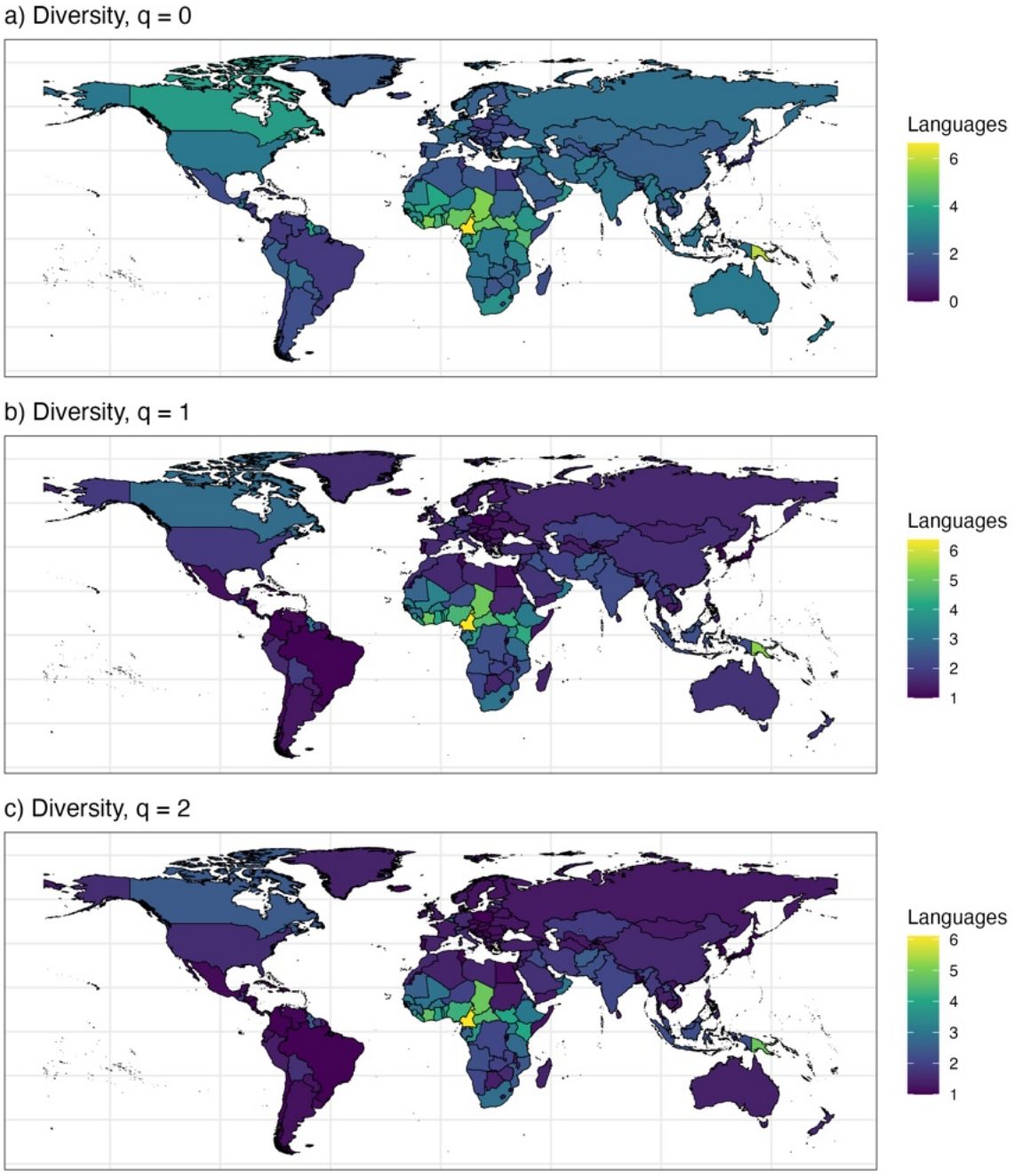

Figure 5.13: Maps depicting non-naive global linguistic diversity weighted by lexical similarity with values of q = a) 0, b) 1, and c) 2. Yellow color indicates more diversity

ilarity, Cameroon is unquestionably the most diverse country in the world. The general global trend that can be observed from taking language similarity into account seems to be a clearer shift towards western and central Sub-Saharan Africa being perceived as the greatest hot spot of linguistic diversity instead of Oceania, where not only numerous and evenly distributed languages are found, but also very dissimilar ones.

Rank	Non-naive Diversity, q = 0	Non-naive Diversity, q = 1	Non-naive Diversity, q = 2
1	Cameroon (6.64)	Cameroon (6.36)	Cameroon (6.10)
2	Papua New Guinea (5.95)	Papua New Guinea (5.42)	Papua New Guinea (4.86)
3	Chad (5.37)	Chad (5.10)	Chad (4.84)
4	Côte D'Ivoire (5.32)	Côte D'Ivoire (5.00)	Côte D'Ivoire (4.69)
5	Nigeria (4.92)	Central African Republic (4.66)	Central African Republic (4.48)
6	Central African Republic (4.89)	Nigeria (4.55)	Nigeria (4.25)
7	Belgium (4.78)	South Sudan (4.35)	kenya (4.06)
8	South Sudan (4.77)	Kenya (4.30)	South Sudan (4.01)
9	Kenya (4.55)	Belgium (4.23)	Liberia (3.85))
10	Ghana (4.35)	Ghana (4.07)	Ghana (3.82)

Table 5.6: A table depicting the 10 most diverse countries according to non-naive diversity for q = 0, 1, and 2 in the Leinster-Cobbold framework. The effective number of languages in each country is written in parentheses.

A country that is perceived as notably less diverse is Vanuatu, which was ranked as the most diverse country in the world for q = 2 in the naive case. When taking similarity into account, it is only ranked as the 11th most diverse country. A hint towards why is given by looking at the mean lexical similarity, which is substantially higher for Vanuatu at 0.26, compared to Papua New Guinea and Cameroon, both having a mean lexical similarity of 0.12. It thus follows that although the languages of Vanuatu are more equally distributed, those languages are also more similar, resulting in the country being correctly perceived as less diverse. Furthermore, the diversity profile for the three countries can be plotted as seen in Figure 5.14, and indeed, for small values of q, both Papua New Guinea and Cameroon are perceived as more diverse than Vanuatu. Whereas in the naive case, Vanuatu overtook Papua New Guinea in diversity at q = 2.5, the sensitivity to dominant languages needs to be cranked to about 5.5 before Papua New Guinea is perceived as less diverse than Vanuatu in the non-naive case. This is because now, not only the number of languages, but also their distinctiveness have to be accounted for when measuring the diversity, which is conceptually necessary if one truly wants to capture linguistic diversity.

Until now, we have considered the impact of going from naive diversity to non-naive diversity using all the data available. Although this gives as exact a picture of global diversity as possible, it suffers from the bias introduced by the underlying data for deriving the similarity measurement. While the ASJP database is extensive, its coverage isn't perfect, and looking back at the country coverage in Figure 4.19, it is clear that across Eurasia, Africa and Oceania, a few countries have a somewhat lower coverage, meaning that some of the more widely spoken languages in those countries, most notably in Mauretania, won't be considered for measuring non-naive diversity, thereby biasing it. However, the impact of considering similarity can still

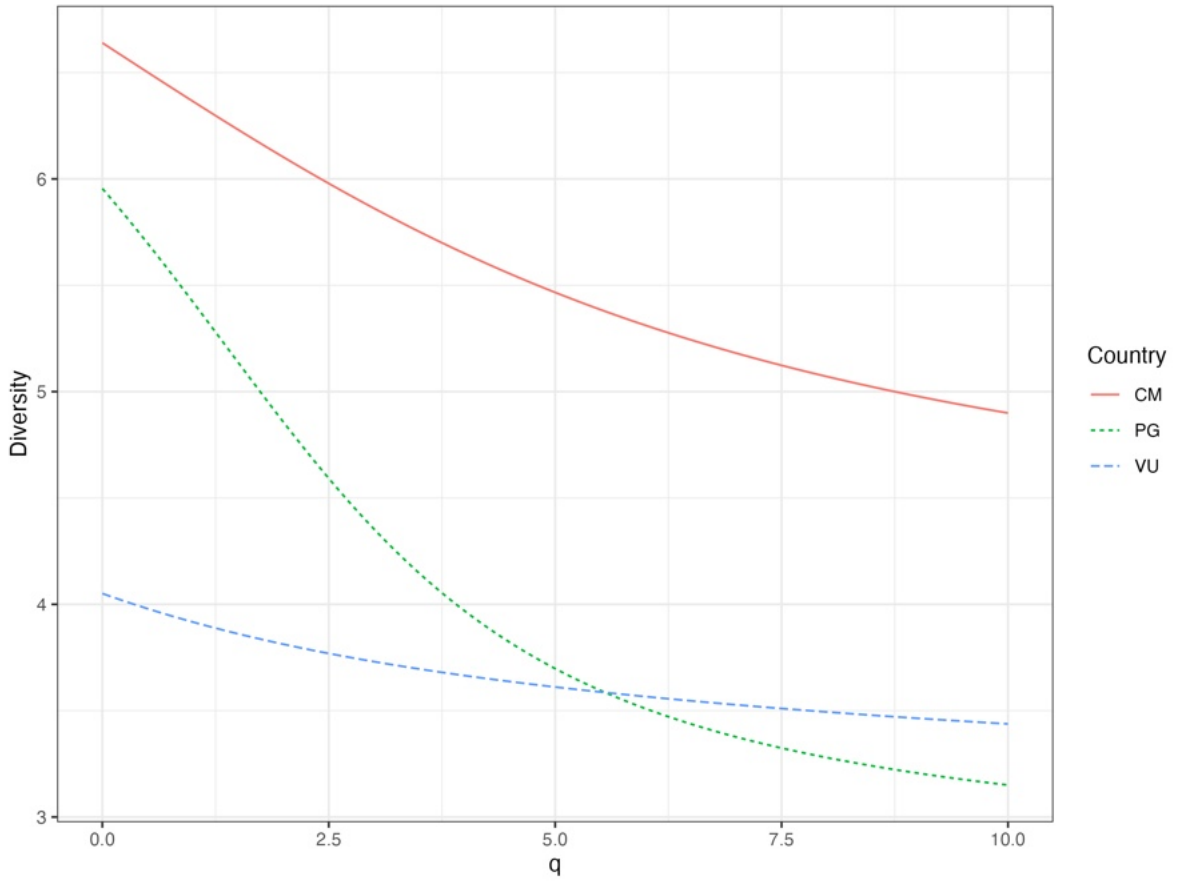

Figure 5.14: A plot showcasing the non-naive diversity profiles of the three countries Cameroon (CM), Papua New Guinea (PG), and Vanuatu (VU).

be estimated by only considering the languages for which similarity measures can be derived. By calculating both naive and non-naive similarity for this smaller set of data, the difference between these measures will be the impact of considering similarity. By staying close to linguistic tradition and looking at q = 2, equivalent to the Hill number version of the A- and B-Index as envisioned by Greenberg by using lexical similarity as a resemblance factor, the relationship between the naive and non-naive diversity is given by ranking all countries according to diversity and plotting against each other as seen in 5.15.

If there was no impact on the perceived diversity by weighting according to similarity, all countries would be expected to stay on the black 45-degree line, signaling that both the naive and non-naive measure assigns all countries the same ranking. As can be seen, this is not the case, and there are quite a lot of countries that gain and lose rankings. By calculating the Spearman rank correlation, the agreement between the measures is given, such that a perfect correlation equals 1, meaning that all languages lie on the black line, and 0, meaning that there is no association between the ranking of the two measures. For the weighted and non-weighted diversity measure, Spearman's ρ equals 0.96[1], 95% CI[0.94, 0.98], and as such, the impact on

[1]A correlogram with all correlations can be found in Table 7.6 in the Appendix

how the diversity is perceived is very small. For comparison, the Spearman's ρ between rankings of naive diversity for q = 0 and q = 2, in effect, the impact of adding relative abundance, amounts to 0.43, 95% CI[0.31, 0.53], which is much larger in comparison. Still, one can see that for a select number of countries, the rankings change quite drastically when measuring diversity and weighting by similarity, the most prominent case of which is Madagascar. According to the naive measure of diversity, Madagascar is assigned rank 197, but when using the non-naive measure, Madagascar is given the rank of 115, resulting in a total loss of 82 ranks, equalling a drastic change in how the diversity of Madagascar is perceived.

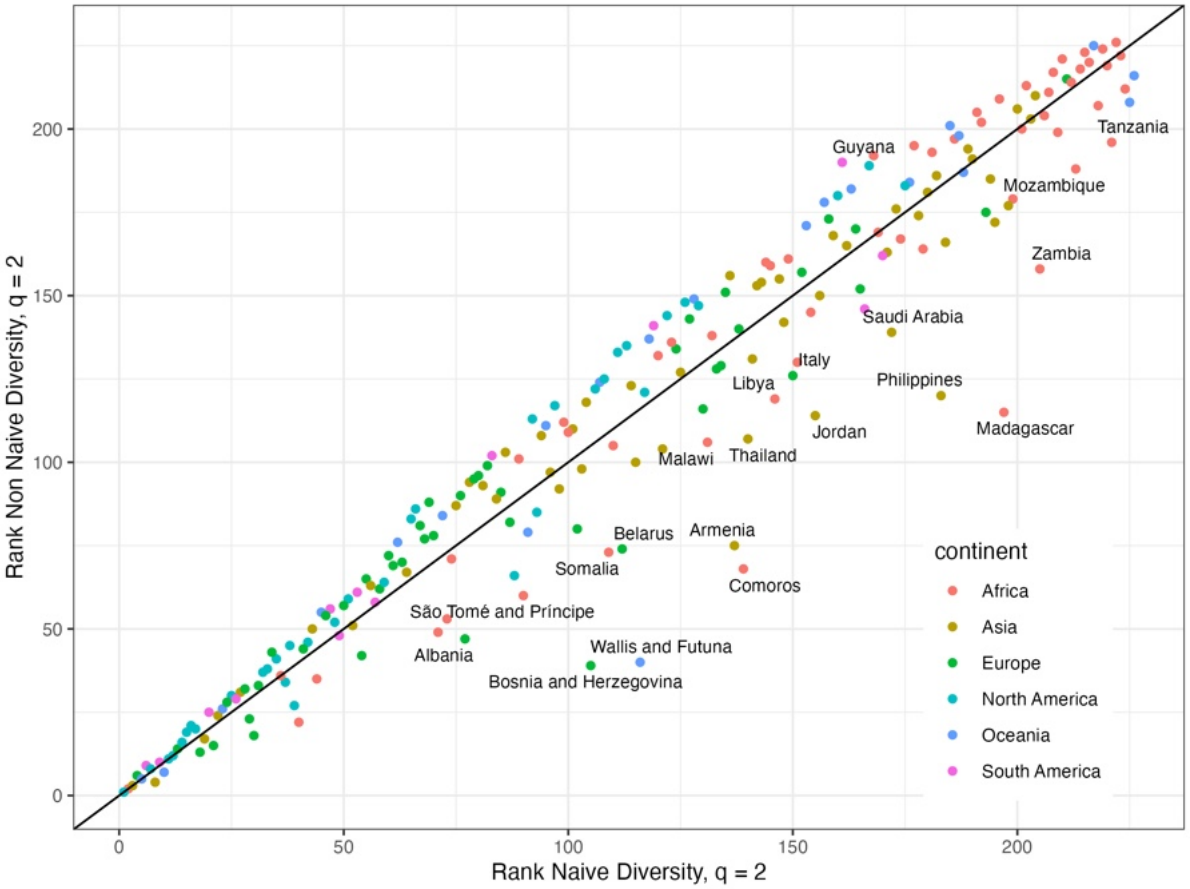

Figure 5.15: A scatterplot showing the relationship between the ranking of countries according to Naive and Non-naive Diversity, q = 2. A higher rank corresponds to more diversity. Countries are colored according to continent, and the names of the 20 countries for which the largest rank difference between the measurements is observed are printed. The straight line shows the expected position of the points if there were no difference in ranking between the two measurements.

The reason for this massive change in perceived diversity of Madagascar lies, of course, in the similarity of the dominating languages spoken there, since Madagascar has a mean lexical similarity of 0.28. The 10 most spoken languages are different varieties of Malagasy, the most dominant being the Merina variety, accounting for 38% of the relative abundance, and altogether,

these very similar languages, although with varying degrees of mutual intelligibility (Bouwer, 2007) make up about 93% of all languages in Madagascar. Again, similar to the argument regarding Czech and Slovak, if one considers the languages as completely separate entities, as with the non-naive measure, the country will appear as very diverse if granting every variety a language status. If one goes in the other direction and lumps all languages together as a single language, which traditionally has been done, then the country is perceived as not diverse at all, since the Malagasy macro language will dominate. This is too is not correct, however, as there are differences between these varieties; the largest lexical similarities found in the Malagasy data are between Northern and Southern Betsimisaraka, which have a similarity of 0.90. At the same time, the varieties of Nara and Masikoro only have a lexical similarity of 0.62. This can be compared to the lexical similarity between Czech and Slovak, which amounts to 0.68. Indeed, by adding the weighting according to similarity and defining language on a spectrum, we can account for differences between languages and thus need not argue between if varieties should be considered different languages or not; if all adequately measured, this framework allows us to measure the diversity while taking similarity into account, which in the case of Madagascar reduces the perceived diversity, reflecting that the dominating varieties indeed are similar.

Further analysis of the impact of measuring similarity non-naively supports the conclusion that the measure provides a better representation in cases where languages are very similar by lowering the perceived diversity. Among the languages losing most ranks, we find, for example, Bosnian and Herzegovina, which is at the epicenter of the debate between dialects and language. While the Southwest Slavic dialect continuum saw the institutionalization of a standardized language called Serbo-Croatian, the breakup of Yugoslavia saw the formation of the nation-states Croatia, Bosnia, Serbia, and Montenegro, and with them, linguistic fractionalization according to nation-ethnic lines. As a result, the Croats view themselves as speaking Croatian, the Bosniaks as speaking Bosnian, and the Serbs as speaking Serbian, although the varieties in themselves are mutually intelligible (Bailyn, 2010). With Bosnia and Herzegovina being a multiethnic state of Bosniaks, Croats, and Serbs, if these languages are considered equally dissimilar, the country will be perceived as very diverse compared to the actual linguistic situation, where most of the speakers in the linguistic structural sense are speaking the same language. In this case, the lexical similarity between the most widely spoken variants, Bosnian and Croatian, amounts to 0.87, resulting in Bosnia and Herzegovina losing 66 ranks from 115 to 82. Here, it should be noted that the impact of considering the similarity on the perceived diversity of Bosnia and Herzegovina is probably underestimated, since the ASJP lacks an entry for Serbian as seen in Figure 4.17, which means this proportion of speakers are not accounted for when assessing the diversity according to any of the measures.

Other examples that showcase how accounting for similarity impacts the perceived similarity and trivializes the problem introduced by dialects as opposed to languages can be found here, as we see Albania, which is split on the Gheg and Tosk varieties of Albanian, see many rank shifts. Equally, the diversity of Armenia, with a majority of speakers split on eastern and western Armenian, is reevaluated as a result of using the non-naive measure, and Italy, with its many regional languages, which are under constant scrutiny of whether they should be considered languages in their own right or dialects. As we have seen, this discussion becomes trivial if only the similarity between the languages is accounted for accurately. Although the impact of accounting for similarity overall is fairly small, this shows that accounting for similarity is neces-

sary to accurately model linguistic diversity, assuming that one can reliably measure similarity. The results of this thesis suggest that the most reliable, yet still extensive, measure is lexical similarity operationalized as normalized Levenshtein similarity between phonetic representations of core vocabulary from the ASJP database. As we have also seen, this data does not cover everything, and there are still blind spots in the data that need to be accounted for to truly assess the entire global linguistic diversity accurately. Hopefully, the rise in development of large-scale cross-linguistic databases that has been seen in recent years will continue, and the data sources explored here to derive measures of linguistic similarity will be expanded further to cover as many languages as possible, as long as there are still languages to be described. Only by collecting more and accurate data are we truly able to correctly model diversity, and thus also employ quantitative techniques to better understand the phenomenon of linguistic diversity. As such, the main focus of further research on linguistic diversity should be to continue solving the bottleneck that is cross-linguistic data.

6 Conclusion

Finally, this last chapter shall conclude the thesis by first synthesizing the most important findings and conclusions from the previous two chapters to answer the research questions posed at the beginning of the thesis. Then, the limitations of the work will be addressed while providing directions for further research.

6.1 Conclusions

This thesis set out to investigate global linguistic diversity by incorporating measures of linguistic similarity. This has been achieved in three parts. First in the descriptive part, a birds-eye view on linguistic diversity has been taken by describing the state of linguistic data that can account for the critically important relative abundance of languages. By considering the Ethnologue and Joshua project data, a baseline is created from which naive linguistic diversity that goes beyond simply richness can be calculated. Having gained insights into the global composition of languages and speakers, a judgment about the coverage of the cross-linguistic databases could be made, allowing multiple important conclusions about the composition of the most extensive cross-linguistic databases to be drawn. Firstly, high language coverage does not necessarily imply high speaker coverage. While Grambank was concluded to cover more languages than PHOIBLE, the missing coverage of languages such as Urdu and Spanish means that the speaker coverage is substantially lower and disregards much of South America, making it highly unsuitable for deriving similarity measures that are to be incorporated with relative abundance data in a global context. Furthermore, the distribution of languages across macroregions in PHOIBLE turned out to be highly unrepresentative, heavily underrepresenting Papunesia, which is a hotspot of linguistic diversity and very important in this context. Of the three databases investigated, the ASJP database turned out to have the best coverage of both language and speakers, covering more than 90% of speakers while only underrepresenting Eurasian languages somewhat, making it the most appropriate database to derive similarity measures in a global context. As such, research question 1:

> what data is available to calculate linguistic similarity/distance on a global scale, and how well does such data cover languages, speakers, and countries?

can reliably be answered: only data to derive lexical similarity from the ASJP database truly exists on a global scale. And even then, the ASJP database has a few countries with particularly low coverage. Still, the insights gained through this extensive descriptive analysis show that there are numerous countries with high coverage in all databases, meaning that if a diversity analysis takes place on a regional scale for select countries, then all databases can be utilized to account for linguistic similarity.

Moving on from the descriptive part of the thesis and studying the behavior of the measures derived from the databases, it was concluded that lexical similarity and morphosyntactic similarity measures meaningfully capture similarity. Primarily similarity induced from common ancestry, but also similarities derived through language contact. It was thereby shown that phylogenetic similarity is insufficient as a measure in itself for the purpose of measuring similarity, since it fails to account for language contact. Still, it has repeatedly been suggested that phylogenetic similarity can be used as a proxy for actual similarity. Through the correlation analysis of pairwise similarities between more than 1000 languages globally, it was found that only weak correlations exist between the various measures, although phylogenetic similarity generally correlates the most strongly, thus partially refuting this claim. To answer research question number 2:

> How do linguistic distance/similarity measures based on different typological aspects of language correlate in a global context? Could any measure be used as a proxy for the other?

It can be concluded that no one measure can be used as a proxy in a global context, but this does not exclude that in specific cases, where most languages are somewhat related or a sprachbund is found, one measure could perhaps be used as a proxy if lacking available data for the other. Globally, however, no strong support suggesting that a measure can work as a proxy is found. Still, it can be concluded that relationships between the measures exist, but it is highly complex, with the different measures bounding each other, restricting the similarity to a certain limit, but never implying its value.

Based on the conclusions derived in the first two parts, the thesis culminated in the third part, where the Leinster-Cobbold framework was implemented using lexical similarity derived from the ASJP database to answer research question number 3:

> How does adding linguistic distance/similarity to diversity measures impact how global linguistic diversity is perceived?

Indeed, it is concluded that adding linguistic similarity is a key ingredient to assessing linguistic diversity, as posited by Grin and Fürst (2022). While linguistic similarity does not radically change the way linguistic diversity is perceived in the same way that taking relative abundance into account does, its impact is obvious in the way it emphasizes Sub-Saharan Africa as the most linguistically diverse region, where not only countries with many evenly distributed languages are found, but also very dissimilar ones, among which Cameroon emerges as notably diverse. And although most countries are perceived as approximately equally diverse relative to other languages when adjusting for similarity, quite a few countries are perceived as radically less diverse if the dominant languages found there indeed are more or less mutually intelligible, as in the case of Bosnia and Herzegovina, and Madagascar. Therefore, accounting for similarity is critical for modeling linguistic diversity accurately, and the Leinster-Cobbold framework together with the ASJP database makes this possible.

6.2 Limitations and Further Research

While this thesis has been an ambitious project, successfully filling an important methodological gap in the literature on linguistic diversity and providing a better overview of the data existing in the field, it comes with some limitations. Due to the extensiveness of its scope, the implementation of measures of linguistic similarity were selected in such a way that they were computationally effective and easy to implement. This means that measures that are better models of similarity exist or can be developed, such as a weighting scheme taking the similarity of phonemes into account for phonemic similarity, or the frequency of occurrence of phonemes in the given language. Also, some of the shortcomings of phylogenetic similarity measures could be mended if the phylogenetic relationship between languages with exact separation times on a large scale could be inferred while also taking loans into account, using e.g., Bayesian phylogenetic analysis (see Hoffmann, Bouckaert, Greenhill, and Kühnert (2021); Neureiter, Ranacher, Efrat-Kowalsky, et al. (2022)). Since these measures are the basis for assessing how similar any two languages are, improving them will be associated with better measures of linguistic diversity, presenting a possible avenue for further research.

Furthermore, the largest limitation of the work comes down to the available data. While the endeavor of this thesis was enabled by the recent expansion in publicly available cross-linguistic data, it has been shown that it is far from all-encompassing, and the way that diversity is calculated in the Leinster-Cobbold framework means that if similarity between any two languages can not be accounted for, they are completely disregarded. As such, unless the data coverage is extended to cover all, or almost all, existing languages, non-naive linguistic diversity can't be fully implemented on a global scale. Therefore, the most pressing need for further research lies in the domain of data collection, extending the existing databases to increase their coverage. If that is done, then it opens up to considering more linguistic domains than only lexicon, potentially creating a combined index of linguistic similarity.

Still, we have shown that there are 1012 languages for which relative abundance data exist, and all similarity measures can be derived. Therefore, the road lies open to explore these languages further to see which countries are adequately covered, possibly investigating how using other similarity measures than lexical similarity or combinations of similarity measures affects how the diversity is perceived.

Finally, we have here taken a wide global view on diversity, showing that the Leinster-Cobbold framework can and should be adapted to linguistics. But as was shown, the linguistic diversity isn't a geographically homogonous concept. Therefore, further research should also look into other domains, such as diversity on a regional level, or in the digital space, and consider its impact in those contexts as well. In the end, measuring something is only the first step of analysis; we now have a conceptually satisfactory way of quantifying linguistic diversity, but to refer back to (Greenberg, 1956) again, we want to

> "...eventually to correlate varying degrees of linguistic diversity with political, economic, geographic, historic, and other non- linguistic factors." (p.1)

While this thesis bridges the gap between ecology, data science and sociolinguistics by providing methodological and empirical advances in how to model linguistic diversity, the next

step must be to correlate these measures with other factors, such as digitization. By doing so, we can increase our understanding of why and how linguistic diversity is changing, and contribute to preserving it, which is not only an academic concern, but our societal responsibility.

7 Appendix

	ISO	Name	Language Phoible (%)	Speaker Phoible (%)	Language Grambank (%)	Speaker Grambank (%)	Language ASJP (%)	Speaker ASJP (%)
1	AD	Andorra	100.00	100.00	66.67	68.10	100.00	100.00
2	AE	United Arab E.	63.33	74.11	56.67	80.35	83.33	94.58
3	AF	Afghanistan	47.27	46.40	20.00	25.92	70.91	86.81
4	AG	Antigua and B.	100.00	100.00	50.00	97.16	50.00	97.16
5	AI	Anguilla	100.00	100.00	66.67	6.67	66.67	6.67
6	AL	Albania	75.00	17.42	66.67	16.93	83.33	99.47
7	AM	Armenia	60.00	63.05	60.00	62.24	86.67	99.93
8	AO	Angola	37.93	50.06	31.03	66.33	67.24	87.95
9	AR	Argentina	86.27	96.00	58.82	9.96	90.20	99.66
10	AS	American Samoa	71.43	14.13	100.00	100.00	100.00	100.00
11	AT	Austria	76.74	15.30	55.81	7.71	88.37	98.10
12	AU	Australia	81.82	97.29	57.44	95.68	80.17	99.61
13	AW	Aruba	75.00	18.33	75.00	7.87	100.00	100.00
14	AZ	Azerbaijan	63.89	98.46	63.89	93.77	97.22	99.76
15	BA	Bosnia and H.	91.67	49.08	58.33	1.38	91.67	68.48
16	BB	Barbados	50.00	51.76	50.00	51.76	50.00	51.76
17	BD	Bangladesh	37.50	87.88	37.50	0.17	64.58	88.03
18	BE	Belgium	76.92	31.94	56.41	21.62	92.31	48.98
19	BF	Burkina Faso	50.63	90.66	32.91	22.85	83.54	95.78
20	BG	Bulgaria	66.67	94.91	58.33	9.56	83.33	96.60
21	BH	Bahrain	78.57	37.40	78.57	89.43	92.86	49.27
22	BI	Burundi	80.00	99.41	60.00	0.23	90.00	99.45
23	BJ	Benin	60.66	67.18	16.39	36.33	55.74	67.54
24	BM	Bermuda	100.00	100.00	100.00	100.00	100.00	100.00
25	BN	Brunei	63.64	78.55	63.64	29.35	90.91	99.46
26	BO	Bolivia	86.05	99.70	62.79	10.56	86.05	99.69
27	BR	Brazil	85.87	97.66	49.46	97.66	85.33	97.89
28	BS	Bahamas	60.00	29.45	100.00	100.00	80.00	37.29
29	BT	Bhutan	45.95	81.36	35.14	72.89	45.95	78.93
30	BW	Botswana	56.76	22.89	45.95	92.12	83.78	96.36
31	BY	Belarus	80.00	25.66	56.00	98.98	100.00	100.00
32	BZ	Belize	81.82	92.52	72.73	36.43	81.82	77.88
33	CA	Canada	60.74	96.23	50.83	91.19	89.67	98.91

	ISO	Name	Language Phoible (%)	Speaker Phoible (%)	Language Grambank (%)	Speaker Grambank (%)	Language ASJP (%)	Speaker ASJP (%)
34	CC	Cocos Islands	100.00	100.00	100.00	100.00	50.00	33.33
35	CD	Congo, DRC	18.92	57.16	31.35	70.29	73.51	97.08
36	CF	Central African R.	24.39	35.12	35.37	33.91	79.27	90.53
37	CG	Congo	24.38	46.72	31.40	37.16	78.10	82.19
38	CH	Switzerland	73.33	94.22	51.11	36.61	86.67	99.05
39	CI	Côte D'Ivoire	53.40	78.86	23.30	26.85	75.73	93.55
40	CK	Cook Islands	16.67	9.45	83.33	97.84	100.00	100.00
41	CL	Chile	76.19	99.83	61.90	2.59	85.71	99.84
42	CM	Cameroon	54.33	79.98	31.14	42.02	75.43	93.37
43	CN	China	25.24	86.95	38.49	86.11	56.78	98.12
44	CO	Colombia	84.44	99.63	58.89	0.80	86.67	99.66
45	CR	Costa Rica	57.14	99.34	85.71	3.37	92.86	99.71
46	CU	Cuba	75.00	99.16	83.33	1.42	91.67	99.69
47	CV	Cabo Verde	60.00	89.56	20.00	1.90	60.00	23.99
48	CW	Curacao	60.00	25.05	50.00	16.37	80.00	96.77
49	CX	Christmas Island	66.67	69.23	100.00	100.00	66.67	53.85
50	CY	Cyprus	78.26	99.16	65.22	94.57	95.65	99.91
51	CZ	Czechia	69.70	99.64	51.52	97.72	87.88	99.76
52	DE	Germany	75.53	88.98	50.00	12.93	93.62	99.07
53	DJ	Djibouti	80.00	84.03	60.00	83.78	80.00	84.03
54	DK	Denmark	79.10	96.55	61.19	93.97	97.01	99.77
55	DM	Dominica	60.00	8.14	80.00	18.57	100.00	100.00
56	DO	Dominican Republic	66.67	93.70	66.67	6.50	77.78	98.97
57	DZ	Algeria	48.15	15.47	37.04	15.19	81.48	98.63
58	EC	Ecuador	74.07	98.76	55.56	6.67	92.59	99.70
59	EE	Estonia	73.68	95.20	57.89	91.68	89.47	96.15
60	EG	Egypt	68.57	68.57	54.29	67.60	85.71	73.75
61	EH	Western Sahara	50.00	1.64	50.00	98.36	50.00	1.64
62	ER	Eritrea	68.75	93.10	56.25	22.43	87.50	98.91
63	ES	Spain	73.61	98.93	52.78	18.76	86.11	99.20
64	ET	Ethiopia	61.39	78.87	57.43	61.88	94.06	98.60
65	FI	Finland	83.33	98.34	64.81	97.97	90.74	99.20
66	FJ	Fiji	47.83	63.45	60.87	60.34	78.26	75.63
67	FK	Falkland Islands	100.00	100.00	100.00	100.00	100.00	100.00
68	FM	Micronesia	33.33	33.65	87.50	97.52	87.50	97.52
69	FO	Faroe Islands	100.00	100.00	100.00	100.00	100.00	100.00
70	FR	France	69.89	95.57	51.61	90.26	90.32	98.81
71	GA	Gabon	19.61	58.97	31.37	13.76	88.24	93.42
72	GB	United Kingdom	78.12	98.64	55.21	94.98	91.67	99.50
73	GD	Grenada	50.00	7.26	50.00	7.26	75.00	9.63
74	GE	Georgia	72.22	94.55	61.11	85.92	88.89	99.46

	ISO	Name	Language Phoible (%)	Speaker Phoible (%)	Language Grambank (%)	Speaker Grambank (%)	Language ASJP (%)	Speaker ASJP (%)
75	GF	French Guiana	55.00	21.78	55.00	30.42	80.00	93.35
76	GH	Ghana	67.33	76.91	31.68	52.24	83.17	95.51
77	GI	Gibraltar	100.00	100.00	50.00	93.29	100.00	100.00
78	GL	Greenland	100.00	100.00	100.00	100.00	100.00	100.00
79	GM	Gambia	57.14	79.67	42.86	70.45	89.29	97.26
80	GN	Guinea	57.78	89.21	33.33	55.20	80.00	95.17
81	GP	Guadeloupe	40.00	2.96	80.00	5.71	100.00	100.00
82	GQ	Equatorial Guinea	55.56	89.10	50.00	18.01	83.33	98.37
83	GR	Greece	64.58	92.51	47.92	91.71	83.33	95.64
84	GT	Guatemala	29.63	92.20	51.85	19.23	92.59	99.67
85	GU	Guam	71.43	88.14	100.00	100.00	100.00	100.00
86	GW	Guinea-Bissau	51.61	50.12	41.94	36.62	87.10	83.10
87	GY	Guyana	72.22	15.33	61.11	14.24	77.78	15.43
88	HK	China, Hong Kong	85.00	98.61	85.00	98.87	95.00	99.80
89	HN	Honduras	50.00	98.16	71.43	2.01	92.86	99.62
90	HR	Croatia	65.22	96.13	43.48	1.58	69.57	93.94
91	HT	Haiti	60.00	0.94	80.00	99.33	100.00	100.00
92	HU	Hungary	80.77	99.58	53.85	95.57	88.46	99.81
93	ID	Indonesia	9.97	76.24	22.19	85.54	68.82	91.57
94	IE	Ireland	88.24	98.73	67.65	97.34	94.12	99.26
95	IL	Israel	53.66	75.26	46.34	91.76	73.17	96.28
96	IM	Isle of Man	50.00	98.03	50.00	98.03	100.00	100.00
97	IN	India	31.88	93.97	33.09	71.20	46.86	93.83
98	IO	British IOT	100.00	100.00	100.00	100.00	100.00	100.00
99	IQ	Iraq	51.72	33.13	44.83	3.77	89.66	99.43
100	IR	Iran	36.25	78.80	25.00	57.12	67.50	93.02
101	IS	Iceland	90.91	99.75	63.64	98.69	81.82	99.25
102	IT	Italy	67.78	67.86	45.56	76.28	90.00	93.65
103	JE	Jersey	0	0	0	0	0	0
104	JM	Jamaica	66.67	98.98	66.67	99.43	77.78	99.73
105	JO	Jordan	64.71	3.40	52.94	62.25	94.12	99.82
106	JP	Japan	61.76	98.89	64.71	99.68	85.29	99.89
107	KE	Kenya	50.57	76.53	40.23	83.17	81.61	97.22
108	KG	Kyrgyzstan	70.59	96.99	52.94	23.96	91.18	99.20
109	KH	Cambodia	47.06	97.27	64.71	99.41	73.53	99.62
110	KI	Kiribati	33.33	2.25	100.00	100.00	100.00	100.00
111	KM	Comoros	58.33	96.40	50.00	2.41	91.67	95.03
112	KN	Saint Kitts and Nevis	100.00	100.00	66.67	4.66	66.67	4.66
113	KP	Korea, North	100.00	100.00	100.00	100.00	100.00	100.00
114	KR	Korea, South	78.95	99.75	78.95	99.76	89.47	99.79
115	KW	Kuwait	68.18	57.90	63.64	91.48	90.91	99.17

	ISO	Name	Language Phoible (%)	Speaker Phoible (%)	Language Grambank (%)	Speaker Grambank (%)	Language ASJP (%)	Speaker ASJP (%)
116	KY	Cayman Islands	75.00	97.37	75.00	90.36	100.00	100.00
117	KZ	Kazakhstan	75.00	38.37	61.67	35.18	93.33	99.99
118	LA	Laos	35.05	79.52	38.14	74.81	54.64	86.73
119	LB	Lebanon	63.16	8.84	63.16	89.84	100.00	100.00
120	LC	Saint Lucia	75.00	5.35	75.00	5.35	100.00	100.00
121	LI	Liechtenstein	71.43	89.70	42.86	7.54	85.71	95.73
122	LK	Sri Lanka	73.68	99.41	52.63	24.19	89.47	99.66
123	LR	Liberia	48.65	64.50	27.03	21.25	56.76	73.29
124	LS	Lesotho	87.50	97.23	62.50	87.22	100.00	100.00
125	LT	Lithuania	66.67	83.14	46.67	83.83	86.67	84.13
126	LU	Luxembourg	88.24	95.30	64.71	31.70	88.24	96.54
127	LV	Latvia	78.95	96.73	57.89	31.52	89.47	89.81
128	LY	Libya	69.70	83.87	63.64	76.55	87.88	96.65
129	MA	Morocco	36.84	91.39	42.11	6.99	78.95	97.68
130	MC	Monaco	91.67	98.85	58.33	73.85	100.00	100.00
131	MD	Moldova	72.73	98.98	63.64	14.44	90.91	96.41
132	ME	Montenegro	68.75	64.32	37.50	4.32	75.00	13.64
133	MG	Madagascar	42.31	40.80	34.62	50.55	88.46	99.00
134	MH	Marshall Islands	75.00	7.98	100.00	100.00	100.00	100.00
135	MK	North Macedonia	64.71	95.05	47.06	93.49	88.24	97.91
136	ML	Mali	39.44	78.02	36.62	63.48	83.10	86.76
137	MM	Myanmar	30.43	85.62	42.03	76.57	52.17	91.35
138	MN	Mongolia	70.59	93.59	70.59	93.27	94.12	99.52
139	MO	China, Macau	77.78	99.47	77.78	99.47	77.78	99.47
140	MP	Northern Mariana I.	70.00	95.83	90.00	99.98	90.00	99.98
141	MQ	Martinique	50.00	4.49	83.33	5.45	100.00	100.00
142	MR	Mauritania	83.33	17.39	75.00	99.77	91.67	17.40
143	MS	Montserrat	100.00	100.00	50.00	2.63	50.00	2.63
144	MT	Malta	87.50	99.96	75.00	99.81	87.50	99.96
145	MU	Mauritius	88.24	99.85	70.59	20.26	94.12	99.92
146	MV	Maldives	100.00	100.00	50.00	1.88	100.00	100.00
147	MW	Malawi	47.83	15.90	39.13	15.89	73.91	85.63
148	MX	Mexico	20.53	96.50	31.13	3.26	60.26	98.78
149	MY	Malaysia	29.19	35.56	40.99	85.36	63.98	81.03
150	MZ	Mozambique	35.19	9.69	37.04	44.12	64.81	82.15
151	NA	Namibia	80.00	98.84	40.00	79.77	86.67	97.97
152	NC	New Caledonia	14.29	61.18	52.38	90.50	47.62	92.17
153	NE	Niger	54.84	95.78	51.61	86.89	77.42	98.86
154	NF	Norfolk Island	50.00	82.35	50.00	82.35	50.00	82.35
155	NG	Nigeria	23.75	88.46	16.67	57.85	66.28	95.74
156	NI	Nicaragua	50.00	97.76	75.00	2.49	83.33	99.64

	ISO	Name	Language Phoible (%)	Speaker Phoible (%)	Language Grambank (%)	Speaker Grambank (%)	Language ASJP (%)	Speaker ASJP (%)
157	NL	Netherlands	71.59	85.65	53.41	82.12	90.91	97.76
158	NO	Norway	87.04	98.13	61.11	10.73	92.59	99.27
159	NP	Nepal	33.03	85.87	44.04	78.89	60.55	81.48
160	NR	Nauru	28.57	19.35	100.00	100.00	100.00	100.00
161	NU	Niue	0	0	0	0	0	0
162	NZ	New Zealand	74.67	95.33	70.67	97.27	93.33	99.67
163	OM	Oman	66.67	36.45	45.45	29.56	78.79	55.13
164	PA	Panama	72.73	94.37	90.91	22.98	95.45	99.78
165	PE	Peru	68.75	94.67	44.79	12.21	78.12	98.06
166	PF	French Polynesia	20.00	71.17	60.00	95.63	70.00	98.76
167	PG	Papua New Guinea	9.35	14.06	30.77	65.68	81.54	93.15
168	PH	Philippines	13.37	84.31	43.32	92.20	60.43	96.62
169	PK	Pakistan	35.14	64.07	13.51	34.80	55.41	83.19
170	PL	Poland	64.52	97.74	45.16	97.67	83.87	98.17
171	PM	Saint Pierre and M.	100.00	100.00	100.00	100.00	100.00	100.00
172	PN	Pitcairn Islands	0	0	0	0	0	0
173	PR	Puerto Rico	85.71	99.68	57.14	2.47	85.71	99.68
174	PS	Palestine	55.56	9.59	66.67	99.93	100.00	100.00
175	PT	Portugal	80.65	99.61	45.16	97.73	80.65	99.23
176	PW	Palau	66.67	97.79	83.33	99.88	83.33	99.88
177	PY	Paraguay	75.00	99.54	58.33	95.25	83.33	99.40
178	QA	Qatar	78.95	36.63	57.89	84.19	100.00	100.00
179	RE	Reunion	92.86	99.93	57.14	39.26	92.86	99.92
180	RO	Romania	74.19	91.87	54.84	5.77	87.10	92.29
181	RS	Serbia	68.75	25.41	56.25	74.38	84.38	75.97
182	RU	Russian Federation	72.54	98.05	57.75	96.64	90.14	99.61
183	RW	Rwanda	70.00	98.81	50.00	0.78	90.00	98.87
184	SA	Saudi Arabia	70.73	76.48	60.98	53.92	87.80	97.86
185	SB	Solomon Islands	13.16	13.46	68.42	85.40	90.79	98.33
186	SC	Seychelles	60.00	5.38	60.00	5.38	80.00	99.46
187	SD	Sudan	59.57	97.92	42.55	13.53	87.23	99.65
188	SE	Sweden	80.56	96.57	59.72	94.79	90.28	99.51
189	SG	Singapore	76.32	82.60	63.16	93.10	89.47	99.43
190	SH	Saint Helena	100.00	100.00	100.00	100.00	100.00	100.00
191	SI	Slovenia	66.67	93.88	46.67	88.55	80.00	96.55
192	SJ	Svalbard	100.00	100.00	50.00	41.38	100.00	100.00
193	SK	Slovakia	85.19	97.62	51.85	10.18	88.89	98.47
194	SL	Sierra Leone	80.77	93.55	42.31	65.81	92.31	97.12
195	SM	San Marino	100.00	100.00	100.00	100.00	100.00	100.00
196	SN	Senegal	53.06	94.42	32.65	72.74	91.84	97.97
197	SO	Somalia	47.37	97.37	36.84	76.58	73.68	98.70

	ISO	Name	Language Phoible (%)	Speaker Phoible (%)	Language Grambank (%)	Speaker Grambank (%)	Language ASJP (%)	Speaker ASJP (%)
198	SR	Suriname	55.00	9.35	60.00	27.38	75.00	47.35
199	SS	South Sudan	55.41	82.83	31.08	24.83	79.73	62.16
200	ST	Sao Tome and P.	0	0	0	0	0	0
201	SV	El Salvador	62.50	99.55	62.50	0.22	87.50	99.59
202	SX	Sint Maarten	62.50	74.64	37.50	29.38	87.50	64.45
203	SY	Syria	72.41	22.59	55.17	82.98	100.00	100.00
204	SZ	Eswatini	55.56	11.65	44.44	3.72	100.00	100.00
205	TC	Turks and Caicos I.	33.33	3.63	66.67	25.17	66.67	25.17
206	TD	Chad	27.82	44.64	30.08	57.31	65.41	78.78
207	TG	Togo	66.67	75.29	21.05	35.32	78.95	78.86
208	TH	Thailand	55.43	96.33	54.35	51.03	80.43	69.82
209	TJ	Tajikistan	61.29	98.93	41.94	16.80	70.97	99.51
210	TK	Tokelau	0	0	0	0	0	0
211	TL	Timor-Leste	17.39	6.70	43.48	70.11	56.52	92.15
212	TM	Turkmenistan	85.11	93.07	59.57	90.67	95.74	95.37
213	TN	Tunisia	37.50	4.67	37.50	4.67	87.50	99.82
214	TO	Tonga	66.67	99.33	100.00	100.00	100.00	100.00
215	TR	Turkey	66.67	96.04	46.38	83.56	92.75	99.14
216	TT	Trinidad and Tobago	54.55	95.66	45.45	95.36	72.73	96.18
217	TV	Tuvalu	0	0	0	0	0	0
218	TW	Taiwan	57.69	97.96	96.15	99.76	92.31	99.75
219	TZ	Tanzania	35.17	58.26	41.38	64.74	82.07	96.18
220	UA	Ukraine	75.00	97.24	56.94	96.53	91.67	99.56
221	UG	Uganda	58.46	61.93	46.15	67.68	83.08	97.84
222	US	United States	58.21	97.66	53.98	83.50	89.30	99.44
223	UY	Uruguay	86.96	99.25	60.87	6.69	91.30	99.64
224	UZ	Uzbekistan	76.79	96.24	60.71	86.80	92.86	99.98
225	VA	Vatican	100.00	100.00	100.00	100.00	100.00	100.00
226	VC	St V, and G.	42.86	9.54	57.14	10.14	85.71	99.50
227	VE	Venezuela	83.33	99.21	64.81	3.40	90.74	99.69
228	VG	British Virgin Islands	66.67	10.83	66.67	10.83	66.67	10.83
229	VI	Virgin Islands (U.S.)	80.00	51.75	60.00	39.43	80.00	51.75
230	VN	Vietnam	28.70	89.09	39.13	89.20	57.39	96.37
231	VU	Vanuatu	9.57	18.98	61.74	86.20	98.26	100.00
232	WF	Wallis and Futuna	33.33	1.28	100.00	100.00	100.00	100.00
233	WS	Samoa	66.67	13.04	100.00	100.00	100.00	100.00
234	XK	Kosovo	75.00	97.78	37.50	94.09	75.00	96.23
235	YE	Yemen	55.56	6.04	44.44	48.04	85.19	60.42
236	YT	Mayotte	57.14	6.83	57.14	6.83	100.00	100.00
237	ZA	South Africa	60.00	70.73	55.56	43.33	91.11	99.14
238	ZM	Zambia	43.40	50.53	52.83	63.40	84.91	98.61

	ISO	Name	Language Phoible (%)	Speaker Phoible (%)	Language Grambank (%)	Speaker Grambank (%)	Language ASJP (%)	Speaker ASJP (%)
239	ZW	Zimbabwe	38.89	78.84	44.44	79.93	86.11	97.79

Table 7.1: A table showing the countries with linguistic data assessed in the thesis together with the respective language and speaker coverage in the databases PHOIBLE, Grambank and ASJP

	Country	Mean Phonemic Similarity	Mean Morphosyntactic Similarity	Mean Lexical Similarity	Mean Phylogenetic Similarity
1	Andorra	0.22	0.77	0.21	0.22
2	United Arab Emirates	0.24	0.65	0.16	0.11
3	Afghanistan	0.28	0.74	0.19	0.22
4	Antigua and Barbuda	0.30	-	-	0.73
5	Anguilla	0.22	0.65	0.14	0.31
6	Albania	0.30	0.76	0.16	0.14
7	Armenia	0.30	0.72	0.13	0.09
8	Angola	0.28	0.70	0.23	0.32
9	Argentina	0.25	0.66	0.13	0.07
10	American Samoa	0.27	0.66	0.20	0.12
11	Austria	0.25	0.68	0.16	0.11
12	Australia	0.29	0.64	0.13	0.03
13	Aruba	0.25	0.70	0.17	0.14
14	Azerbaijan	0.25	0.73	0.12	0.09
15	Bosnia and Herzegovina	0.28	0.74	0.18	0.16
16	Barbados	0.19	0.65	0.14	0.14
17	Bangladesh	0.27	0.70	0.17	0.10
18	Belgium	0.24	0.66	0.12	0.06
19	Burkina Faso	0.47	0.70	0.13	0.17
20	Bulgaria	0.25	0.69	0.16	0.11
21	Bahrain	0.32	0.67	0.14	0.05
22	Burundi	0.30	0.68	0.14	0.07
23	Benin	0.49	0.74	0.15	0.20
24	Bermuda	0.26	0.64	0.08	0.03
25	Brunei	0.24	0.70	0.19	0.10
26	Bolivia	0.32	0.64	0.13	0.03
27	Brazil	0.35	0.65	0.12	0.03
28	Bahamas	0.16	0.74	0.10	0.11
29	Bhutan	0.33	0.73	0.18	0.21
30	Botswana	0.22	0.74	0.18	0.16
31	Belarus	0.24	0.73	0.14	0.09
32	Belize	0.19	0.67	0.12	0.07

	Country	Mean Phonemic Similarity	Mean Morphosyntactic Similarity	Mean Lexical Similarity	Mean Phylogenetic Similarity
33	Canada	0.24	0.64	0.10	0.03
34	Cocos Islands	0.21	0.73	-	0.00
35	Congo, Democratic Republic Of	0.40	0.76	0.19	0.28
36	Central African Republic	0.43	0.73	0.16	0.18
37	Congo	0.40	0.73	0.18	0.25
38	Switzerland	0.25	0.69	0.15	0.10
39	Côte D'Ivoire	0.45	0.67	0.13	0.09
40	Cook Islands	-	0.73	0.50	0.44
41	Chile	0.22	0.65	0.11	0.04
42	Cameroon	0.39	0.72	0.12	0.20
43	China	0.24	0.76	0.12	0.07
44	Colombia	0.37	0.66	0.13	0.03
45	Costa Rica	0.22	0.65	0.13	0.09
46	Cuba	0.22	0.68	0.16	0.12
47	Cabo Verde	0.26	-	0.11	0.11
48	Curacao	0.22	0.70	0.20	0.20
49	Christmas Island	0.21	0.73	0.10	0.24
50	Cyprus	0.23	0.66	0.13	0.09
51	Czechia	0.24	0.70	0.18	0.13
52	Germany	0.24	0.67	0.13	0.08
53	Djibouti	0.23	0.63	0.14	0.14
54	Denmark	0.23	0.66	0.13	0.08
55	Dominica	0.22	0.65	0.18	0.14
56	Dominican Republic	0.24	0.70	0.17	0.11
57	Algeria	0.21	0.66	0.19	0.17
58	Ecuador	0.31	0.71	0.21	0.07
59	Estonia	0.26	0.75	0.16	0.10
60	Egypt	0.25	0.67	0.14	0.10
61	Western Sahara	-	-	-	0.88
62	Eritrea	0.29	0.66	0.17	0.18
63	Spain	0.24	0.69	0.15	0.11
64	Ethiopia	0.31	0.68	0.13	0.06
65	Finland	0.22	0.67	0.12	0.05
66	Fiji	0.23	0.68	0.21	0.19
67	Falkland Islands	-	-	-	-
68	Micronesia	0.24	0.74	0.21	0.32
69	Faroe Islands	0.18	0.77	0.38	0.62
70	France	0.23	0.65	0.11	0.05
71	Gabon	0.33	0.72	0.21	0.30
72	United Kingdom	0.23	0.66	0.12	0.07
73	Grenada	0.19	0.65	0.15	0.18

	Country	Mean Phonemic Similarity	Mean Morphosyntactic Similarity	Mean Lexical Similarity	Mean Phylogenetic Similarity
74	Georgia	0.26	0.72	0.11	0.06
75	French Guiana	0.26	0.65	0.14	0.08
76	Ghana	0.45	0.72	0.14	0.17
77	Gibraltar	0.16	0.65	0.11	0.05
78	Greenland	0.13	0.64	0.17	0.10
79	Gambia	0.40	0.67	0.12	0.10
80	Guinea	0.42	0.68	0.15	0.13
81	Guadeloupe	0.30	0.65	0.25	0.27
82	Equatorial Guinea	0.27	0.70	0.15	0.24
83	Greece	0.24	0.68	0.14	0.10
84	Guatemala	0.38	0.75	0.34	0.28
85	Guam	0.25	0.69	0.16	0.16
86	Guinea-Bissau	0.41	0.67	0.12	0.15
87	Guyana	0.26	0.66	0.13	0.07
88	China, Hong Kong	0.24	0.68	0.13	0.05
89	Honduras	0.21	0.65	0.11	0.02
90	Croatia	0.29	0.75	0.28	0.23
91	Haiti	0.22	0.70	0.18	0.22
92	Hungary	0.26	0.68	0.20	0.15
93	Indonesia	0.35	0.69	0.16	0.09
94	Ireland	0.24	0.70	0.15	0.11
95	Israel	0.25	0.67	0.14	0.09
96	Isle of Man	-	-	0.11	0.11
97	India	0.32	0.73	0.15	0.09
98	British Indian Ocean Territory	0.36	0.60	0.05	0.00
99	Iraq	0.26	0.69	0.17	0.12
100	Iran	0.25	0.69	0.19	0.22
101	Iceland	0.25	0.68	0.17	0.12
102	Italy	0.27	0.67	0.15	0.11
103	Jersey	-	-	-	-
104	Jamaica	0.19	0.68	0.15	0.07
105	Jordan	0.27	0.65	0.15	0.09
106	Japan	0.21	0.70	0.15	0.07
107	Kenya	0.35	0.69	0.17	0.16
108	Kyrgyzstan	0.25	0.75	0.16	0.07
109	Cambodia	0.26	0.76	0.16	0.09
110	Kiribati		0.67	0.17	0.12
111	Comoros	0.34	0.65	0.23	0.24
112	Saint Kitts and Nevis	0.22	0.65	0.14	0.31
113	Korea, North	0.18	0.63	0.12	0.02
114	Korea, South	0.25	0.67	0.12	0.02

	Country	Mean Phonemic Similarity	Mean Morphosyntactic Similarity	Mean Lexical Similarity	Mean Phylogenetic Similarity
115	Kuwait	0.25	0.66	0.15	0.11
116	Cayman Islands	0.28	0.65	0.12	0.12
117	Kazakhstan	0.22	0.74	0.13	0.06
118	Laos	0.33	0.83	0.19	0.10
119	Lebanon	0.28	0.70	0.16	0.12
120	Saint Lucia	0.22	0.68	0.20	0.21
121	Liechtenstein	0.25	0.64	0.19	0.24
122	Sri Lanka	0.31	0.72	0.20	0.14
123	Liberia	0.43	0.67	0.15	0.13
124	Lesotho	0.19	0.65	0.18	0.18
125	Lithuania	0.26	0.74	0.18	0.17
126	Luxembourg	0.22	0.73	0.20	0.16
127	Latvia	0.27	0.73	0.14	0.11
128	Libya	0.25	0.67	0.14	0.10
129	Morocco	0.21	0.69	0.19	0.18
130	Monaco	0.22	0.78	0.20	0.21
131	Moldova	0.28	0.74	0.16	0.11
132	Montenegro	0.30	0.74	0.25	0.25
133	Madagascar	0.26	0.66	0.28	0.17
134	Marshall Islands	0.35	0.62	0.09	0.03
135	North Macedonia	0.31	0.73	0.19	0.17
136	Mali	0.45	0.70	0.16	0.07
137	Myanmar	0.26	0.72	0.15	0.09
138	Mongolia	0.23	0.75	0.14	0.08
139	China, Macau	0.22	0.65	0.12	0.05
140	Northern Mariana Islands	0.24	0.67	0.13	0.12
141	Martinique	0.21	0.65	0.20	0.18
142	Mauritania	0.34	0.64	0.10	0.05
143	Montserrat	0.30	-	-	0.73
144	Malta	0.29	0.67	0.13	0.06
145	Mauritius	0.25	0.67	0.19	0.10
146	Maldives	0.31	0.76	0.26	0.25
147	Malawi	0.25	0.71	0.21	0.23
148	Mexico	0.26	0.70	0.13	0.10
149	Malaysia	0.29	0.72	0.21	0.13
150	Mozambique	0.23	0.71	0.21	0.34
151	Namibia	0.26	0.70	0.26	0.25
152	New Caledonia	0.23	0.73	0.15	0.37
153	Niger	0.36	0.68	0.14	0.08
154	Norfolk Island	-	-	-	1.00
155	Nigeria	0.41	0.70	0.12	0.14

	Country	Mean Phonemic Similarity	Mean Morphosyntactic Similarity	Mean Lexical Similarity	Mean Phylogenetic Similarity
156	Nicaragua	0.24	0.65	0.13	0.04
157	Netherlands	0.24	0.66	0.14	0.08
158	Norway	0.23	0.69	0.13	0.08
159	Nepal	0.30	0.75	0.21	0.16
160	Nauru	0.21	0.69	0.16	0.28
161	Niue	-	-	-	-
162	New Zealand	0.23	0.64	0.14	0.06
163	Oman	0.26	0.67	0.18	0.15
164	Panama	0.23	0.67	0.13	0.07
165	Peru	0.37	0.68	0.19	0.06
166	French Polynesia	0.24	0.74	0.39	0.52
167	Papua New Guinea	0.31	0.67	0.12	0.05
168	Philippines	0.36	0.75	0.37	0.24
169	Pakistan	0.29	0.76	0.22	0.28
170	Poland	0.26	0.71	0.20	0.20
171	Saint Pierre and Miquelon	0.25	0.75	0.14	0.08
172	Pitcairn Islands	-	-	-	-
173	Puerto Rico	0.22	0.63	0.15	0.07
174	Palestine	0.25	0.78	0.16	0.15
175	Portugal	0.23	0.70	0.17	0.17
176	Palau	0.32	0.67	0.13	0.13
177	Paraguay	0.25	0.65	0.12	0.06
178	Qatar	0.28	0.67	0.17	0.09
179	Reunion	0.24	0.63	0.18	0.13
180	Romania	0.25	0.71	0.18	0.13
181	Serbia	0.27	0.72	0.20	0.16
182	Russian Federation	0.20	0.73	0.11	0.04
183	Rwanda	0.37	0.72	0.17	0.10
184	Saudi Arabia	0.24	0.65	0.15	0.09
185	Solomon Islands	0.23	0.75	0.26	0.30
186	Seychelles	0.18	0.68	0.15	0.10
187	Sudan	0.33	0.65	0.12	0.02
188	Sweden	0.23	0.69	0.13	0.07
189	Singapore	0.27	0.67	0.15	0.06
190	Saint Helena	-	-	-	-
191	Slovenia	0.30	0.75	0.26	0.20
192	Svalbard and Jan Mayen	0.15		0.16	0.22
193	Slovakia	0.24	0.68	0.20	0.16
194	Sierra Leone	0.41	0.66	0.14	0.10
195	San Marino	-	-	-	-
196	Senegal	0.38	0.67	0.13	0.15

	Country	Mean Phonemic Similarity	Mean Morphosyntactic Similarity	Mean Lexical Similarity	Mean Phylogenetic Similarity
197	Somalia	0.26	0.67	0.18	0.19
198	Suriname	0.25	0.66	0.13	0.08
199	South Sudan	0.36	0.72	0.13	0.09
200	Sao Tome and Principe	-	-	-	-
201	El Salvador	0.26	0.64	0.11	0.02
202	Sint Maarten	0.27	0.78	0.23	0.28
203	Syria	0.24	0.67	0.13	0.09
204	Eswatini	0.21	0.66	0.20	0.22
205	Turks and Caicos Islands	-	0.74	0.12	0.26
206	Chad	0.38	0.71	0.13	0.09
207	Togo	0.48	0.71	0.15	0.20
208	Thailand	0.31	0.77	0.14	0.06
209	Tajikistan	0.24	0.71	0.15	0.13
210	Tokelau	-	-	-	-
211	Timor-Leste	0.23	0.73	0.25	0.24
212	Turkmenistan	0.23	0.73	0.13	0.07
213	Tunisia	0.14	0.65	0.23	0.24
214	Tonga	0.13	0.67	0.31	0.23
215	Turkey	0.24	0.67	0.13	0.06
216	Trinidad and Tobago	0.23	0.67	0.16	0.14
217	Tuvalu	-	-	-	-
218	Taiwan	0.24	0.70	0.17	0.14
219	Tanzania	0.36	0.76	0.28	0.40
220	Ukraine	0.22	0.71	0.12	0.06
221	Uganda	0.34	0.70	0.15	0.11
222	United States	0.24	0.65	0.10	0.02
223	Uruguay	0.24	0.72	0.19	0.14
224	Uzbekistan	0.22	0.73	0.13	0.06
225	Vatican	-	-	-	-
226	St Vincent and Grenadines	0.18	0.65	0.13	0.06
227	Venezuela	0.29	0.63	0.12	0.04
228	British Virgin Islands	0.19	0.65	0.14	0.31
229	Virgin Islands (U.S.)	0.21	0.68	0.16	0.19
230	Vietnam	0.28	0.80	0.16	0.07
231	Vanuatu	0.25	0.75	0.26	0.43
232	Wallis and Futuna	-	0.66	0.35	0.30
233	Samoa	0.25	0.61	0.10	0.00
234	Kosovo	0.34	0.72	0.19	0.20
235	Yemen	0.23	0.67	0.20	0.23
236	Mayotte	0.32	0.66	0.21	0.19
237	South Africa	0.21	0.66	0.14	0.09

	Country	Mean Phonemic Similarity	Mean Morphosyntactic Similarity	Mean Lexical Similarity	Mean Phylogenetic Similarity
238	Zambia	0.28	0.75	0.28	0.28
239	Zimbabwe	0.20	0.71	0.24	0.31

Table 7.2: A table with values for mean phonemic, morphosyntactic, lexical and phylogenetic similarity per country.

	Country	Naive Diversity, q = 0	Naive Diversity, q = 1	Naive Diversity, q = 2	Non-naive Diversity, q = 0	Non-naive Diversity, q = 1	Non-naive Diversity, q = 2
1	Andorra	9	3.97	3.17	2.00	1.89	1.84
2	United Arab Emirates	30	9.19	4.93	3.18	2.89	2.68
3	Afghanistan	55	7.79	5.63	2.84	2.66	2.53
4	Antigua and Barbuda	2	1.14	1.06	1.00	1.00	1.00
5	Anguilla	3	1.31	1.14	1.42	1.26	1.18
6	Albania	12	1.72	1.42	1.20	1.14	1.12
7	Armenia	15	2.46	2.11	1.52	1.32	1.25
8	Angola	58	12.11	6.70	2.57	2.38	2.25
9	Argentina	51	1.88	1.27	1.60	1.28	1.18
10	American Samoa	7	1.79	1.35	1.77	1.41	1.27
11	Austria	43	2.85	1.49	1.94	1.47	1.31
12	Australia	242	3.45	1.54	2.68	1.75	1.43
13	Aruba	8	2.09	1.50	1.61	1.37	1.28
14	Azerbaijan	36	2.38	1.44	2.10	1.48	1.29
15	Bosnia and H.	12	3.02	2.61	1.20	1.12	1.09
16	Barbados	4	2.11	2.04	1.09	1.04	1.02
17	Bangladesh	48	1.77	1.31	1.03	1.02	1.01
18	Belgium	78	8.92	3.83	4.78	4.23	3.81
19	Burkina Faso	79	10.99	5.00	3.87	3.38	3.00
20	Bulgaria	24	1.82	1.35	1.64	1.34	1.22
21	Bahrain	14	5.76	3.40	3.91	3.70	3.53
22	Burundi	10	1.08	1.02	1.02	1.01	1.01
23	Benin	61	26.08	15.94	4.31	4.04	3.81
24	Bermuda	3	1.26	1.12	1.43	1.17	1.10
25	Brunei	22	6.87	3.57	2.77	2.31	2.04
26	Bolivia	43	2.71	1.85	2.34	1.92	1.68
27	Brazil	184	1.24	1.07	1.08	1.03	1.02
28	Bahamas	5	2.41	2.07	1.79	1.57	1.45
29	Bhutan	37	9.24	4.59	2.62	2.26	2.06
30	Botswana	37	3.74	1.84	1.78	1.50	1.39
31	Belarus	25	2.48	1.74	1.39	1.29	1.25
32	Belize	11	4.23	2.70	2.93	2.10	1.71

	Country	Naive Diversity, q = 0	Naive Diversity, q = 1	Naive Diversity, q = 2	Non-naive Diversity, q = 0	Non-naive Diversity, q = 1	Non-naive Diversity, q = 2
33	Canada	242	8.43	3.30	3.64	2.89	2.46
34	Cocos Islands	2	1.89	1.80	1.00	1.00	1.00
35	Congo, DRC	185	11.51	3.81	2.61	2.27	2.06
36	Central African Republic	82	36.01	23.17	4.89	4.66	4.48
37	Congo	242	47.59	20.87	4.22	3.90	3.66
38	Switzerland	45	5.74	2.90	2.83	2.39	2.14
39	Côte D'Ivoire	103	25.42	13.30	5.32	5.00	4.69
40	Cook Islands	6	2.09	1.50	1.57	1.35	1.26
41	Chile	21	1.22	1.07	1.34	1.10	1.05
42	Cameroon	289	53.59	20.96	6.64	6.36	6.10
43	China	317	4.38	2.01	2.05	1.72	1.55
44	Colombia	90	1.12	1.03	1.11	1.03	1.02
45	Costa Rica	14	1.25	1.08	1.30	1.11	1.06
46	Cuba	12	1.13	1.04	1.11	1.04	1.02
47	Cabo Verde	5	2.25	1.69	2.21	2.08	2.00
48	Curacao	10	3.13	1.95	1.98	1.66	1.50
49	Christmas Island	3	2.88	2.77	1.82	1.80	1.79
50	Cyprus	23	3.46	2.08	2.85	2.23	1.89
51	Czechia	33	1.33	1.10	1.14	1.06	1.04
52	Germany	94	4.94	2.07	2.41	1.80	1.55
53	Djibouti	10	3.51	2.78	1.99	1.75	1.67
54	Denmark	67	2.40	1.34	2.13	1.45	1.27
55	Dominica	5	1.95	1.48	1.54	1.31	1.22
56	Dominican Republic	9	1.42	1.17	1.30	1.16	1.11
57	Algeria	27	2.28	1.62	1.82	1.58	1.44
58	Ecuador	27	1.55	1.19	1.60	1.26	1.16
59	Estonia	19	2.93	2.01	1.92	1.72	1.60
60	Egypt	35	3.11	2.13	1.23	1.16	1.13
61	Western Sahara	2	1.09	1.03	1.00	1.00	1.00
62	Eritrea	16	5.23	3.38	2.61	2.29	2.08
63	Spain	72	3.92	2.02	1.93	1.59	1.45
64	Ethiopia	101	12.29	6.84	3.44	3.21	3.05
65	Finland	54	1.94	1.29	1.78	1.35	1.21
66	Fiji	23	7.80	4.95	3.01	2.60	2.31
67	Falkland Islands	1	1.00	1.00	1.00	1.00	1.00
68	Micronesia	24	8.28	4.65	2.92	2.61	2.40
69	Faroe Islands	2	1.14	1.06	1.06	1.04	1.04
70	France	93	2.97	1.46	2.24	1.56	1.34
71	Gabon	51	11.17	4.73	3.40	2.96	2.65
72	United Kingdom	96	2.68	1.42	2.11	1.52	1.32
73	Grenada	4	1.52	1.22	2.28	2.26	2.25

	Country	Naive Diversity, q = 0	Naive Diversity, q = 1	Naive Diversity, q = 2	Non-naive Diversity, q = 0	Non-naive Diversity, q = 1	Non-naive Diversity, q = 2
74	Georgia	36	2.64	1.66	2.35	1.71	1.47
75	French Guiana	20	6.99	3.85	2.81	2.32	2.03
76	Ghana	101	17.30	8.22	4.35	4.07	3.82
77	Gibraltar	4	1.47	1.20	1.72	1.30	1.18
78	Greenland	3	1.86	1.50	1.99	1.63	1.44
79	Gambia	28	6.70	4.06	3.51	3.08	2.76
80	Guinea	45	7.21	4.72	3.09	2.92	2.81
81	Guadeloupe	5	1.33	1.12	1.13	1.07	1.06
82	Equatorial Guinea	18	4.45	2.40	3.19	2.42	1.99
83	Greece	48	2.53	1.45	2.02	1.46	1.28
84	Guatemala	27	3.51	1.95	2.59	2.06	1.76
85	Guam	14	6.11	4.33	3.44	3.09	2.83
86	Guinea-Bissau	31	10.17	7.60	4.03	3.82	3.67
87	Guyana	18	3.51	2.55	3.80	3.02	2.49
88	China, Hong Kong	20	1.96	1.32	1.62	1.34	1.23
89	Honduras	14	1.21	1.06	1.24	1.08	1.04
90	Croatia	23	1.64	1.20	1.12	1.05	1.04
91	Haiti	5	1.07	1.02	1.06	1.03	1.02
92	Hungary	26	1.38	1.12	1.50	1.18	1.10
93	Indonesia	712	17.80	6.15	2.45	2.30	2.22
94	Ireland	34	2.03	1.31	1.84	1.38	1.24
95	Israel	41	4.71	2.58	2.70	2.16	1.88
96	Isle of Man	2	1.10	1.04	1.15	1.06	1.04
97	India	414	11.57	4.77	2.55	2.25	2.06
98	British IOT	2	1.28	1.14	1.61	1.23	1.13
99	Iraq	29	5.90	4.05	2.72	2.38	2.15
100	Iran	80	6.88	3.05	2.28	2.00	1.83
101	Iceland	11	1.43	1.15	1.39	1.17	1.11
102	Italy	90	6.05	2.80	1.93	1.67	1.55
103	Jersey	1	1.00	1.00	0.00	1.00	1.00
104	Jamaica	9	1.53	1.19	1.57	1.24	1.15
105	Jordan	17	3.39	2.61	1.69	1.50	1.45
106	Japan	34	1.24	1.07	1.21	1.09	1.05
107	Kenya	87	22.11	12.55	4.55	4.30	4.06
108	Kyrgyzstan	34	2.98	1.87	1.93	1.51	1.37
109	Cambodia	34	1.76	1.27	1.69	1.34	1.22
110	Kiribati	3	1.18	1.07	1.18	1.09	1.06
111	Comoros	12	2.88	2.39	1.32	1.25	1.23
112	Saint Kitts and Nevis	3	1.25	1.10	1.75	1.74	1.74
113	Korea, North	4	1.05	1.01	1.05	1.02	1.01
114	Korea, South	19	1.25	1.08	1.30	1.11	1.07

	Country	Naive Diversity, q = 0	Naive Diversity, q = 1	Naive Diversity, q = 2	Non-naive Diversity, q = 0	Non-naive Diversity, q = 1	Non-naive Diversity, q = 2
115	Kuwait	22	6.18	4.26	2.73	2.51	2.38
116	Cayman Islands	4	1.66	1.32	1.81	1.41	1.27
117	Kazakhstan	60	3.53	2.28	2.30	2.01	1.86
118	Laos	97	10.92	3.77	3.06	2.45	2.07
119	Lebanon	19	2.19	1.41	1.96	1.48	1.31
120	Saint Lucia	4	1.28	1.11	1.26	1.14	1.09
121	Liechtenstein	7	2.42	1.59	1.78	1.39	1.27
122	Sri Lanka	19	1.84	1.60	1.60	1.50	1.43
123	Liberia	37	20.08	14.02	4.30	4.04	3.85
124	Lesotho	8	1.75	1.35	1.47	1.31	1.24
125	Lithuania	15	2.54	1.81	1.56	1.36	1.26
126	Luxembourg	17	4.52	2.67	2.36	2.08	1.90
127	Latvia	19	3.67	2.54	1.90	1.76	1.68
128	Libya	33	5.05	2.51	1.85	1.60	1.48
129	Morocco	19	2.90	1.92	1.95	1.73	1.59
130	Monaco	12	5.62	4.62	2.34	2.22	2.16
131	Moldova	22	2.15	1.43	1.68	1.38	1.26
132	Montenegro	16	3.25	2.38	2.45	2.23	2.08
133	Madagascar	26	8.51	5.29	1.64	1.50	1.45
134	Marshall Islands	4	1.39	1.17	1.57	1.25	1.16
135	North Macedonia	17	3.06	2.16	2.34	2.02	1.83
136	Mali	71	15.53	7.47	3.96	3.39	2.97
137	Myanmar	138	7.29	2.66	2.68	2.15	1.84
138	Mongolia	17	2.35	1.54	2.13	1.51	1.33
139	China, Macau	9	2.04	1.42	1.86	1.46	1.31
140	Northern Mariana I.	10	5.35	4.28	3.54	3.22	2.98
141	Martinique	6	1.32	1.12	1.20	1.10	1.07
142	Mauritania	12	2.02	1.45	3.01	2.75	2.58
143	Montserrat	2	1.13	1.05	1.00	1.00	1.00
144	Malta	8	1.22	1.07	1.26	1.10	1.06
145	Mauritius	17	2.62	1.59	2.06	1.61	1.42
146	Maldives	6	1.19	1.06	1.10	1.06	1.04
147	Malawi	23	4.43	2.70	1.55	1.45	1.40
148	Mexico	302	1.69	1.15	1.58	1.20	1.11
149	Malaysia	161	11.83	5.06	2.94	2.49	2.17
150	Mozambique	54	19.24	12.27	2.51	2.45	2.42
151	Namibia	15	6.04	4.23	2.37	2.18	2.06
152	New Caledonia	42	7.83	3.24	3.47	2.71	2.24
153	Niger	31	4.47	2.51	2.73	2.28	2.01
154	Norfolk Island	2	1.59	1.41	1.00	1.00	1.00
155	Nigeria	522	21.86	8.67	4.92	4.55	4.25

	Country	Naive Diversity, q = 0	Naive Diversity, q = 1	Naive Diversity, q = 2	Non-naive Diversity, q = 0	Non-naive Diversity, q = 1	Non-naive Diversity, q = 2
156	Nicaragua	12	1.18	1.06	1.22	1.08	1.05
157	Netherlands	88	5.54	2.13	2.61	1.85	1.57
158	Norway	54	2.73	1.42	2.37	1.55	1.32
159	Nepal	109	6.87	3.11	2.10	1.77	1.59
160	Nauru	7	3.92	2.95	2.79	2.49	2.26
161	Niue	1	1.00	1.00	1.00	1.00	1.00
162	New Zealand	75	3.68	1.67	2.76	1.88	1.54
163	Oman	33	8.07	4.30	3.40	3.18	3.01
164	Panama	22	2.71	1.66	2.27	1.74	1.51
165	Peru	96	2.65	1.49	2.07	1.57	1.37
166	French Polynesia	10	2.49	1.87	1.92	1.76	1.64
167	Papua New Guinea	845	96.65	13.83	5.95	5.42	4.86
168	Philippines	187	10.52	4.50	1.60	1.53	1.50
169	Pakistan	74	7.34	5.25	2.40	2.31	2.25
170	Poland	31	1.20	1.06	1.06	1.02	1.01
171	Saint Pierre and M.	2	1.22	1.11	1.27	1.14	1.09
172	Pitcairn Islands	1	1.00	1.00	0.00	1.00	1.00
173	Puerto Rico	7	1.16	1.06	1.18	1.07	1.05
174	Palestine	9	1.48	1.23	1.24	1.18	1.14
175	Portugal	31	1.42	1.11	1.23	1.11	1.07
176	Palau	6	3.24	2.51	2.68	2.33	2.09
177	Paraguay	36	1.81	1.26	1.80	1.35	1.21
178	Qatar	19	7.19	4.71	2.99	2.59	2.34
179	Reunion	14	3.88	2.58	2.50	1.97	1.72
180	Romania	31	1.90	1.37	1.58	1.25	1.15
181	Serbia	32	2.20	1.71	1.20	1.08	1.05
182	Russian Federation	142	2.49	1.37	2.44	1.55	1.31
183	Rwanda	10	1.38	1.15	1.07	1.05	1.05
184	Saudi Arabia	41	5.91	3.71	2.05	1.79	1.67
185	Solomon Islands	76	42.31	29.93	3.65	3.48	3.34
186	Seychelles	5	1.34	1.13	1.34	1.14	1.09
187	Sudan	94	2.86	1.49	2.17	1.56	1.35
188	Sweden	72	2.28	1.30	2.09	1.41	1.23
189	Singapore	38	10.34	5.74	3.71	3.36	3.08
190	Saint Helena	1	1.00	1.00	1.00	1.00	1.00
191	Slovenia	15	2.01	1.35	1.25	1.14	1.10
192	Svalbard	2	1.97	1.94	1.71	1.70	1.69
193	Slovakia	27	1.87	1.34	1.66	1.35	1.24
194	Sierra Leone	26	8.32	5.47	4.04	3.73	3.48
195	San Marino	1	1.00	1.00	1.00	1.00	1.00
196	Senegal	49	7.88	4.71	3.68	3.32	3.08

	Country	Naive Diversity, q = 0	Naive Diversity, q = 1	Naive Diversity, q = 2	Non-naive Diversity, q = 0	Non-naive Diversity, q = 1	Non-naive Diversity, q = 2
197	Somalia	19	2.34	1.72	1.54	1.31	1.25
198	Suriname	20	7.34	5.70	2.50	1.98	1.74
199	South Sudan	74	18.83	9.75	4.77	4.35	4.01
200	Sao Tome and P.	5	2.75	2.20	1.59	1.27	1.19
201	El Salvador	8	1.05	1.01	1.03	1.01	1.01
202	Sint Maarten	8	5.05	4.16	2.40	2.33	2.28
203	Syria	29	2.92	1.80	2.26	1.75	1.53
204	Eswatini	9	1.82	1.35	1.44	1.21	1.15
205	Turks and Caicos I.	3	1.95	1.65	1.57	1.38	1.28
206	Chad	133	35.16	18.11	5.37	5.10	4.84
207	Togo	57	18.03	10.08	4.22	3.87	3.56
208	Thailand	92	5.21	3.31	1.98	1.56	1.41
209	Tajikistan	31	1.99	1.48	1.70	1.49	1.37
210	Tokelau	1	1.00	1.00	1.00	1.00	1.00
211	Timor-Leste	23	9.42	5.97	2.43	2.25	2.12
212	Turkmenistan	47	3.24	1.77	2.00	1.49	1.33
213	Tunisia	8	1.40	1.17	1.11	1.09	1.07
214	Tonga	3	1.08	1.03	1.09	1.03	1.02
215	Turkey	69	2.53	1.53	2.53	1.64	1.41
216	Trinidad and Tobago	11	1.38	1.13	1.16	1.06	1.04
217	Tuvalu	2	1.06	1.02	1.03	1.02	1.01
218	Taiwan	26	3.15	2.44	2.10	1.81	1.69
219	Tanzania	145	45.31	19.80	3.11	2.90	2.79
220	Ukraine	72	2.32	1.61	1.55	1.33	1.26
221	Uganda	65	20.63	12.81	3.72	3.37	3.11
222	United States	402	3.56	1.76	2.63	1.89	1.59
223	Uruguay	23	1.69	1.21	1.36	1.18	1.13
224	Uzbekistan	56	2.71	1.57	1.79	1.44	1.31
225	Vatican	1	1.00	1.00	1.00	1.00	1.00
226	St Vincent and G.	7	1.76	1.30	1.72	1.34	1.22
227	Venezuela	54	1.29	1.08	1.26	1.10	1.06
228	British Virgin Islands	3	1.51	1.25	1.71	1.65	1.60
229	Virgin Islands (U.S.)	5	2.85	2.55	1.84	1.66	1.54
230	Vietnam	115	2.82	1.46	1.89	1.44	1.28
231	Vanuatu	115	54.34	37.24	4.05	3.92	3.81
232	Wallis and Futuna	3	1.99	1.82	1.19	1.12	1.10
233	Samoa	3	1.52	1.30	1.63	1.37	1.26
234	Kosovo	8	1.48	1.17	1.41	1.12	1.07
235	Yemen	27	3.91	2.84	1.70	1.45	1.35
236	Mayotte	7	2.45	1.86	1.94	1.74	1.62
237	South Africa	45	10.11	7.76	3.29	3.04	2.86

	Country	Naive Diversity, q = 0	Naive Diversity, q = 1	Naive Diversity, q = 2	Non-naive Diversity, q = 0	Non-naive Diversity, q = 1	Non-naive Diversity, q = 2
238	Zambia	53	12.65	6.75	2.23	2.05	1.98
239	Zimbabwe	36	5.53	2.60	1.98	1.70	1.59

Table 7.3: A table with values for Naive and Non-naive Diversity for q = 0, 1 and 2 per country.

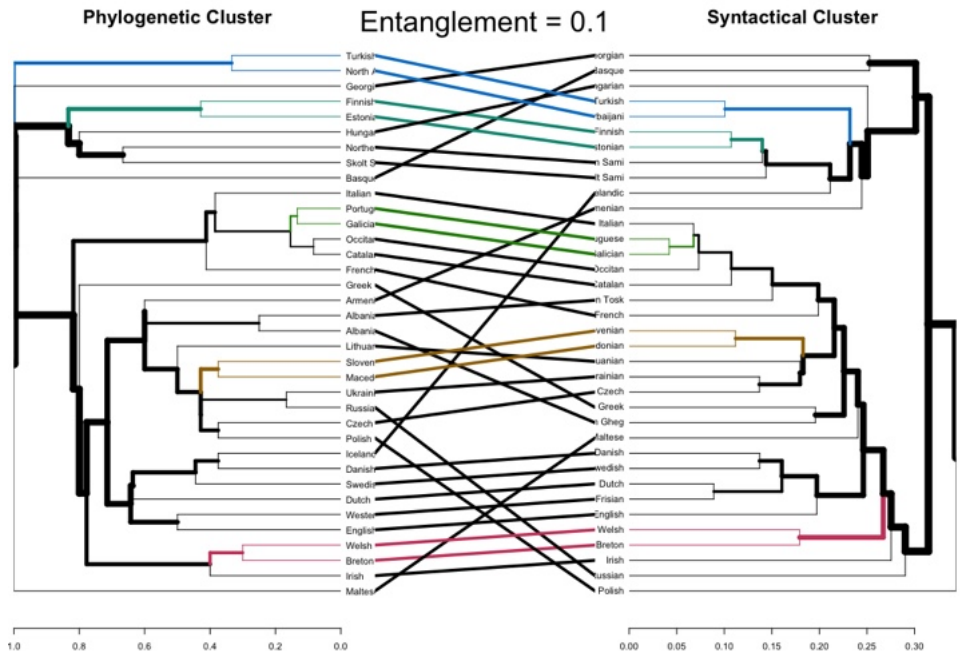

Figure 7.1: A tanglegram showcasing the structural alignment between the dendrograms resulting from hierarchical clustering based on phylogenetic and morphosyntactic similarities. Colored branches indicate identical substructures in the respective dendrograms. The dendrograms were untangled using the step2side algorithm.

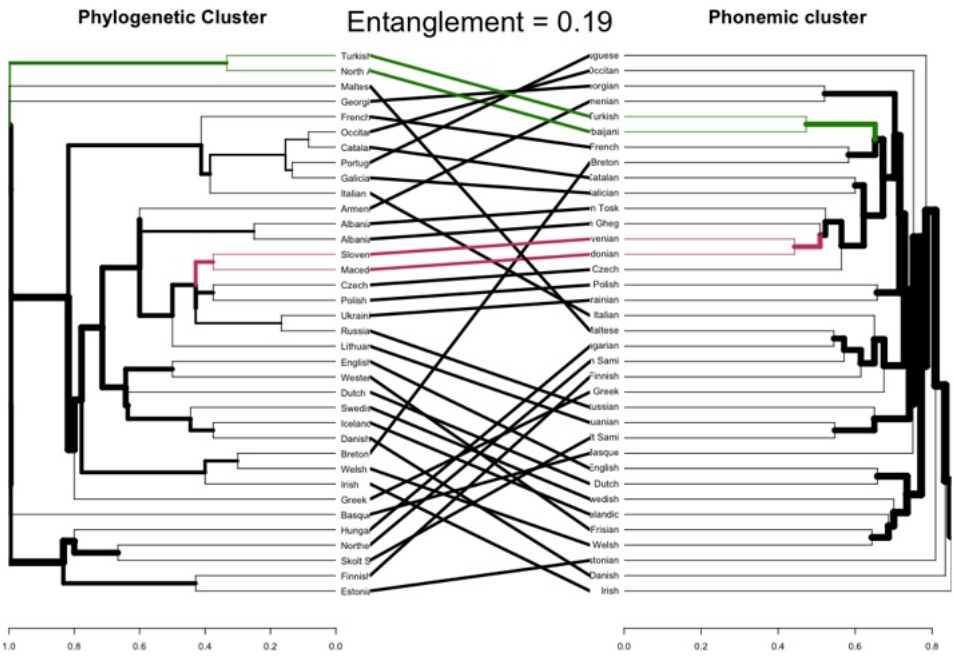

Figure 7.2: A tanglegram showcasing the structural alignment between the dendrograms resulting from hierarchical clustering based on phylogenetic and phonemic similarities. Colored branches indicate identical substructures in the respective dendrograms. The dendrograms were untangled using the step2side algorithm.

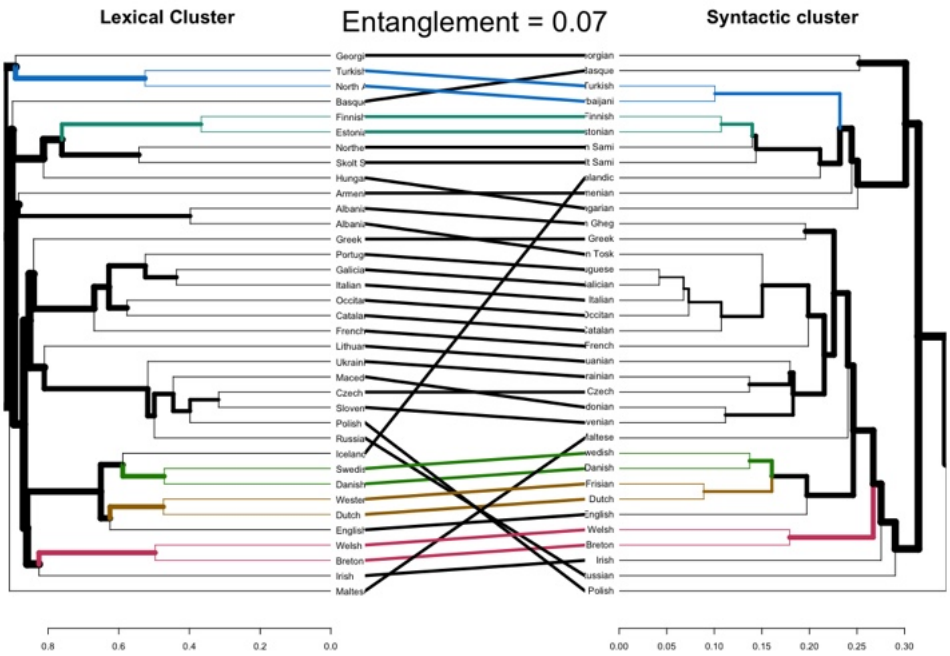

Figure 7.3: A tanglegram showcasing the structural alignment between the dendrograms resulting from hierarchical clustering based on lexical and morphosyntactic similarities. Colored branches indicate identical substructures in the respective dendrograms. The dendrograms were untangled using the step2side algorithm.

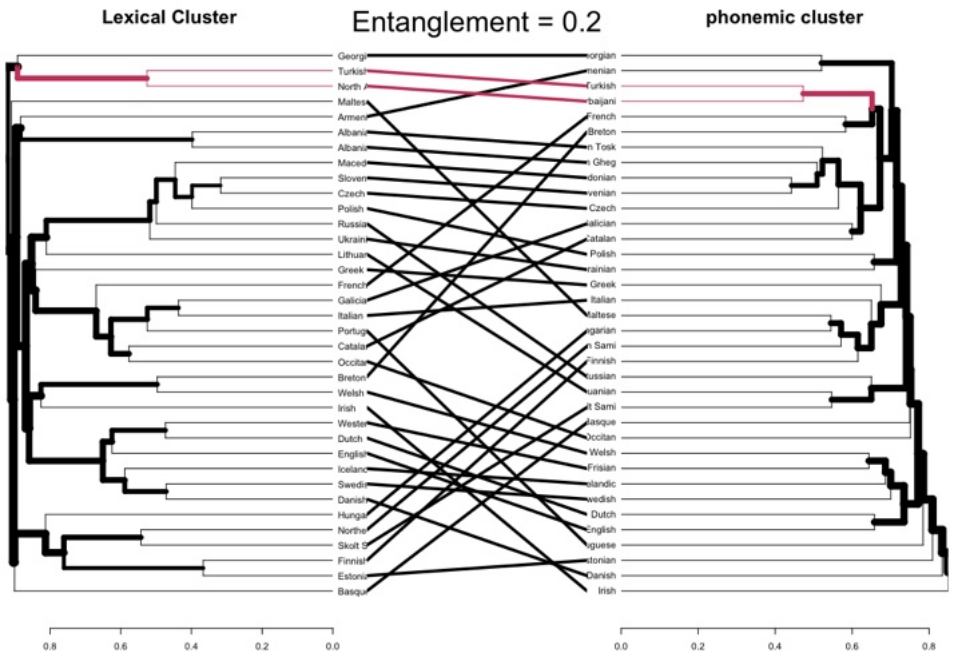

Figure 7.4: A tanglegram showcasing the structural alignment between the dendrograms resulting from hierarchical clustering based on phonemic and lexical similarities. Colored branches indicate identical substructures in the respective dendrograms. The dendrograms were untangled using the step2side algorithm.

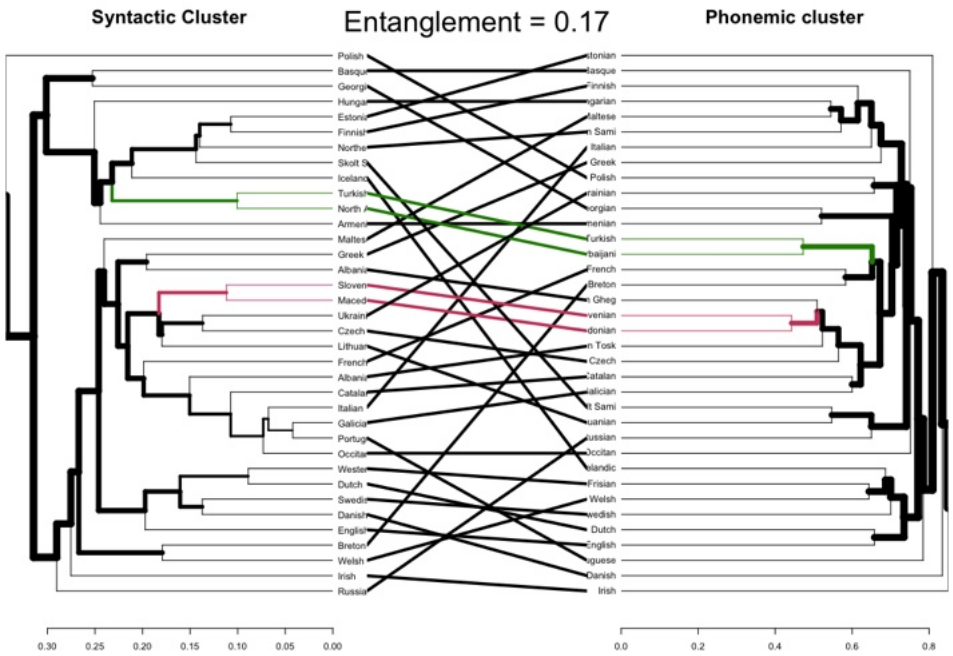

Figure 7.5: A tanglegram showcasing the structural alignment between the dendrograms resulting from hierarchical clustering based on phonemic and morphosyntactic similarities. Colored branches indicate identical substructures in the respective dendrograms. The dendrograms were untangled using the step2side algorithm.

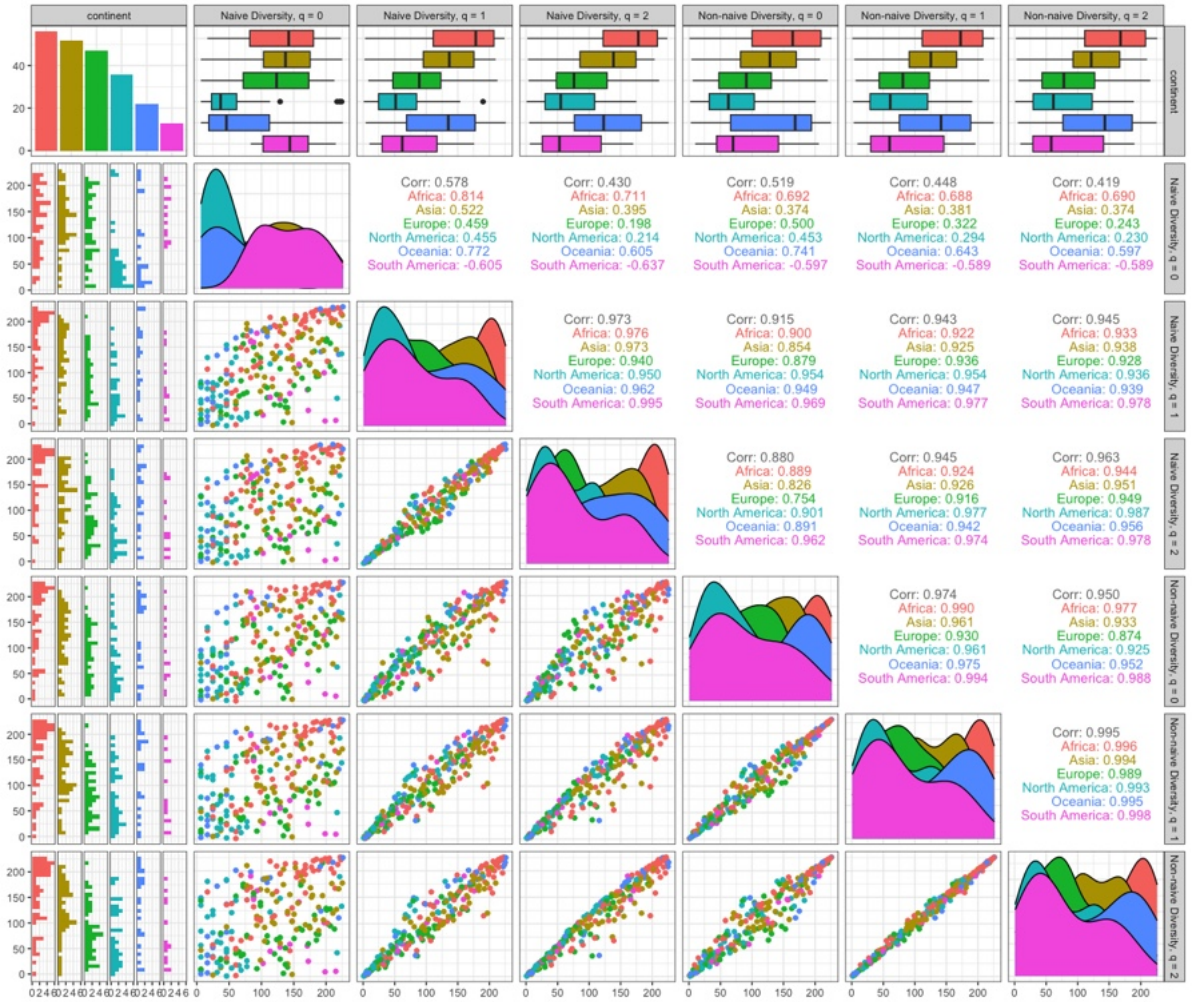

Figure 7.6: A correlogram showing the relationship between naive and and non-naive diversity for q = 0, 1, and 2, disaggregated on continents. The correlation coefficient given is Spearman's ρ

List of Figures

List of Tables

Bibliography

Ammon, U. (2010). World languages: Trends and futures. In N. Coupland (Ed.), *The handbook of language and globalization* (pp. 101–122). Oxford, UK: Wiley-Blackwell.

Anderson, C., Tresoldi, T., Greenhill, S. J., Forkel, R., Gray, R. D., & List, J.-M. (2021, September). Measuring variation in phoneme inventories. Retrieved from `https://doi.org/10.21203/rs.3.rs-891645/v1` (Preprint) doi: 10.21203/rs.3.rs-891645/v1

Archibald, J. (2022). Using jaccard distance to measure the linguistic i-proximity of phonological inventories in a contrastive hierarchy. In *Workshop on l3 processing and acquisition.* King's College London.

Automated Similarity Judgment Program (ASJP). (2024). *Asjp help page.* Retrieved from `https://asjp.clld.org/help` (Accessed: 2025-04-16)

Açıkalın, N. (2024). The institutionalization of turkish diaspora and its impact on domestic and foreign agenda. *Osservatario Turchia, CESPI*.

Bailyn, J. F. (2010). To what degree are croatian and serbian the same language? evidence from a translation study. *Journal of Slavic Linguistics*, *18*(2), 181–219. Retrieved from `http://www.jstor.org/stable/24600116`

Bakker, D., Müller, A., Velupillai, V., Wichmann, S., Brown, C. H., Brown, P., … Holman, E. W. (2009, May). Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, *13*(1), 169–181. Retrieved from `https://doi.org/10.1515/LITY.2009.009` doi: 10.1515/LITY.2009.009

Bouckaert, R., Redding, D., Sheehan, O., Kyritsis, T., Gray, R., Jones, K. E., & Atkinson, Q. (2022, July 20). *Global language diversification is linked to socio-ecology and threat status.* Retrieved from `https://doi.org/10.31235/osf.io/f8tr6` (Preprint, OSF) doi: 10.31235/osf.io/f8tr6

Bouwer, L. (2007). Intercomprehension and mutual intelligibility among southern malagasy languages. *Language Matters*, *38*(2), 253–274. Retrieved from `https://doi.org/10.1080/10228190701794624` doi: 10.1080/10228190701794624

Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., … Hua, X. (2022, February). Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology Evolution*, *6*(2), 163–173. doi: 10.1038/s41559-021-01604-y

Brown, C. H., Holman, E. W., Wichmann, S., & Velupillai, V. (2008). Automated classification of the world's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals*, *61*(4), 285–308. doi: 10.1524/stuf.2008.0026

Börstell, C., Crasborn, O., & Whynot, L. (2020). Measuring lexical similarity across sign languages in global signbank. In *Proceedings of the lrec2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of*

the language community, technological challenges and application perspectives (pp. 21–26). Marseille, France: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/2020.signlang-1.4/

Cambridge University Press. (n.d.). *Diversity.* Cambridge University Press. Retrieved from https://dictionary.cambridge.org/dictionary/english/diversity (Accessed: 2025-05-20)

Carr, J. R. C. (2022). Linguistic landscapes. *Oxford Bibliographies*. Retrieved from https://www.oxfordbibliographies.com/abstract/document/obo-9780199772810/obo-9780199772810-0251.xml doi: 10.1093/obo/9780199772810-0251

Chao, A., Chiu, C.-H., & Jost, L. (2014, November). Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through hill numbers. *Annual Review of Ecology, Evolution, and Systematics*, *45*(Volume 45, 2014), 297–324. doi: 10.1146/annurev-ecolsys-120213-091540

Chao, A., Chiu, C.-H., & Jost, L. (2016). Phylogenetic diversity measures and their decomposition: A framework based on hill numbers. In R. Pellens & P. Grandcolas (Eds.), *Biodiversity conservation and phylogenetic systematics: Preserving our evolutionary heritage in an extinction crisis* (p. 141–172). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-22461-9$_8$ doi: 10.1007/978-3-319-22461-9$_8$

Clarke, K. R., Gorley, R. N., Somerfield, P. J., & Warwick, R. M. (2014). *Change in marine communities: An approach to statistical analysis and interpretation* (3rd ed.). Plymouth, UK: PRIMER-E.

Clarke, K. R., Somerfield, P. J., & Gorley, R. N. (2016, October). Clustering in non-parametric multivariate analyses. *Journal of Experimental Marine Biology and Ecology*, *483*, 147–155. doi: 10.1016/j.jembe.2016.07.010

Dahl, (2008, August). An exercise in a posteriori language sampling. *Sprachtypologie und Universalienforschung: STUF*, *61*(3), 208–220.

Deacon, C. (2016). *Portugal: Europe's mistaken identity* (Master's thesis, University of Amsterdam, Amsterdam). Retrieved from https://www.fon.hum.uva.nl/archive/2016/2016-MA-ChrisDeacon.pdf (Supervised by Silke Hamann)

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online (v2020.4)* [Data set]. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.13950591 doi: 10.5281/zenodo.13950591

Dunn, M., Levinson, S. C., Lindström, E., Reesink, G., & Terrill, A. (2008). Structural phylogeny in historical linguistics: Methodological explorations applied in island melanesia. *Language*, *84*(4), 710–759. doi: 10.1353/lan.0.0062

Dyen, I., Kruskal, J. B., & Black, P. (1992). *An indoeuropean classification: A lexicostatistical experiment* (Vol. 82) (No. 5). Philadelphia: The American Philosophical Society.

Eberhard, G. F. S., David M., & Fennig, C. D. (2025). *Ethnologue: Languages of the world. twenthy-eight edition*. Dallas, Texas: SIL International.

Elgendy, R., Younes, A., Abu-Donia, H. M., & Farouk, R. M. (2024). Efficient quantum algorithms for set operations. *Scientific Reports*, *14*, 7015. Retrieved from https://www.nature.com/articles/s41598-024-56860-2 doi: 10.1038/s41598-

024-56860-2

Enfield, N. J., & Comrie, B. (Eds.). (2015). *Languages of mainland southeast asia: The state of the art*. Berlin, München, Boston: De Gruyter Mouton. Retrieved from https://doi.org/10.1515/9781501501685 doi: 10.1515/9781501501685

Ethnologue. (2025a). *About ethnologue.* Retrieved 2025-03-18, from http://ethnologue.com/about/ (Accessed: 2025-03-18)

Ethnologue. (2025b). *About ethnologue.* Retrieved from https://www.ethnologue.com/methodology/Status (Accessed: 2025-03-06)

Everett, C. (2017). Languages in drier climates use fewer vowels. *Frontiers in Psychology*, *8*, 1285. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01285/full doi: 10.3389/fpsyg.2017.01285

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Chichester, UK: Wiley.

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth*, *8*(2), 195–222.

Forkel, R., & Haspelmath, M. (2023). *Clld: Cross-linguistic linked data.* Max Planck Institute for Evolutionary Anthropology. Retrieved from https://clld.org

Fought, C. (2011). Language and ethnicity. In R. Mesthrie (Ed.), *The cambridge handbook of sociolinguistics* (pp. 238–258). Cambridge University Press. Retrieved from https://www.cambridge.org/core/books/abs/cambridge-handbook-of-sociolinguistics/language-and-ethnicity/61E592C820248EF08082276E387BD39F doi: 10.1017/CBO9780511997068.019

Galili, T. (2015). dendextend: an r package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*. Retrieved from https://doi.org/10.1093/bioinformatics/btv428 doi: 10.1093/bioinformatics/btv428

Gamallo, P., Pichel, J. R., & Alegria, I. (2020, April). Measuring language distance of isolated european languages. *Information*, *11*(44), 181. doi: 10.3390/info11040181

Giacalone, C. G. (2016). Sicilian language usage: Language attitudes and usage in sicily and abroad. *Italica*, *93*(2), 305–316.

Gibson, H., Guérois, R., Mapunda, G., & Marten, L. (Eds.). (2024). *Morphosyntactic variation in east african bantu languages: Descriptive and comparative approaches* (Vol. 8). Berlin: Language Science Press. Retrieved from https://langsci-press.org/catalog/book/383 doi: 10.5281/zenodo.10453704

Ginsburgh, V., & Weber, S. (2020). The economics of language. *Journal of Economic Literature*, *58*(2), 348–404. Retrieved from https://www.jstor.org/stable/27030435

Glottolog. (2024). *Glottolog information.* Retrieved from https://glottolog.org/glottolog/glottologinformation (Accessed: 2025-03-12)

Glottolog Project. (2024). *Glottolog glossary.* https://glottolog.org/meta/glossary. (Accessed: 2025-04-17)

Goldin-Meadow, S., & Brentari, D. (2017). Gesture, sign, and language: The coming of age

of sign language and gesture studies. *The Behavioral and Brain Sciences*, *40*, e46. doi: 10.1017/S0140525X15001247

Gooskens, C., & Heuven, V. J. v. (2020). *How well can intelligibility of closely related languages in europe be predicted by linguistic and non-linguistic variables?* (Vol. 10) (No. 3). John Benjamins. doi: https://doi.org/10.1075/lab.17084.goo

Gooskens, C., van Heuven, V. J., Golubović, J., Schüppert, A., Swarte, F., , & Voigt, S. (2018, April). Mutual intelligibility between closely related languages in europe. *International Journal of Multilingualism*, *15*(2), 169–193. doi: 10.1080/14790718.2017.1350185

Gooskens, C., van Heuven, V. J., van Bezooijen, R., & Pacilly, J. J. A. (2010). Is spoken danish less intelligible than swedish? *Speech Communication*, *52*(11-12), 1022–1037. Retrieved from `https://doi.org/10.1016/j.specom.2010.06.005` doi: 10.1016/j.specom.2010.06.005

Grambank. (2024a). *Background of the grambank questionnaire: Questionnaire history.* `https://github.com/grambank/grambank/wiki/Background-of-the-Grambank-questionnairequestionnaire-history`. (Accessed: 2025-03-25)

Grambank. (2024b). *Binarised features.* `https://github.com/grambank/grambank/wiki/Binarised-features`. (Accessed: 2025-03-25)

Grambank. (2024c). *Faq (general).* `https://github.com/grambank/grambank/wiki FAQ-(general)`. (Accessed: 2025-03-25)

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

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

Gurevich, T., Herman, P. R., Toubal, F., & Yotov, Y. V. (2025, March). A dataset on linguistic connectivity across and within countries. *Scientific Data*, *12*(1), 542. doi: 10.1038/s41597-025-04692-8

Hammarström, H. (2015). Glottolog: A free, online, comprehensive bibliography of the world's languages. In E. Kuzmin (Ed.), *Proceedings of the 3rd international conference on linguistic and cultural diversity in cyberspace* (pp. 183–188). Moscow, Russia: UNESCO. Retrieved from `https://www.mpi.nl/publications/item2354764/glottolog-free-online-comprehensive-bibliography-worlds-languages`

Hammarström, H. (2016, January). Linguistic diversity and language evolution. *Journal of Language Evolution*, *1*(1), 19–29. doi: 10.1093/jole/lzw002

Hammarström, H., & Donohue, M. (2014, January). Some principles on the use of macro-areas in typological comparison. Retrieved from `https://brill.com/view/journals/ldc/4/1/article-p167`$_5$`.xml` doi: 10.1163/22105832-00401001

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2025). *Glottolog 5.2.* Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://glottolog.org` (Available online at http://glottolog.org, Accessed on 2025-06-03) doi: 10.5281/zenodo.15525265

Hammarström, H., & O'Connor, L. (2013, September). Dependency-sensitive typological distance. In L. Borin & A. Saxena (Eds.), *Approaches to measuring linguistic differences* (p. 329–352). DE GRUYTER. Retrieved from

https://www.degruyter.com/document/doi/10.1515/9783110305258.329/html doi: 10.1515/9783110305258.329

Harmon, D., & Loh, J. (2010). The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages. *Language Documentation & Conservation*, *4*, 97–151. Retrieved from http://hdl.handle.net/10125/4474

Hartmann, F., Roberts, S. G., Valdes, P., & Grollemund, R. (2024, January). Investigating environmental effects on phonology using diachronic models. *Evolutionary Human Sciences*, *6*, e8. doi: 10.1017/ehs.2023.33

Hassenstein, M. J., & Vanella, P. (2022). Data quality—concepts and problems. *Encyclopedia*, *2*(1), 498–510. doi: 10.3390/encyclopedia2010032

Heeringa, W., Gooskens, C., & van Heuven, V. J. (2023, May). Comparing germanic, romance and slavic: Relationships among linguistic distances. *Lingua*, *287*, 103512. doi: 10.1016/j.lingua.2023.103512

Heeringa, W., Johnson, K., & Gooskens, C. (2009). Measuring dialect distance phonetically. *Speech Communication*, *51*(2), 167-183. doi: 10.1016/j.specom.2008.07.006.

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*(2), 427–432. doi: 10.2307/1934352

Hoberman, R. D. (2007). Maltese morphology. In A. S. Kaye (Ed.), *Morphologies of asia and africa* (pp. 257–281). Winona Lake, Indiana: Eisenbrauns. Retrieved from

Hoffmann, K., Bouckaert, R., Greenhill, S. J., & Kühnert, D. (2021, July). Bayesian phylogenetic analysis of linguistic data using beast. *Journal of Language Evolution*, *6*(2), 119–135. Retrieved from https://doi.org/10.1093/jole/lzab005 doi: 10.1093/jole/lzab005

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., & Bakker, D. (2008). Explorations in automated language classification. *Folia Linguistica*, *42*(3-4), 331–354. doi: 10.1515/FLIN.2008.331

Hooper, D. U., III, F. S. C., Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., … Wardle, D. A. (2005). Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecological Monographs*, *75*(1), 3–35. Retrieved from https://esajournals.onlinelibrary.wiley.com/doi/10.1890/04-0922 doi: 10.1890/04-0922

International Phonetic Association. (1999). *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet*. Cambridge, UK: Cambridge University Press. Retrieved from https://www.cambridge.org/highereducation/books/handbook-of-the-international-phonetic-association/6A48F0458C3A1F172D5ED09D41E754B6

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37-50. Retrieved from https://doi.org/10.1111/j.1469-8137.1912.tb05611.x doi: 10.1111/j.1469-8137.1912.tb05611.x

Joshua Project. (2025). *Joshua project: People groups of the world.* Retrieved from https://www.joshuaproject.net/ (Accessed: 2025-04-08)

Jost, L. (2018, 07). What do we mean by diversity? the path towards quantification. *Metode*, *2019*. doi: 10.7203/metode.9.11472

Jost, L., & Chao, A. (n.d.). *Diversity analysis: a fresh approach.*

Kennard, H. J. (2021). Variation in breton word stress: new speakers and the influence of french. *Phonology*, *38*(3), 363–399. Retrieved from https://doi.org/10.1017/S0952675721000245 doi: 10.1017/S0952675721000245

Khan, A., Shipton, M., Anugraha, D., Duan, K., Hoang, P. H., Khiu, E., … Lee, E.-S. A. (2025, January). URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 6937–6952). Abu Dhabi, UAE: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.coling-main.463/

Koch, H. (2014). Language contact and the comparative method in australian linguistics. In H. Koch & R. Nordlinger (Eds.), *The languages and linguistics of australia: A comprehensive guide* (pp. 579–624). De Gruyter Mouton. Retrieved from https://www.degruyter.com/document/doi/10.1515/9783110279771.23/html doi: 10.1515/9783110279771.23

Kortmann, B., & van der Auwera, J. (Eds.). (2011a). *The languages and linguistics of europe: A comprehensive guide*. Berlin, Boston: De Gruyter Mouton. Retrieved from https://doi.org/10.1515/9783110220261 doi: 10.1515/9783110220261

Kortmann, B., & van der Auwera, J. (Eds.). (2011b). *The languages and linguistics of europe: A comprehensive guide*. Berlin, Boston: De Gruyter Mouton. Retrieved from https://doi.org/10.1515/9783110220261 doi: 10.1515/9783110220261

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. *Ecology*, *93*(3), 477–489. doi: 10.1890/10-2402.1

Leslie, L. M., Kim, Y. L., & Ye, E. R. (2025). Diversity initiatives: Intended and unintended effects. *Current Opinion in Psychology*, *61*, 101942. doi: https://doi.org/10.1016/j.copsyc.2024.101942

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710. (Originally published in Russian in 1965)

List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., & Gray, R. D. (2022, June). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, *9*(1), 316. doi: 10.1038/s41597-022-01432-0

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., & Levin, L. (2017, April). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers* (pp. 8–14). Valencia, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/E17-2002/

Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, *31*(19), R1174–R1177. doi: https://doi.org/10.1016/j.cub.2021.07.049

Majtczak, T. (2014). From analytic to synthetic – origin of some early japanese volitive expressions. *Comparative Legilinguistics*, *19*(3), 97-112. Retrieved from https://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.ojs-nameId-49b94d03-b5da-3b30-b733-c85e11ddfcc0

-year-2014-volume-19-issue-3-article-4899

Mari, L., Wilson, M., & Maul, A. (2023). What is measured? In *Measurement across the sciences*. Springer, Cham. Retrieved from https://doi.org/10.1007/978-3-031-22448-5$_5$ doi: 10.1007/978-3-031-22448-5$_5$

Moran, S., & McCloy, D. (Eds.). (2019). *Phoible 2.0*. Jena: Max Planck Institute for the Science of Human History. Retrieved from https://phoible.org/

Moran, S. P. (2013, April). Phonetics information base and lexicon. Retrieved from http://hdl.handle.net/1773/22452

National Research Council (US) Committee on Noneconomic and Economic Value of Biodiversity. (1999). *Perspectives on biodiversity: Valuing its role in an everchanging world*. Washington (DC): National Academies Press (US). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK224405/ (Accessed: 2025-05-20)

Nerbonne, J., & Heeringa, W. (2002, 12). Measuring dialect distance phonetically.

Nerbonne, J., & Hinrichs, E. (2006). Linguistic distances. In *Proceedings of the workshop on linguistic distances* (p. 1–6). USA: Association for Computational Linguistics.

Neureiter, N., Ranacher, P., Efrat-Kowalsky, N., et al. (2022). Detecting contact in language trees: a bayesian phylogenetic model with horizontal transfer. *Humanities and Social Sciences Communications*, *9*(205). Retrieved from https://doi.org/10.1057/s41599-022-01211-7 doi: 10.1057/s41599-022-01211-7

Nguyen, N., Chawshin, K., Berg, C. F., & Varagnolo, D. (2022). Shuffle & untangle: Novel untangle methods for solving the tanglegram layout problem. *Bioinformatics Advances*, *2*(1), vbac014. Retrieved from https://doi.org/10.1093/bioadv/vbac014 doi: 10.1093/bioadv/vbac014

Nordhoff, S., & Hammarström, H. (2011). Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the first international workshop on linked science 2011 (lisc2011).* Bonn, Germany. Retrieved from https://ceur-ws.org/Vol-783/paper7.pdf

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., … Borman, T. (2025). vegan: Community ecology package [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=vegan (R package version 2.6-10)

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *187*, 253–318. Retrieved from https://doi.org/10.1098/rsta.1896.0007 doi: 10.1098/rsta.1896.0007

Pearson's correlation coefficient. (2008). In *Encyclopedia of public health* (p. 1090–1091). Dordrecht: Springer Netherlands. Retrieved from https://doi.org/10.1007/978-1-4020-5614-7$_2$569 doi: 10.1007/978-1-4020-5614-7$_2$569

Pereltsvaig, A. (2020). Languages of the caucasus [Book chapter]. In *Languages of the world: An introduction* (pp. 146–180). Cambridge: Cambridge University Press.

Petroni, F., & Serva, M. (2010, June). Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications*, *389*(11), 2280–2283. doi: 10.1016/j.physa.2010.02.004

PHOIBLE. (2025). *Phoible github repository.* https://github.com/phoible.

Ploeger, E., Poelman, W., Høeg-Petersen, A. H., Schlichtkrull, A., Lhoneux, M. d., & Bjerva, J. (2025, March). A principled framework for evaluating on typologically diverse languages. (arXiv:2407.05022). Retrieved from http://arxiv.org/abs/2407.05022 (arXiv:2407.05022 [cs]) doi: 10.48550/arXiv.2407.05022

Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., … Korhonen, A. (2019, September). Modeling language variation and universals: A survey on typological linguistics for natural language processing. , *45*, 559–601. doi: 10.1162/coli_a_00357

R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, *21*(1), 24–43. doi: https://doi.org/10.1016/0040-5809(82)90004-1

Reesink, G., Singer, R., & Dunn, M. (2009). Explaining the linguistic diversity of sahul using population models. *PLoS Biology*, *7*(11), e1000241. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC2770058/ (Accessed: 2025-03-25) doi: 10.1371/journal.pbio.1000241

Rojo, J., Cervigón, P., Ferencova, Z., Cascón, , Galán Díaz, J., Romero-Morte, J., … Gutiérrez-Bustillo, A. M. (2024, March). Assessment of environmental risk areas based on airborne pollen patterns as a response to land use and land cover distribution. *Environmental Pollution*, *344*, 123385. doi: 10.1016/j.envpol.2024.123385

Rowley, A. (2011). Bavarian: Successful dialect or failed language? In J. Fishman & O. García (Eds.), *Handbook of language and ethnic identity. volume 2: The success-failure continuum in language and ethnic identity efforts* (pp. 299–309). Oxford University Press.

Saraçli, S., Doğan, N., & Doğan, (2013, April). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, *2013*(1), 203. doi: 10.1186/1029-242X-2013-203

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

SIL. (2025). *About sil.* Retrieved 2025-03-06, from https://www.sil.org/about (Accessed: 2025-03-06)

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

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., … Gray, R. D. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, *9*(16), eadg6175. Retrieved from https://doi.org/10.1126/sciadv.adg6175 doi: 10.1126/sciadv.adg6175

Skirgård, H., Haynie, H. J., Hammarström, H., Barlow, R., Blasi, D. E., Collins, J., … Gray, R. D. (2023). *Grambank v1.0.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.7844558 doi: 10.5281/zenodo.7844558

Skirgård, H., van der Meer, S., & Hammarström, H. (2014). The nijmegen typological survey (nts). In *44th colloquium on african languages and linguistics.* Leiden, the Netherlands. (Conference presentation)

Spearman, C. (1904). The proof and measurement of association between two

things. *The American Journal of Psychology*, *15*(1), 72–101. Retrieved from http://www.jstor.org/stable/1412159

Swadesh, M. (1955). Towards greater accuracy in lexicostatistical dating. *International Journal of American Linguistics*, *21*(2), 121–137. (S2CID: 144581963) doi: 10.1086/464321

Swadesh, M. (1971). *The origin and diversification of language* (J. Sherzer, Ed.). Chicago: Aldine.

Trask, R. (1997). *The history of basque* (1st ed.). Routledge. Retrieved from https://doi.org/10.4324/9781315004358 doi: 10.4324/9781315004358

Trudgill, P. (2020). Sociolinguistic typology and the speed of linguistic change. *Journal of Historical Sociolinguistics*, *6*(2). Retrieved from https://doi.org/10.1515/jhsl-2019-0015 doi: 10.1515/jhsl-2019-0015

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003, March). *Language vitality and endangerment.* Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000183699 (Document submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages, Paris, 10–12 March 2003)

van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*, 111-122. Retrieved from https://CRAN.R-project.org/package=stringdist

Vaux, B., & Sayeed, O. (2017). The evolution of armenian. In J. Klein, B. Joseph, & M. Fritz (Eds.), *Handbook of comparative and historical indo-european linguistics* (Vol. 2, pp. 1143–1165). Berlin, Boston: De Gruyter Mouton. Retrieved from https://doi.org/10.1515/9783110523874-021 doi: 10.1515/9783110523874-021

Wang, T., Wichmann, S., Xia, Q., & Ran, Q. (2023). Temperature shapes language sonority: Revalidation from a large dataset. *PNAS Nexus*, *2*(12), pgad384. Retrieved from https://doi.org/10.1093/pnasnexus/pgad384 doi: 10.1093/pnasnexus/pgad384

Wichmann, S., & Holman, E. W. (2010). Pairwise comparisons of typological profiles. In *Rethinking universals* (pp. 241–264). De Gruyter Mouton. doi: 10.1515/9783110220933.241

Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010, September). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, *389*(17), 3632–3639. doi: 10.1016/j.physa.2010.05.011

Wichmann, S., Holman, E. W., & Brown, C. H. (2022). *The asjp database (version 20).* https://asjp.clld.org/. (Accessed: 2025-04-14)

Wilk, M. B., & Gnanadesikan, R. (1968, March). Probability plotting methods for the analysis of data. *Biometrika*, *55*(1), 1–17. doi: 10.1093/biomet/55.1.1

Worldometer. (2024). *World population by year.* Retrieved from https://www.worldometers.info/world-population/world-population-by-year/ (Accessed: March 13, 2025)

Yutong, D. (2024). A study on the trend of english changing from synthetic language to analytic language from the perspective of language economics. *Francis Academic Press*. Retrieved from https://francis-press.com/papers/17386

Zeh, K. (2025). *Quantifying the relationship between digitization and linguistic diversity: A multilevel statistical analysis using r* (Master's thesis, University of Vienna). doi: 10.25365/thesis.78469