1. Supplementary Materials

These are the supplementary materials accompanying the paper *Phylogeny as a proxy for structural linguistic diversity: a reasonable approach?* submitted to EVOLANG 2026.

1.1. *Preprocessing*

We here describe the preprocessing applied to the databases to enable the analysis in the paper.

All CLLD datasets were accessed from Zeonodo (European Organization For Nuclear Research & OpenAIRE, 2013) and read into R (R Core Team, 2024) using the rcldf package (Greenhill, 2025), through which all subsequent processing and analysis took place. Since the databases use different unique identifiers, we grouped all languages according to their assigned Glottocode, and if one Glottocode matches multiple unique identifiers, we selected the most data-complete entry. If two languages were equally data-complete, we introduced a bias by selecting the first one as it appears alphabetically. In instances where Glottocodes were lacking, we removed the entries: there were two such cases in PHOIBLE, none in Grambank, and 423 in ASJP. To evaluate completeness, we used all 195 features in Grambank and the 40-item Swadesh list entries in ASJP (Holman et al., 2008). Regarding phoneme inventory, the idea of 'completeness' does not exist; as such, we always selected the smallest inventory when more than one inventory existed to get a conservative estimation of phoneme inventories.

For the coverage analysis, we intersected the datasets on their uniquely identifying Glottocode, yielding a new dataset containing the largest number of languages between which linguistic similarity can be estimated along all dimensions. For the datasets and their intersection, we counted the number of languages and expressed the coverage as the percentage of all languages in Glottolog. For speaker numbers, we joined each dataset with the Ethnologue and Joshua project data according to their Glottocode. We counted the total number of speakers in each dataset, expressing the coverage as the percentage of all speakers in the Ethnologue and Joshua project data. Since the Ethnologue and Joshua data did not cover all languages from Glottolog, there were some mismatches, and as a result, the speaker analysis failed to account for 394 languages in PHOIBLE, 555 languages in ASJP, 319 languages in Grambank, and 94 languages in their intersection. Regarding macroarea, we compared the proportion of languages assigned to each macroarea in Glottolog across the databases relative to the distribution in Glottolog.

1.2. *Visualizations*

In this part of the supplementary material, we present visualizations of our data that could not be included in the submission.

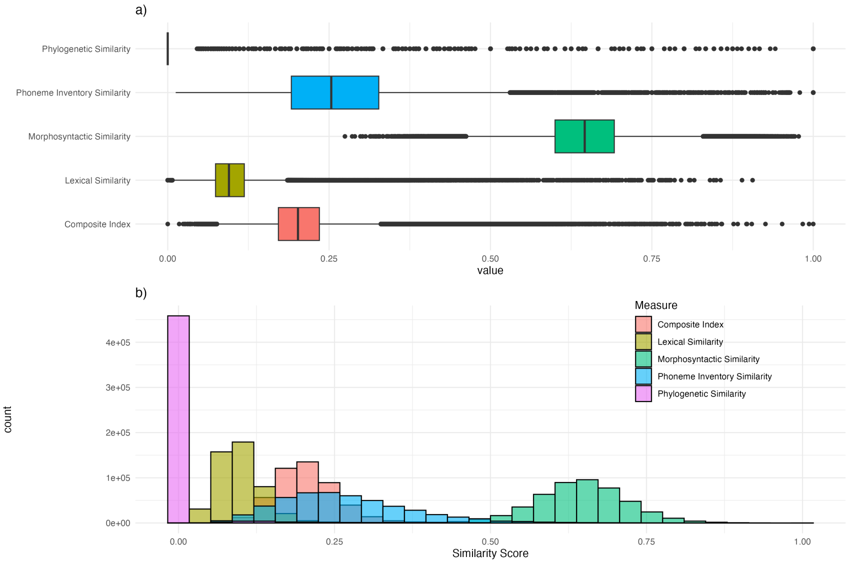

Figure 1. A visualization of the distribution of all pairwise similarities between the languages found in the intersection between PHOIBLE, Grambank, ASJP and Glottolog through a) boxplots and b) histograms. Note that most phylogenetic similarity values are 0 as a result of most languages not being related.

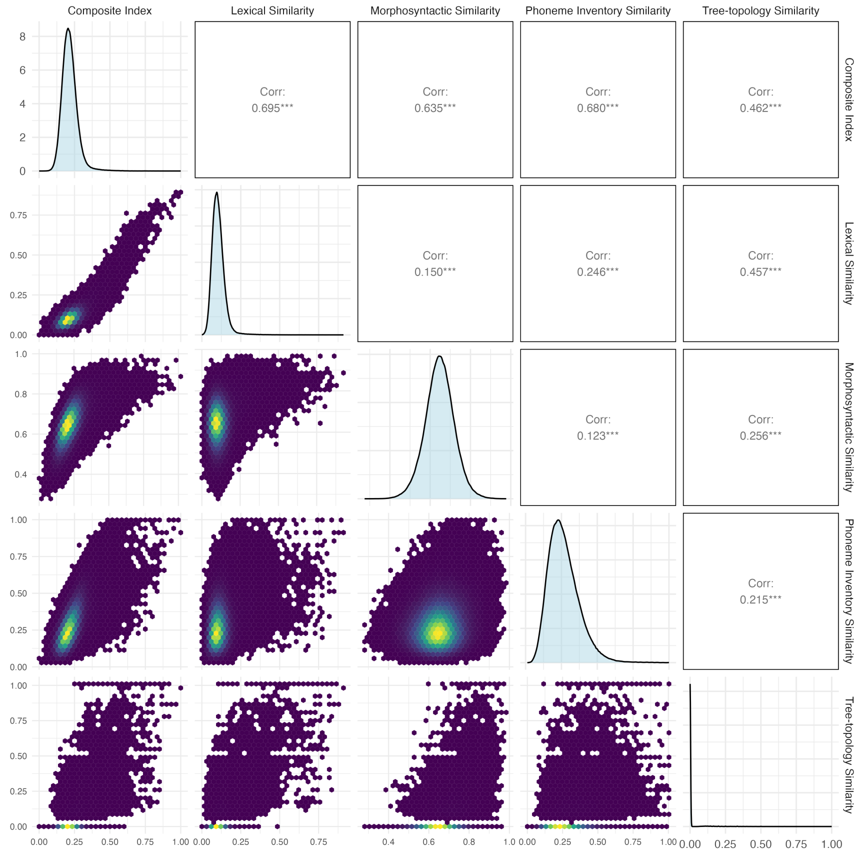

Figure 2. A correlationplot showing the correlation between all pairwise similarities derived from the databases. The upper triangle displays the correlation coefficient measured using Pearson's r. *** signifies $p < 0.001$. The diagonal contains a density plot of the respective measure, and the under triangle a hexbin plot. Yellow color corresponds to a larger count of observations falling in that bin. The hexbin plots show that lexical similarity best captures variance in the composite index, and that the relationship between the similarity measures are complex with a bounding relationship. For example, lexical similarity bounds morphosyntactical similarity such that high lexical similarity implies high morphosyntactical similarity, but high morphosyntactical similarity does not imply high lexical similarity.

1.3. *Similarity measures*

We here formally present the measures of similarity employed in the paper.

1.3.1. *Phoneme inventory similarity*

Let the languages A and B be represented by the set of phonemes a and b. Then, their phoneme inventory similarity s is defined as

$$s(A, B) = \frac{|a \cap b|}{|a \cup b|}. \tag{1}$$

1.3.2. *Morphosyntactic similarity*

Let the two languages A and B be represented by two feature vectors taking on categorical values. The morphosyntactical similarity s between A and B is then defined as

$$s(A, B) = \frac{\#_{i \in DEF(A,B)} A[i] = B[i]}{|DEF(A, B)|} \tag{2}$$

where $DEF(A, B)$ is the set of features for which both A and B are defined.

1.3.3. *Lexical similarity*

Let c be a concept defined for the languages A and B and 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 fewest insertions, deletions, and substitutions necessary to turn a_c into b_c, denoted $d_l(a_c, b_c)$. The normalized Levenshtein distance (LDN) is then defined as

$$LDN(a_c, b_c) = \frac{d_l(a_c, b_c)}{max(l(a_c), l(b_c))} \tag{3}$$

where *l(*a_c*)* and *l(*b_c*)* are the lengths of the strings. Then, the lexical similarity s is defined as

$$s(A, B) = 1 - \frac{1}{|C|} \sum_{c \in C} LDN(a_c, b_c). \tag{4}$$

1.3.4. *Tree-topology similarity*

Let the set of intermittent phylogenetic stages from the root to the node of language A be a, and the equivalent set for the language B be b. Then, the tree-topology similarity between A and B, s, is defined as

$$s(A, B) = \frac{|a \cap b|}{|a \cup b|}. \tag{5}$$

Acknowledgements

References

European Organization For Nuclear Research, & OpenAIRE. (2013). *Zenodo.* CERN.

Greenhill, S. J. (2025). *rcldf: Read linguistic data in the cross linguistic data format (cldf).* (R package version 1.3.1, commit ab9554e763c646a5ea6a49fc0989cf9277322443)

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., & Bakker, D. (2008). Explorations in automated language classification. *42*(3–4), 331–354.

R Core Team. (2024). *R: A language and environment for statistical computing.* Vienna, Austria.