Replicating the Work of Hartmann & Walkdeen (2024) on Phylogenetic Signal in Syntactic Data with Maximum Parsimony and NeighborNet

By Lofty-John C. Anyanwu

Introduction

Phylogenetic methods have long been applied to linguistic data (lexical, phonological, morphosyntactic) to infer historical relationships. The use of syntactic features for this purpose remains controversial: some argue that syntax is more conservative and thus useful for deep-time inference, while others find that syntactic similarity often reflects contact or convergent tendencies. Various studies have tested these ideas using large syntactic databases (Dunn et. al, 2005; Hartmann & Walkden, 2024). Notably, analyzed 110 languages from the Syntactic Structures of the World's Languages (SSWL) database, applying Bayesian phylogenetic inference, and reported generally *weak* signal: deep subgroupings (e.g. major families) did not emerge with high support (Koopman, 2012). Their focus was on quantifying *"apparent"* signal from computational models of syntax, acknowledging that borrowing or typological biases may obscure true genealogical signal.

Previous work offers mixed expectations. For instance, Dunn et al. (2005, 2008) and Wichmann & Saunders (2007) successfully used syntactic characters to resolve certain language families (Papuan/Oceanic; Americas), but other studies (e.g. Shu et al. 2018) found shallow trees from raw syntactic data. In particular, Shu et al. (2018) applied Hamming distances and Neighbor-Joining to a subset of SSWL and obtained a very unresolved global tree – they noted only modest success for a small "Romance" subset. More generally, Gray et al. (2010) and have highlighted that NeighborNet graphs and statistics like delta (d) and Q-residual can quantify how non-"tree-like" linguistic datasets are.

Our goal is to replicate and evaluate Hartmann & Walkden's findings using two standard phylogenetic methods: Maximum Parsimony (a character-based tree method) and NeighborNet (a distance-based network). We use the same dataset ("onlyextant.nex" from Hartmann's repository) to generate (1) a parsimonious tree with bootstrap support, and (2) a NeighborNet splits graph with network metrics. This paper proceeds as follows: the second part describes the dataset and methods, the third part reports results from parsimony and NeighborNet, the fourth part discusses how these compare to the original findings, and the fifth part concludes with implications and future directions.

Data and Methods

Dataset and Preprocessing

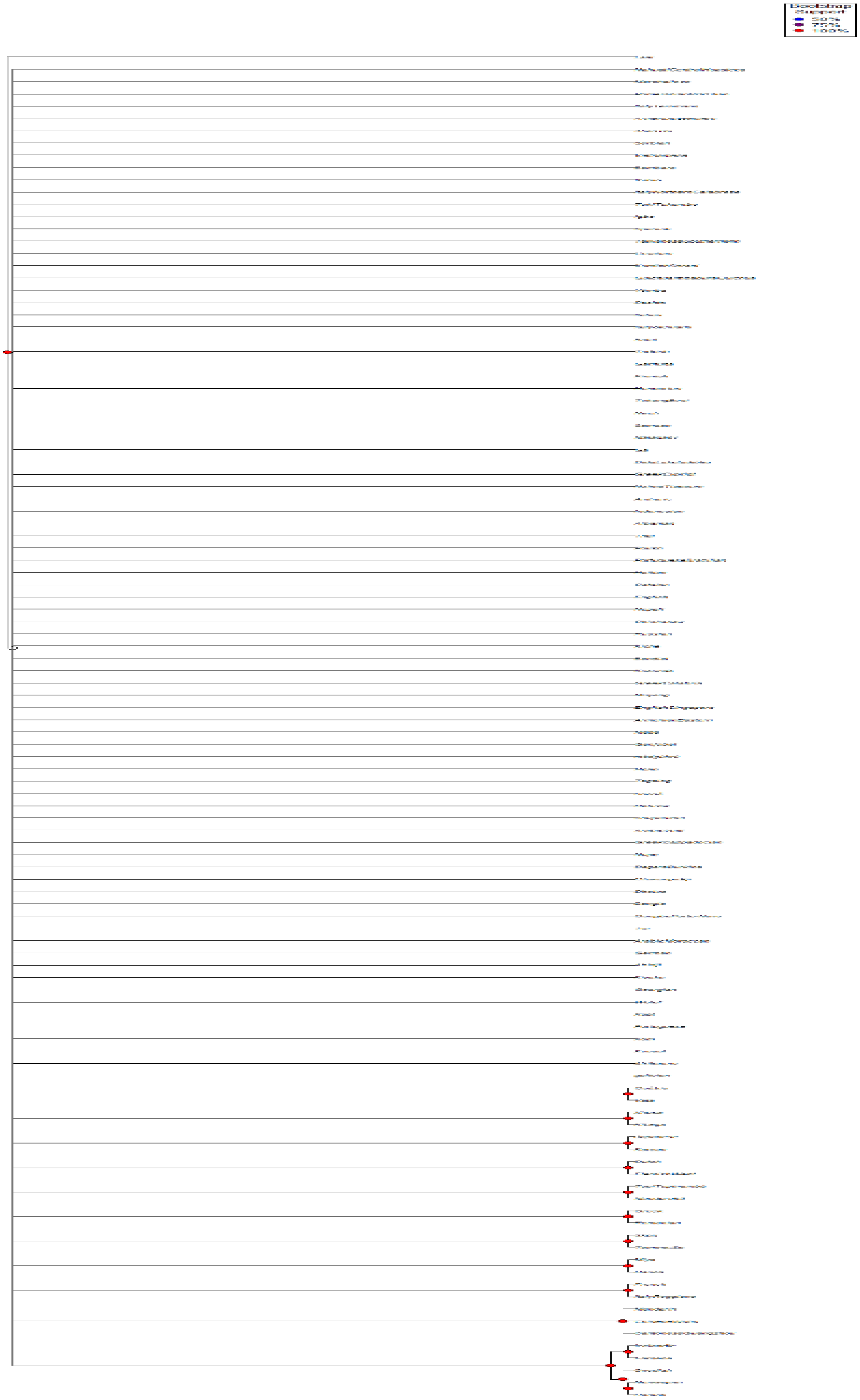
We used the SSWL syntactic dataset of 110 modern languages as compiled by Hartmann & Walkden (2024). These languages span multiple families worldwide. The raw SSWL database (Koopman 2012–; TerraLing group) codes 129 syntactic features as three-valued characters: *"Yes"*, *"No"* or *"Not Applicable"* (NA). In the Nexus file provided (onlyextant.nex), states are encoded as 1 (yes), 2 (no), 3 (NA), with missing data as ?. We converted the data to a fully sequential (non-interleaved) format for analysis. Care was taken to preserve the coding: every property where a language lacks the relevant construction was coded 3, and any unknown entries as ?. No further gap symbols were used except as above. Thus the dataset is a **mixed** binary/multistate matrix of 110 taxa × 129 characters (states "1,2,3,?"), reflecting different syntactic traits across languages.

The choice of SSWL means the features are not pre-filtered for phylogenetic informativeness: they include many typological or typologically dependent traits (subject to entailment). For context, Hartmann & Walkden note that SSWL properties are narrower than classic parameters and often have implicational dependencies. We follow their approach of treating all features equally (no exclusions) but record tree metrics (Consistency Index, Retention Index) later to gauge homoplasy. All 110 languages are extant; no extinct or reconstructed taxa are included, so we rooted our analyses using Laal (an isolate of Chad) as a functional outgroup, similarly to other typology studies where a linguistic isolate or distantly related language is used to orient the tree.

Maximum Parsimony

(Notes: The maximum parsimony trees cannot be included in this paper due to size issues; there are 110 taxa and the graphs are too small to show the clades but the R Script and the knitted file in the repo will show these. One is shown below as a reference)

For Method #1, we performed a maximum parsimony (MP) analysis using the *phangorn* package in R. The data matrix was input as phyDat format. We ran a parsimony ratchet search (*pratchet* in phangorn) to find the minimal-length tree. The single best MP tree (Supplementary Material S2) had 1316 total steps. We rooted this tree on Laal (our outgroup). Clade support was assessed by nonparametric bootstrapping: we generated 100 bootstrap replicates of the data, reran pratchet on each, and summarized the results as a majority consensus tree (branches present in ≥50% of replicates; Supplementary Materials). Bootstrap support values at each node were recorded. We also computed the parsimony fit indices on the MP tree: the Consistency Index (CI), Retention Index (RI), and their product (Rescaled Consistency, RC). These indices quantify homoplasy: CI = (minimum possible changes / observed changes), RI = [(max changes – observed changes)/(max – min)], and RC = CI·RI (Farris 1989; Pant et. al. 2022). Low CI/RI indicate many convergences or parallelisms.


Bootstrap
Support
50%
75%
100%

NeighborNet (SplitsTree)

For Method #2, we used SplitsTree (App v6) to compute a NeighborNet from the same data. We first computed a pairwise P-distance matrix among the 110 languages (treating all non-? character differences equally, i.e. Hamming distance). In SplitsTree, we loaded the distance matrix and ran the NeighborNet algorithm (Bryant & Moulton 2004) to produce a planar splits network (Fig. 1) and then rooted at the midpoint (Fig. 2). The resulting graph represents incompatible splits as parallel edges ("webbing"), visualizing conflicting phylogenetic signals.

We quantified network "tree-likeness" using two metrics. The **delta score** (d; Holland et al. 2002) measures, for each taxon, average deviation from the four-point condition on all quartets containing that taxon. A delta of 0 indicates perfect tree likeness for that taxon; higher δ implies more conflict. The Q-residual (Gray et al. 2010) similarly averages absolute quartet deviations for each language. We computed each taxon's δ and Q. (Delta and Q effectively range 0–1, with 0 for a tree-like dataset.) The number of splits in the NeighborNet (the graph's resolution) was also noted.

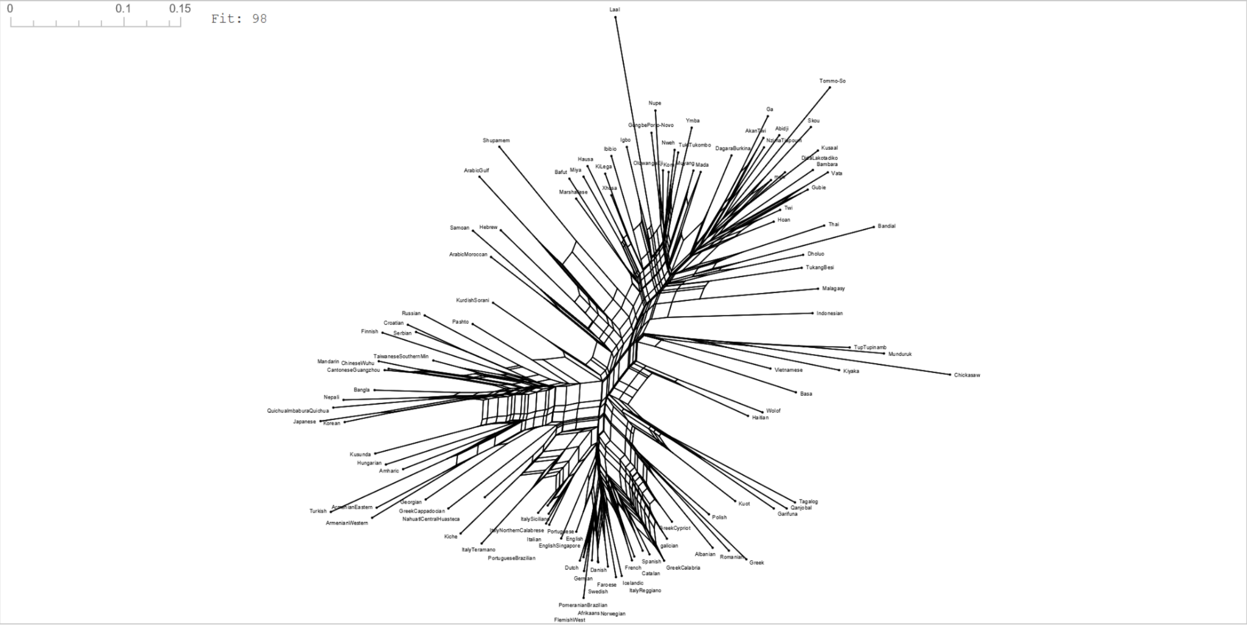


Figure 1: NeighbourNet Split Graph created using SplitsTree v6, unrooted.

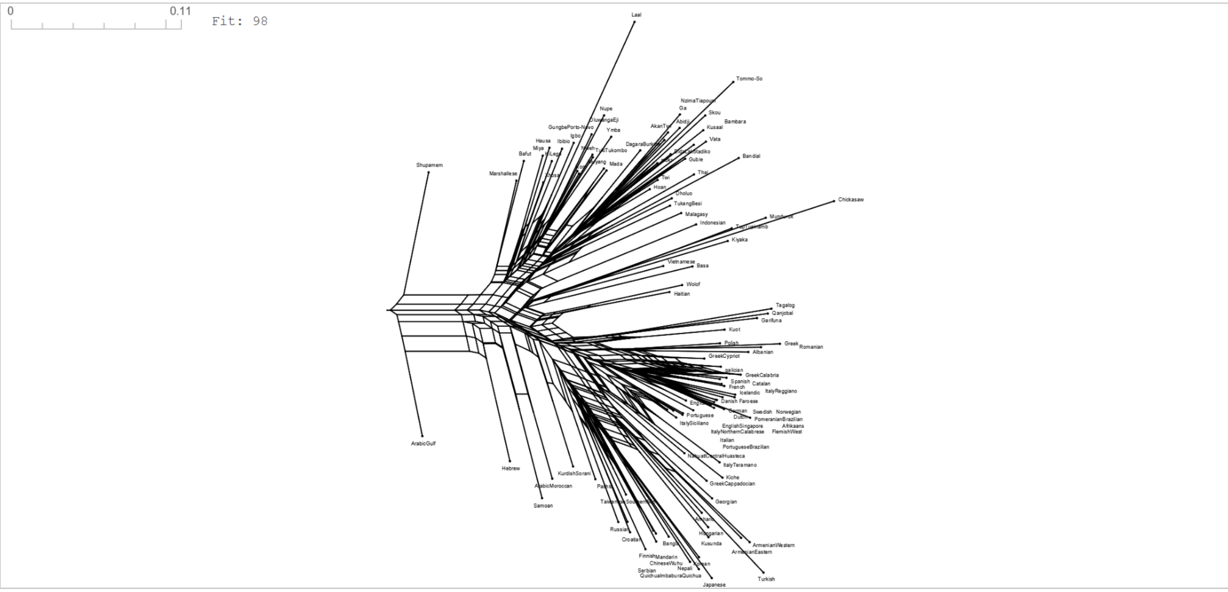


Figure 2: NeighbourNet Split Graph created using SplitsTree v6, rooted at the midpoint.

Results

Maximum Parsimony Results

The MP search yielded a single (most-parsimonious) tree. Its topology shows *some* shallow family-level clusters, but deep structure is largely unresolved. Notably, all North Germanic and West Germanic languages cluster together: for example, Swedish, Icelandic, and others form a strongly supported Germanic clade, with English and Dutch grouping with them (consistent with "Northwestern Indo-European" in the literature). Finno-Ugric languages (Hungarian, Finnish) form a clade, and a few other subfamilies (e.g. Chinese dialects; Niger-Congo subgroups) cluster. However, at higher taxonomic levels (Indo-European vs. Uralic vs. Afroasiatic, etc.), the tree has many long unresolved branches.

Bootstrap support is generally low. Out of all nontrivial clades, only the Germanic subgroups and a handful of local clusters exceed 70% support; most deeper nodes have <50% support. The majority-rule consensus tree collapses most higher nodes into polytomies, reflecting the low resolution. For example, the node subtending all Indo-European languages (including Germanic, Romance, Slavic, etc.) has <10% bootstrap and so is unresolved. In sum, the parsimony tree highlights few well-supported genealogical groupings beyond very recent splits.

The parsimony indices are low, indicating considerable homoplasy. The calculated Consistency Index (CI) was about 0.131 , the Retention Index (RI) about 0.613, yielding an RC of approximately 0.08. (As a reference, a CI near 1 would mean almost no homoplasy, whereas values 0.3–0.5 are typical for noisy linguistic traits.) These values imply that many features change multiple times across the tree. In other words, the syntactic characters do not fit a single tree well, consistent with the poor bootstrap support.

NeighborNet Results

The NeighborNet graph (Figure 1 & 2) reveals **extensive reticulation**. The outline is very "webby": multiple conflicting splits crisscross the graph, and there is no single dominant backbone. Some clades appear as "loops" or parallel edges: for instance, the Germanic languages form one connected cluster, but with several internal boxes indicating conflicting affinities (perhaps due to areal effects or parallel changes). Other languages (Finno-Ugric, Niger-Congo, Sinitic, etc.) similarly sit in roughly coherent regions but with broad networks around them. In general, the network shows few long branches branching cleanly; instead, many shorter edges interwoven with boxes. This visual impression matches that of a dataset with low tree likeness.

Quantitatively, the network has an average delta score across all taxa is ≈ 0.38, and the average Q-residual is ≈ 0.06. For context, studies of other language families find similar numbers. Here, $\delta \neq 0$ and $Q \neq 0$ indicate significant deviation from a perfect tree. Indeed, 0.38 is a fairly high delta (near the upper range for human datasets), implying that on average each language participates in many quartet incompatibilities. The Q value of ~0.08, while numerically small, is nonzero and typical for real linguistic data; it confirms that residual (non-tree) signal is substantial.

Together, these results show that the syntactic data are *not* well approximated by a single tree. The network's rectilinear webbing explicitly depicts conflicting phylogenetic signals (e.g. due to contact or parallel development) for many language clusters. In summary, both the visual graph and the metric scores indicate high reticulation / low tree-likeness in the dataset.

Discussion

Our parsimony and network analyses largely agree with Hartmann & Walkden's original findings of weak deep signal in syntax. The MP tree recovered strong support only for a few shallow groups (notably the Germanic subclades and several narrow family clusters) but failed to recover any well-supported large families. This mirrors Hartmann & Walkden's Bayesian results, which also placed languages into clusters mainly at low time depth. In both studies, higher-level nodes (e.g. "all Indo-European", "all Niger-Congo") collapse. The parsimony bootstraps were low across the board, in line with their finding that posterior probabilities for deep clades were low. Likewise, our low CI/RI values underscore pervasive homoplasy, echoing their conclusion that syntax contains much non-phylogenetic (noise) variation.

The NeighborNet provides complementary insight. Its highly reticulate structure confirms that multiple conflicting histories exist in the data. This network suggests that shared syntactic traits are often better explained by areal spread or parallelism than by clean inheritance. For example, the Germanic cluster appears with some conflict that might hint at language contact (e.g. shared word-order traits across West Germanic and some Romance, as noted by Hartmann & Walkden). Our δ and Q scores quantify this: their moderate values indicate that the dataset deviates considerably from tree-likeness. These findings align with previous literature: Gray et al. (2010) and others note that a nonzero δ or Q implies network-like evolution, as we observe.

There is strong concordance between methods. Both MP and NeighborNet point to the same core conclusion: the SSWL syntactic data show weak genealogical signal. Neither method recovered deep branches with confidence. The few agreements (e.g. Germanic, Chinese varieties, Kwa languages) tend to be very recent splits that are known to share many specific syntactic innovations. One notable difference is that NeighborNet suggests more structure than the MP consensus: even where MP collapsed, the network often shows *some* grouping (albeit with parallelograms). This is expected, since NeighborNet does not enforce a single tree and can display even weak splits. However, those splits are not strongly supported by characters (hence parallel edges). In practice, the network reinforces the parsimony picture rather than overturning it.

Methodologically, each approach has strengths and limits here. Maximum parsimony is model-free and can point to the overall least-change hypothesis, but it may be misled if character change is very non-random (which high homoplasy suggests). Its bootstrap supports are a simple way to gauge robustness, but for data with dependencies, they can be underestimated. NeighborNet is exploratory and does not provide clade support values; it excels at visualizing conflict but can be hard to interpret quantitatively beyond δ and Q. In any case, both methods highlight that low clade support in the tree corresponds to high reticulation in the network. This finding echoes the concern raised by Hartmann & Walkden that shared syntax may often reflect parallel contact rather than shared ancestry.

A potential discrepancy is that Hartmann & Walkden focused on Bayesian inference (stochastic Dollo model) and argued that their results were *minimally biased*. Our deterministic parsimony gave a similar outcome, which suggests robustness of the conclusion. Neither method produced any strong contradictory result. One might worry that rooting on an isolate (Laal) could affect deep splits; however, our outgroup choice seems not to have artificially merged otherwise distinct families. In sum, the methodological consistency strengthens the conclusion that the SSWL syntactic features alone carry limited phylogenetic signal.

Conclusion

In replicating Hartmann & Walkden (2024), we find strong concordance: the 110-language SSWL syntactic dataset yields no strongly supported deep phylogenetic tree under either parsimony or NeighborNet analysis. The MP tree (Supplementary Materials) mostly resolves only recent subgroupings (Germanic, Uralic, etc.), and even those have moderate bootstrap support (Fig. 2). The majority consensus is highly unresolved. The NeighborNet graph is correspondingly highly reticulate, with nonzero delta and Q scores indicating substantial conflict. These outcomes confirm that the syntactic data, as is, do not encode clear higher-order genealogical relationships. This suggests caution in relying on raw SSWL syntax for deep phylogeny: much of the signal is obscured by convergent or contact-induced similarity.

Our study highlights the importance of method choice: both tree and network analyses should be used in tandem. Future work might explore how to improve phylogenetic signal in syntax. One direction is to incorporate covariation models or admixture networks that explicitly model contact. Combining syntax with lexical/phonological data could also boost resolution. Ultimately, understanding when and why syntactic characters fail to preserve deep inheritance (as seen here) remains an open question in linguistic phylogeny.

References

Bryant, D. & Moulton, V. (2004). Neighbor-Net: An agglomerative method for the construction of planar phylogenetic networks. In Algorithms in Bioinformatics (WABI 2002) (LNCS 2452:375–391).

Dunn, M., et al. (2005). Structural phylogenetics and the reconstruction of ancient language history. Science 309:2072–2075.

Gray, R. D., S. J. Greenhill & Y. Zhang (2010). The shape and fabric of human history. Philosophical Transactions of the Royal Society B 365:3923–3933 (neighbor-net, delta/Q measures).

Hartmann, F. & G. Walkden (2024). The strength of the phylogenetic signal in syntactic data. Glossa: a journal of general linguistics 9(1):10598. doi:10.16995/glossa.10598.

Holland, B. R., D. Moulton & P. J. Sanderson (2002). Estimating the phylogenetic signal in binary characters. In Mathematical Phylogenetics (2006) – see Gray et al. (2010) discussion.

Huson, D. H. & D. Bryant (2006). Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23:254–267 (splits graph methods).

James S. Farris, The Retention Index and Homoplasy Excess, Systematic Biology, Volume 38, Issue 4, December 1989, Pages 406–407, https://doi.org/10.2307/2992406

Koopman, Hilda. (2012–). Syntactic Structures of the World’s Languages (SSWL) database. https://terraling.com/groups/ Accessed on 5th May 2025

Pant, S., Kumar, A., Ram, M., Klochkov, Y., & Sharma, H. K. (2022). Consistency Indices in Analytic Hierarchy Process: A Review. Mathematics, 10(8), 1206. https://doi.org/10.3390/math10081206

Shu, Kevin & Aziz, Sharjeel & Huynh, Vy-Luan & Warrick, David & Marcolli, Matilde. 2018. Syntactic phylogenetic trees. In Kouneiher, Joseph (ed.), Foundations of mathematics and physics one century after Hilbert, 417–441. Amsterdam: Springer. DOI: http://doi.org/10.1007/978-3-319-64813-2_14

Wichmann, S., & Saunders, A. (2007). How to use typological databases in historical linguistic research. Diachronica, 24(2), 373–404. https://doi.org/10.1075/dia.24.2.06wic

Supplementary Materials

The R script, Maximum Parsimony Trees, bootstrapped trees, table for delta and Q-scores, and nexus filescan be found in this github repository. Note that the maximum parsimony trees could not be included due to their sizes.