Science
AAAS

Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss.

Authors:

Skirgård, H.[1,2,3,4]*†, Haynie, H. J.[5]†, Blasi, D. E.[6,1,7,8]†, Hammarström, H.[9,4]†, Collins, J.[10,4], Latarche, J. J.[11], Lesage, J.[4,1,12,13,14], Weber, T.[15,1], Witzlack-Makarevich, A.[16], Passmore, S.[17,18,19], Chira, A.[1], Maurits, L.[20], Dinnage, R.[21], Dunn, M.[9,4], Reesink, G.[22], Singer, R.[2,23], Bowern, C.[24], Epps, P.[25], Hill, J.[26]‡, Vesakoski, O.[27,28], Robbeets, M.[29], Abbas, N. K.[11], Auer, D.[1], Bakker, N. A.[15,1], Barbos, G.[11], Borges, R. D.[30], Danielsen, S.[31,32,33], Dorenbusch, L.[1,34], Dorn, E.[11], Elliott, J.[35], Falcone, G.[9], Fischer, J.[15,1], Ghanggo Ate, Y.[36,37], Gibson, H.[38], Göbel, H.-P.[15,1,39], Goodall, J. A.[11], Gruner, V.[1], Harvey, A.[40], Hayes, R.[11], Heer, L.[15], Herrera Miranda, R. E.[41,42,34,13], Hübler, N.[1,15], Huntington-Rainey, B. H.[11], Ivani, J. K.[43], Johns, M.[15,1], Just, E.[43], Kashima, E.[2,3], Kipf, C.[15,1], Klingenberg, J. V.[15,1], König, N.[15], Koti, A.[9], Kowalik, R. G. A.[44], Krasnoukhova, O.[45,46], Lindvall, N. L.[9], Lorenzen, M.[15,1], Lutzenberger, H.[10,47], Martins, T. R.[11], Mata German, C.[11], van der Meer, S.[4], Montoya Samamé, J.[48], Müller, M.[1], Muradoglu, S.[2], Neely, K.[25], Nickel, J.[15,1], Norvik, M.[49,50], Oluoch, C. A.[15,1], Peacock, J.[10,4], Pearey, I. O.[11], Peck, N.[2,51], Petit, S.[11], Pieper, S.[15,1], Poblete, M.[48,52], Prestipino, D.[2], Raabe, L.[15,1], Raja, A.[11], Reimringer, J.[1], Rey, S. C.[11], Rizaew, J.[11], Ruppert, E.[53], Salmon, K. K.[1], Sammet, J.[15,1], Schembri, R.[2,54], Schlabbach, L.[15,1], Schmidt, F. W.[55], Skilton, A.[56], Smith, W. D.[25], de Sousa, H.[4,57], Sverredal, K.[9], Valle, D.[58], Vera, J.[48], Voß, J.[15,1], Witte, T.[15,1], Wu, H.[2], Yam, S.[2,59], Ye 葉婧婷, J.[60,1], Yong, M.[11], Yuditha, T.[61,62], Zariquiey, R.[48,1], Forkel, R.[1], Evans, N.[2,3], Levinson, S. C.[4], Haspelmath, M.[1], Greenhill, S. J.[63], Atkinson, Q. D.[64], Gray, R. D.[1,64]*

Affiliations:

[1] Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology; Leipzig, Germany.
[2] ARC Centre of Excellence for the Dynamics of Language, College of Asia and the Pacific, Australian National University; Canberra, Australia.
[3] Department of Linguistics, School of Culture, History and Language, College of Asia and the Pacific, Australian National University; Canberra, Australia.
[4] Department of Language and Cognition, Max Planck Institute for Psycholinguistics; Nijmegen, the Netherlands
[5] Department of Linguistics, University of Colorado Boulder; Boulder, United States of America.
[6] Department of Human Evolutionary Biology, Harvard University; Cambridge, United States of America.
[7] Linguistic Convergence Laboratory, School of Linguistics, Faculty of Humanities, Higher School of Economics University; Moscow, Russia
[8] Human Relation Area Files, Yale University; New Haven, United States of America.
[9] Department of Linguistics and Philology, Uppsala University; Uppsala, Sweden.
[10] Department of Linguistics, Faculty of Arts, Radboud University; Nijmegen, The Netherlands.
[11] Department of Linguistics, School of Languages, Cultures and Linguistics, School of Oriental and African Studies (SOAS), University of London; London, The United Kingdom.
[12] Langage, langues et cultures d'Afrique (LLACAN), Centre national de la Recherche Scientifique (CNRS); Villejuif, France.
[13] Institut national des langues et civilisations orientales (INALCO); Paris, France.
[14] Department of Asian and African Studies, Humboldt-Universität zu Berlin; Berlin, Germany.
[15] Institute for Scandinavian Studies, Frisian and General Linguistics, Department of General Linguistics, Christian-Albrechts-Universität zu Kiel; Kiel, Germany.
[16] Department of Linguistics, Faculty of Humanities, The Hebrew University of Jerusalem; Jerusalem, Israel
[17] Evolution of Cultural Diversity Initiative, School of Culture, History and Language, College of Asia and the Pacific, The Australian National University, Canberra, Australian Capital Territory, Australia
[18] Faculty of Environment and Information Studies, Keio University SFC (Shonan Fujisawa Campus); Tokyo, Japan
[19] Department of Anthropology and Archaeology, Faculty of Arts, University of Bristol; Bristol, The United Kingdom.

[20] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology; Leipzig: Germany.
[21] Institute of Environment, Department of Biological Sciences, Florida International University, Miami, FL, United States of America.
[22] Houten
[23] Research Unit for Indigenous Language, School of Languages and Linguistics, University of Melbourne; Melbourne, Australia.
[24] Department of Linguistics, Yale University; New Haven, United States of America.
[25] Department of Linguistics, University of Texas at Austin; Austin, United States of America.
[26] School of Anthropology, University of Arizona; Tucson, United States of America.
[27] Department of Biology, Turku University; Turku, Finland.
[28] Department of Finnish and Finno-Ugric languages, University of Turku, Turku, Finland
[29] Department of Archaeology, Max Planck Institute for the Science of Human History; Jena, Germany.
[30] Institute of Slavic Studies, Polish Academy of Sciences; Warsaw, Poland.
[31] Zentrum für kleine und regionale Sprachen, Friesisches Seminar, Europa-Universität Flensburg; Flensburg, Germany
[32] Centro de Investigaciones Históricas y Antropológicas (CIHA); Santa Cruz de la Sierra, Bolivia.
[33] Europa-Universität Flensburg (EUF)
[34] Institute of Linguistics, Leipzig University; Leipzig, Germany.
[35] Department of Linguistics, University of Hawai'i at Mānoa; Honolulu, United States of America,
[36] Schoolf of Culturel, History and Language, College of Asia and the Pacific, Australian National University; Canberra, Australia.
[37] Universitas Katolik Weetebula; Sumba Island, Indonesia
[38] Department of Languages and Linguistics, University of Essex; Essex, The United Kingdom.
[39] Department of Linguistics, University of Cologne; Cologne, Germany.
[40] Faculty of Languages and Literatures, University of Bayreuth; Bayreuth, Germany.
[41] Structure et Dynamique des Langues (SeDyl), Centre national de la recherche scientifique (CNRS); Villejuif, France.
[42] Sprachwissenschaftliches Seminar, Georg-August-Universität Göttingen; Göttingen, Germany.
[43] Department of Comparative Linguistics, University of Zürich; Zürich, Switzerland
[44] Department of Linguistics, Stockholm University; Stockholm, Sweden.
[45] Centre for Linguistics, Leiden University; Leiden, The Netherlands
[46] Department of Linguistics, University of Antwerpen; Antwerpen, Belgium.
[47] Department of English Language and Linguistics, University of Birmingham; Birmingham, United Kingdom
[48] Facultad de Letras y Ciencias Humanas, Pontificia Universidad Católica del Perú, Lima, Perú
[49] Institute of Estonian and General Linguistics, University of Tartu, Tartu, Estonia
[50] Department of Modern Languages, Uppsala University; Uppsala, Sweden.
[51] University of Freiburg; Freiburg, Germany
[52] Universidad de Chile; Santiago, Chile.
[53] Department of Linguistics, Quantitative Lexicology and Variational Linguistics (QLVL), KU Leuven; Leuven, Belgium.
[54] Division of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia
[55] Department of Social Anthropology, University of Cambridge; Cambridge, The United Kingdom.
[56] Department of Linguistics, Cornell University; Ithaca, New York, United States of America
[57] Centre de recherches linguistiques sur l'Asie orientale (CRLAO), École des hautes études en sciences sociales (EHESS; Aubervilliers, France
[58] Department of Modern Languages, University of Mississippi; Oxford, United States of America.
[59] Institute for General Linguistics, Westfälische Wilhelms-Universität Münster; Münster, Germany
[60] Department of Chinese Language and Literature, Fudan University, Shanghai, China
[61] Department of Spanish, Linguistics, and Theory of Literature (Linguistics) Faculty of Philology. University of Seville, Seville, Spain
[62] Department of Languages, Faculty of Education, Atma Jaya Catholic University; Jakarta, Indonesia
[63] Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History; Jena, Germany.
[64] School of Psychology, University of Auckland; Auckland, New Zealand

Notes:

*Corresponding author

†Equal first authors

‡Deceased.

Abstract: While global patterns of human genetic diversity are increasingly well characterized, the diversity of human languages remains less systematically described. Here we outline the Grambank database. With over 400,000 data points and 2,400 languages, Grambank is the largest comparative grammatical database available. The comprehensiveness of Grambank allows us to quantify the relative effects of genealogical inheritance and geographic proximity on the structural diversity of the world's languages, evaluate constraints on linguistic diversity, and identify the world's most unusual languages. An analysis of the consequences of language loss reveals that the reduction in diversity will be strikingly uneven across the major linguistic regions of the world. Without sustained efforts to document and revitalize endangered languages, our linguistic window into human history, cognition and culture will be seriously fragmented.

One-Sentence Summary: We use Grambank to quantify the effects of genealogy and geography on linguistic diversity, evaluate constraints on this diversity, identify the world's most unusual languages, and highlight the impact of language loss.

There are approximately 7,000 spoken languages in the world (*1*). These languages vary widely in their structural properties. They vary by the order in which they arrange words and the constructions they use to combine segments in higher-order units. They can also differ markedly in how information is grammatically expressed. Some languages always mark categories such as gender, number, case and tense, while some never or only optionally mark these categories. Furthermore, sentences that consist of many words in some languages can be translated by a single word in other languages, while the preferred word order varies widely. This linguistic diversity is not randomly distributed. We expect it to be shaped by human cognition (*2*, *3*), geographical proximity (*4*, *5*) and genealogical descent (*6*, *7*). However, an accurate understanding of the actual structural diversity of languages, the factors that shape that variation, and what is at stake when the world loses languages has been hampered by the lack of accessible, systematically sampled, global data. For example, the World Atlas of Language Structures (WALS, *8*) has incomplete genealogical coverage (*9*), and 84% missing data (see Fig. S1).

Here we introduce Grambank - a systematic sample of the structural diversity of the world's languages. Grambank is designed to be used to investigate the global distribution of features, language universals, functional dependencies, language prehistory and interactions between language, cognition, culture and environment. The Grambank database currently covers 2,467 language varieties, capturing a wide range of grammatical phenomena in 195 features, from word order to verbal tense, nominal plurals, and many other well-studied comparative linguistic variables. The dataset includes both varieties classified as "languages" and "dialects" (70 dialects representing 46 languages, resulting in a total of 2,430 unique languages, *1*). The coverage spans 215 different language families and 101 isolates from all inhabited continents and geographic regions (see Fig. S2).

Languages are important to cultural identity, health, the preservation of traditional knowledge and institutions, and as a unique window into human history, culture and cognition (*10–12*). However, languages are vanishing at a rate that rivals our biodiversity crisis (*13*, *14*). It is estimated that without intervention approximately one language will be lost every month in the next 40 years (*15*). This tragic situation and its detrimental consequences has prompted the United Nations to recently announce the UN Decade of Indigenous Languages (*16*). The Grambank dataset is uniquely positioned to showcase the diversity of the world's languages and the knowledge that we are currently in danger of losing.

Here we use the Grambank data to answer four long-standing questions about global linguistic diversity that have previously been difficult to answer in a rigorous quantitative manner. What are the relative roles of genealogical inheritance and geographical diffusion in shaping grammatical diversity? How constrained is grammatical evolution? What are the world's most unusual languages, and what will the consequences of language loss be on our understanding of linguistic diversity?

Genealogy versus geography

One of the oldest debates in the field of linguistics concerns the relative roles of genealogical inheritance and geographical diffusion in shaping patterns of linguistic diversity. Proponents of the tree model of linguistic relationships dating back to at least Schleicher in the 1800s have claimed that nested patterns of inherited linguistic features show that genealogy trumps geographic diffusion (*17*). In contrast, defenders of the "wave model" developed by Schmidt (*18*) have argued that cross-cutting patterns of features reflect waves of linguistic diffusion.

Considerable dispute still exists today about the relative importance of genealogy versus geography for explaining variation in the grammatical features of the world's languages (*19*). Nichols (*20*) has claimed that while features such as a distinction between inclusive and exclusive pronouns are genealogically stable, others such as word order are consistent with primarily geographic influences. Campbell (*21*) has questioned whether genealogical signals can be reliably identified in the structural characteristics of languages, given the potential influences of geographic diffusion, homoplasy, and cognitive constraints on these features. Another dimension of this debate focuses on the temporal depth of genealogical and geographic signals in grammar. Dunn et al. (*22*) propose that structural features of language may bear the signals of deep genealogical relationships in Island Melanesia. Matsumae et al. (*23*) find an association between the variation in grammatical structures and genetic variation in northeast Asia that further supports the idea that structural features reflect deep relationships between populations. Ultimately the dynamics of grammatical feature evolution may be complex, with a small set of features showing stability on language phylogenies and a large number evolving rapidly and showing bursts of contact-related change (*24*).

To go beyond qualitative impressions and *a priori* commitments to either genealogical inheritance or geographic diffusion as the primary factor shaping grammatical diversity, we estimated the magnitude of spatial and phylogenetic effects jointly using approximate Bayesian Inference for Latent Gaussian Models (*25*). We used a Maximum Clade Credibility Tree from a recent Bayesian phylogenetic analysis of all extant languages (*26*) to represent language history. Spatial relations were derived from the language locations documented in Glottolog (*1*). While the effect of phylogeny varies dramatically between Grambank features, ranging from very strong (0.98) to almost non-existent (<0.01), overall it is consistently greater than that of space

(mean phylogeny = 0.72, standard deviation = 0.26 vs. mean space = 0.03, standard deviation = 0.06; see Table S1). Figures S3-5 illustrate the features with the strongest phylogenetic signal in a tree-plot with ancestral state reconstruction and Figures S6-8 are maps showing the features with the strongest spatial signal. The feature with the strongest phylogenetic signal (0.98) was GB133: "Is a pragmatically unmarked constituent order verb-final for transitive clauses?". The feature with the lowest phylogenetic signal (<0.01) was GB129: "Is there a notably small number, i.e. about 100 or less, of verb roots in the language?". We note that the strong phylogenetic effects should be interpreted with the caveat that it can be difficult to estimate the independent effects of space and phylogeny because language diversification is itself a spatial process (and indeed the global phylogeny (*26*) was informed by language location). However, only the global phylogeny captures information on established ancestral relationships between languages. The fact that the phylogeny so consistently and decisively outperforms space as a predictor suggests that the modern patterns of linguistic diversity are shaped by genealogical inheritance more than geographical diffusion.

The relative influences of genealogy and geography may not be uniform across different elements of grammar, however. Linguists (*27*, *28*) have suggested that language contact may have different outcomes for the verbal, pronominal and nominal domains of grammar in contact languages. Grambank features cover many different domains of grammar (e.g. clausal, nominal, pronominal and verbal), and thus enable us to test the generality of this claim. Interestingly, we do not find statistical differences across domains in terms of spatial or phylogenetic effects (see Fig. 1 and Table S2). Nichols (*20*) makes more specific claims about the areal diffusibility vs. phylogenetic inheritance of specific grammatical features in language change in non-contact languages. We matched her predictions with features in Grambank and their respective spatial

and phylogenetic effects. We do find support for several features she predicted to show strong phylogenetic effects, however the same is not true for those predicted to be areal (see Fig. S9).

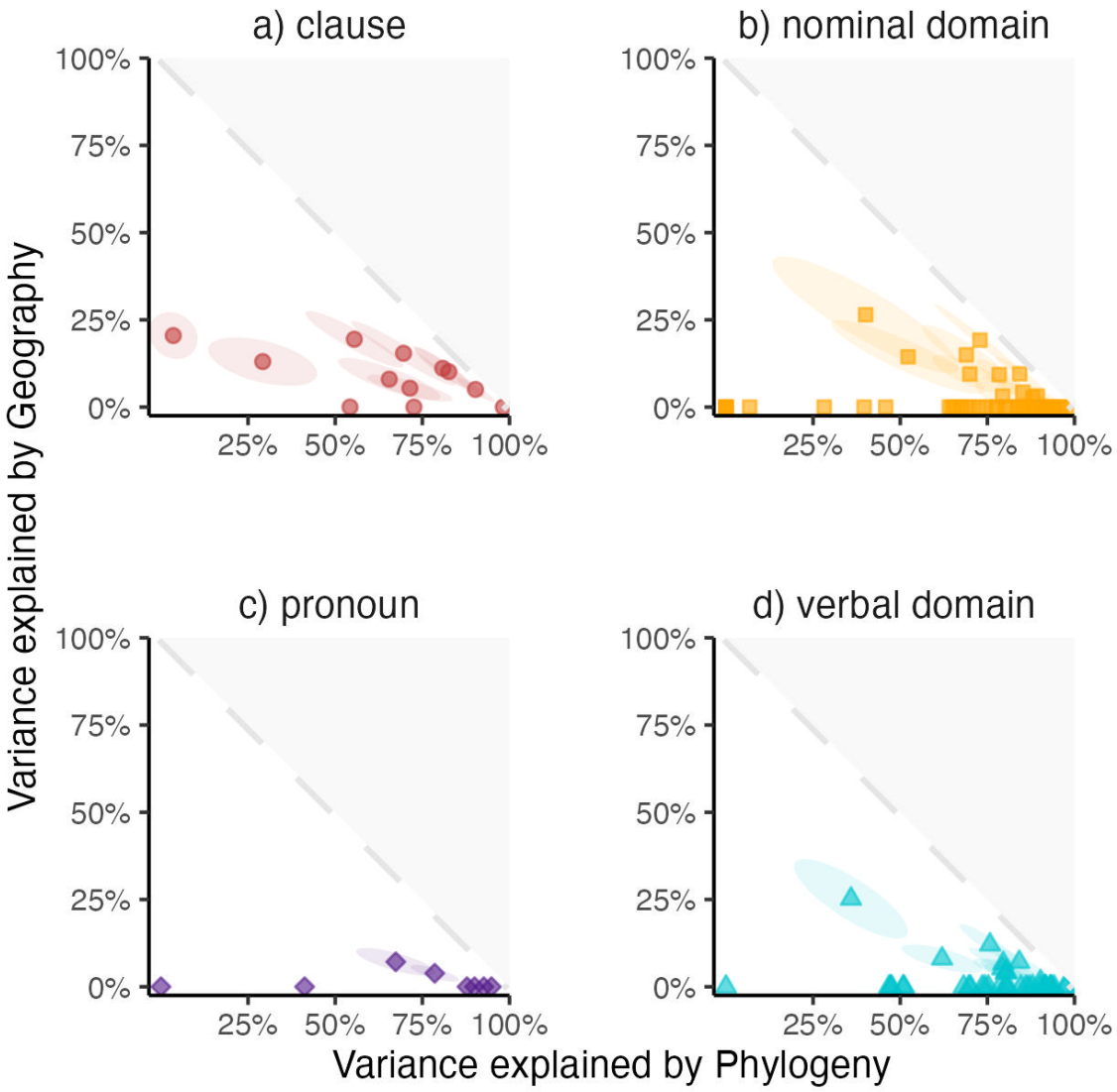

Figure 1: Variance explained by phylogeny and geography. Each point is a Grambank feature. The panels represent different domains of grammar that the features are associated with: a = clausal, b = nominal domain, c = pronominal domain and d = verbal domain. A high value indicates that a large part of the variance is explained by either space (y axis) or phylogeny (x axis). The ellipses represent the standard deviation of the joint posterior, tilted for the covariance.

Constraints on grammar

The Grambank dataset focuses on 195 core grammatical features (see Table S4). Even this basic set of features represents an astronomical number ($>10^{34}$) of possible grammars - the possible "design space" (sensu Dennett, *29*). How constrained is the distribution of the world's actual realized grammars within this total design space and what are the most important axes of variation? Some have claimed that languages are tightly constrained systems - "*un système où tout se tient"* (*a system where everything fits together, 30*). Many generative linguists assert that human cognition imposes strong constraints on grammatical variation such that only a small number of underlying factors are required to explain the observed diversity (*31–33*). In contrast, others have argued that distinct components of language can vary individually - "*All parts of a language appear in principle to be independently mobile*" (*34*). Grambank's broad suite of logically independent traits (see Supplementary Material 1:1), systematically coded across a global sample of languages (see Supplementary Material 1:6), makes it an ideal resource for exploring these claims.

We use Principal Component Analysis (PCA) to reduce the dimensions of the Grambank data to a set of orthogonal variables representing the underlying patterns of variation among the grammatical features we consider (see Supplementary Material 1:9). A non-graphical Cattel's Scree test (*35*) shows that the optimal number of components is 19, explaining 49% of the variation among grammars. The first three components returned by the PCA capture only 21% of the variation (9%, 7%, and 5%, respectively). These results can be compared to similar studies of musical and genetic variation. A recent analysis of cross-cultural musical behavior found that

only three components optimally described the variation (*36*). In contrast, an analysis of human genetic variation across Europe in the form of single nucleotide polymorphisms (SNPs) found that the first and second principal components explained under 1% of the variation (0.3% and 0.15%, respectively, *37*). This indicates that language structures have greater combinatorial flexibility than musical behavior, but far less than genetic evolution. Grammatical systems are thus neither tightly constrained nor entirely free to vary.

Having eliminated nearly all strict logical dependencies from our dataset (see Supplementary Material 1:1), the sizable fraction of grammatical variation that is explained by a limited set of dimensions could reflect functional or historical constraints on grammar. However, even our broader set of 19 principal components still leaves more than half of the variation unexplained, suggesting there is also a high degree of flexibility in grammatical structures, rather than tight constraints determined by a small number of underlying factors.

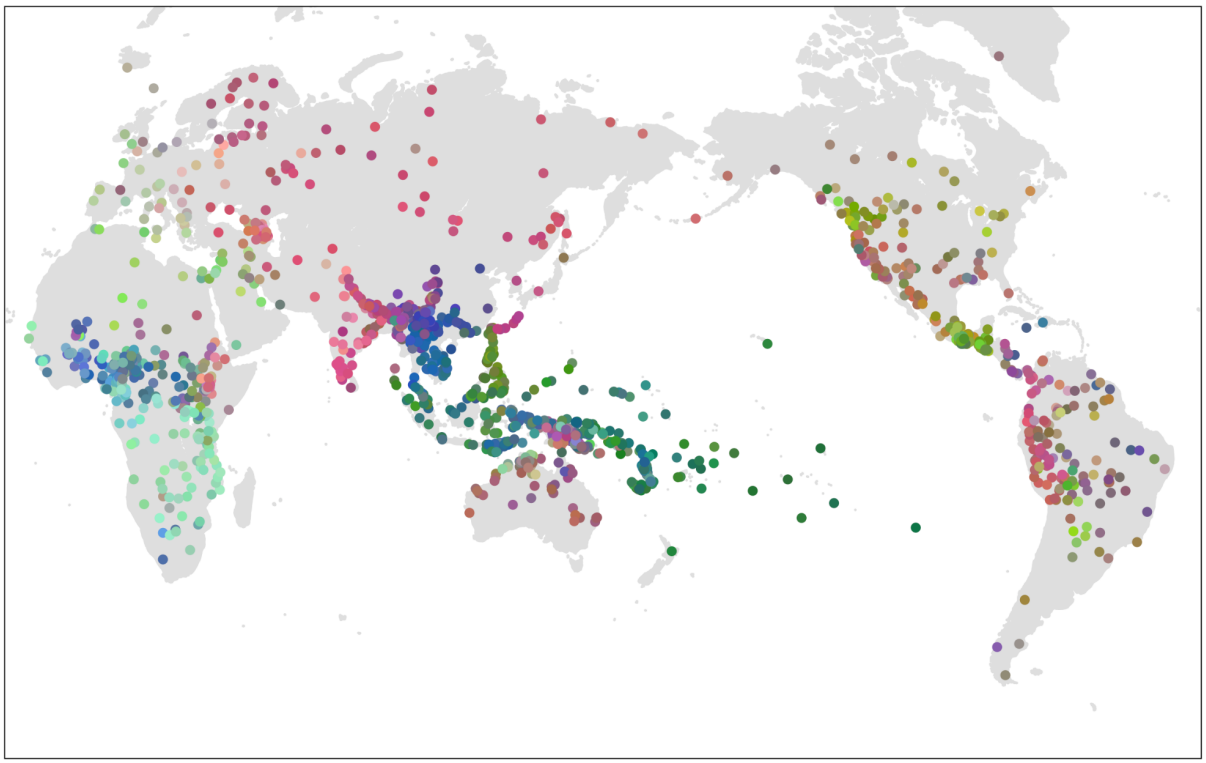

Figure 2. Grammatical similarity in the Grambank sample of languages. The color coding represents the distribution of languages according to the first three principal components of a Principal Component Analysis mapped onto RGB color space (PC1 = Red, PC2 = Green and PC3 = Blue). Similarity in color indicates similarity in grammatical structure on the first three dimensions. See Fig. S14 for loading of Grambank features on these three components.

It is possible that the principal components we infer are simply clusters of traits that are associated due to shared phylogenetic history, rather than functional constraints on these linguistic systems. In order to establish whether they correspond to meaningful aspects of design space, we compared these data-driven dimensions to metrics we developed to capture factors linguists have commonly used to describe grammatical variation. The metrics were word order (*38*, *39*), locus of marking (the degree to which a language mainly features head or dependent marking, as described by *40*), morphological typology (expression by phonologically fused vs.

freestanding morphemes, which we call “fusion” (*41*) - not to be confused with Sapir’s notion of “fusional” languages (*42*)), and flexivity (degree of allomorphic variation, as described by *41*). In addition, we calculated an index for use of noun class/gender to further probe this important component of the flexivity score (Supplementary Material 1:9 and Table S6). We found PC1 correlated most strongly with features capturing fusion, while PC2 correlated most strongly with noun class/gender features (see Fig. S15). PC3 did not show a clear strong association with any of the metrics (see Table S3). Hence, while much of the variation in our data falls outside of these constructs, our analysis indicates that at least the first two dimensions of variation in the world’s grammars do have a clear linguistic interpretation, corresponding to the extent to which languages combine elements through ‘fusion’ and use noun class/gender.

Next, we use these dimensions to examine how history constrains the evolution of languages through this design space. Fig. 2 plots the location of the languages in our sample colored according to the first three principal components. Consistent with our spatiophylogenetic analysis above, this reveals macro-scale spatial patterns around the globe that appear to mirror the distribution of some major language families. For example, most Austronesian languages in the Pacific are colored green, while the Bantu languages in sub-Saharan Africa share a bright turquoise. To examine the connection between history and design space more closely, we map the 15 largest language families in the world onto plots of design space defined by the first two PCs (Fig. 3). Language families such as Austronesian, Nuclear Trans-New Guinea and Dravidian are tightly packed together, suggesting strong phylogenetic inertia in this part of the design space. However, other families like Afro-Asiatic or Indo-European are more spread out in the Grambank design space, demonstrating high within-family diversity in these dimensions. Within Indo-European, for example, there are two clusters largely corresponding to contact languages

and non-contact languages (see Fig. S16). The Austroasiatic language family also shows two distinct clusters: languages of the Munda sub-branch and the rest of the family (see Fig. S17). Language families, then, can be both distinct and diverse samples from the design space.

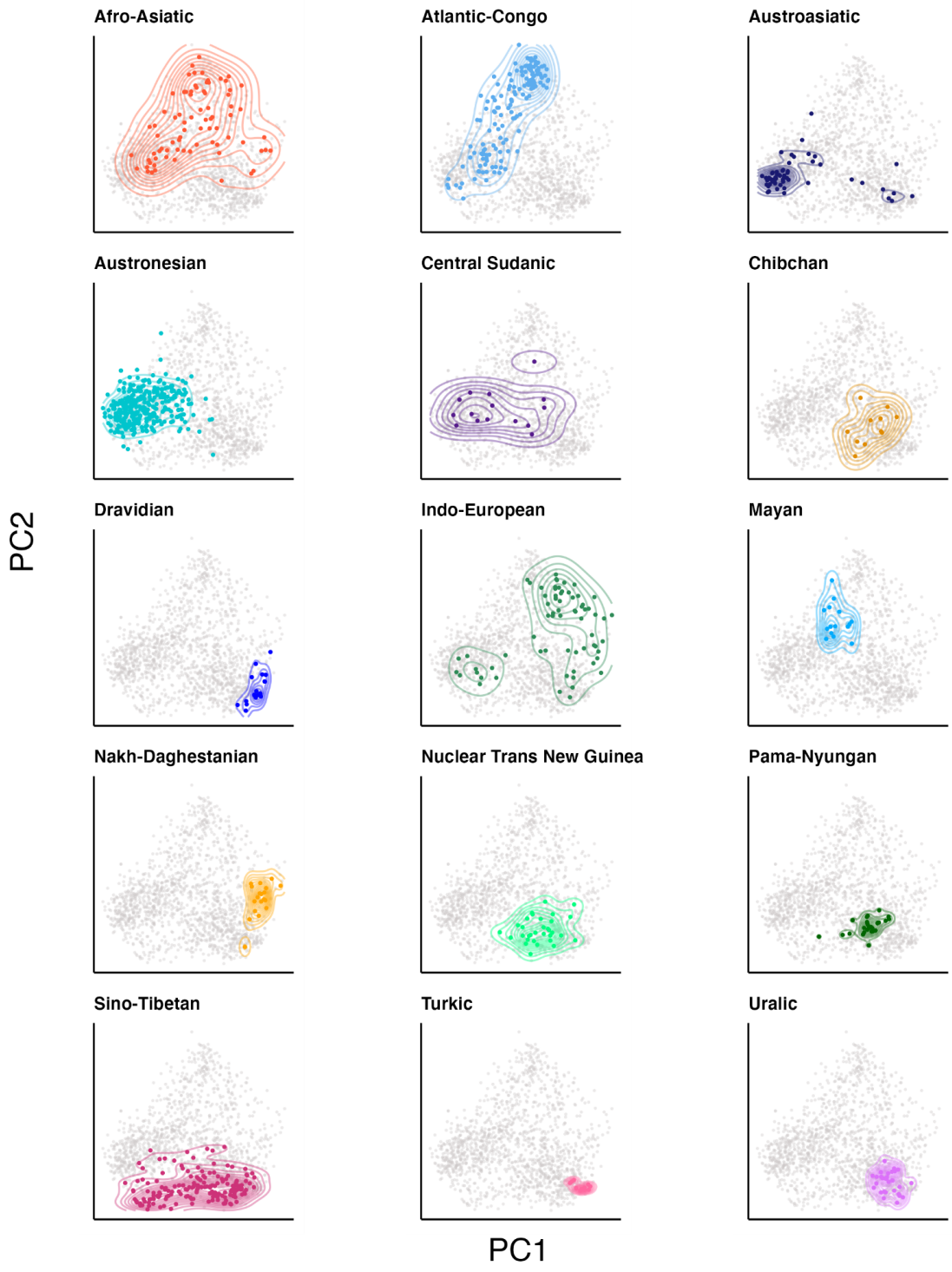

Figure 3. Distribution of the 12 largest families in our dataset in Grambank design space. The x-axis represents the first principal components and the y-axis the second. All languages are plotted, and for each facet one family is highlighted in a different color. Austronesian languages, which are known for lacking gender and having little morphology, are found on the far left.

This mix of both distinctness and diversity within families raises the question: "Is the evolution of the world's languages through this grammatical design space determined by a set of universal and enduring design constraints, or is the process historically contingent, canalized by culturally evolved, inherently unpredictable and lineage-specific basins of attraction?" For example, we find few languages overall in the upper left corner of Fig. 3, where we would expect (given the loadings on the PCA) languages with little morphology but robust noun class/gender systems. This question about constraints parallels Stephen J. Gould's work exploring the role of historical contingency in biological evolution (*43*). Gould asks, if we were to "replay the tape of life" over and over again, what patterns of current diversity would reliably recur (reflecting universal constraints) vs. never evolve again (reflecting historical contingency)? While Gould laments no such experiment exists in the natural world, the evolution of the world's languages does contain a natural experiment of this kind. The current linguistic diversity of the Americas has emerged over the last 15-30kya, essentially 'replaying the tape' of language evolution from a small number of founder lineages.

To answer Gould's question, we computed pairwise cultural fixation scores (*44*) based on the Grambank data for languages of the world divided into 24 linguistic areas (8 in the Americas and 16 elsewhere) (*45*). Cultural fixation scores are preferable to raw (Gower) distances because they take into account feature prevalence and inter- as well as intra-group variation. A low cultural

fixation score indicates a close affinity, and a high score indicates greater differentiation. These pairwise scores can be visualized in a network (Fig. S20), and a modularity score can be calculated to assess the relative independence of network components (see Table S7). The low fixation scores between some areas in the Americas reflect shared history, but the negative modularity of the American component of this network (-0.061) indicates that the Americas do not form a separate community cluster from the rest of the world (see Fig. S20).

These findings suggest that while history clearly matters a lot for explaining global language diversity, there nevertheless appear to be some enduring constraints that shape the cultural evolution of languages over many thousands of years towards predictable regions of grammatical design space.

Unusual languages

Our understanding of how languages work as systems is strongly informed by the cross-linguistic frequency of grammatical features and their combinations. Prolific language groups (such as the Austronesian or Atlantic-Congo families), as well as functional pressures (e.g. the tendency towards harmonic word orders), drive the overall prevalence of certain features and combinations of features. Languages with uncommon features or combinations of features are informative for the study of language because they show the limits of what is possible. They can also represent rare survivors of deep linguistic lineages.

We investigate unusual combinations of grammatical features by introducing a metric – "unusualness" – that generalizes the notion of cross-linguistic frequency from individual features or combinations of features to entire grammars (see Supplementary Material 1:11). According to

our metric, a language is more unusual than another if (a) some of its features and/or (b) some of its combinations of features are more infrequent, comparatively speaking. It should be stressed that this operationalization of unusualness is necessarily restricted to the features present in Grambank - in other words, we make no claims about the unusualness of languages with respect to linguistic features not covered in the database.

The global distribution of unusualness is richly structured (Fig. S22). The most unusual languages are most often not members of the largest language families, or if they are, they are found at the geographic periphery of their expansion. In particular, several of the most unusual languages are isolates with no known connection to any established language family (e.g. Movima [movi1243], Kuot [kuot1243], Hadza [hadz1240], Yélî Dnye [yele1255]). Isolates represent 4% of Grambank's languages in total, but they make up 19% of the most unusual languages. In addition, the distribution of grammatical unusualness displays areal patterning beyond language families, with cultural and historical regions revealing consistent values of unusualness from low (Southeast Asia), mid (southern Africa) to high (Northern Africa and Europe) - see Fig. S23.

To assess the accuracy of these inferences, we built a model to predict unusualness based on language families and cultural-historic regions (see Table S8). The model performs well (Bayesian R^2=0.75, see Fig. S24 and Table S8), which suggests that language families and regions are strongly predictive of a given language's unusualness. In other words, historical factors that have driven regional patterns of lineage loss, such as the expansion of language families and colonial empires, are likely to have been more important in structuring patterns of unusualness than general constraints on grammar.

The existence of unusual languages should not overshadow the fact that all languages in our sample are typically very different from each other. Very few pairs of languages share the same Grambank description (only five; see Manhattan distances in Fig. S25). Given that these descriptions are centered on core grammatical features (i.e. where languages are more likely to be effectively compared), this entails that each and every language enshrines a unique and irreplaceable source of linguistic knowledge. Thus, in addition to the social and humanitarian consequences (*10*, *11*), each endangered language poses a threat to the understanding of language at large.

Language loss

We investigate the potential loss of linguistic knowledge using contemporary estimates of language endangerment and a new way of quantifying language diversity. Our goal is to provide a bird's-eye view of this at both global and regional levels. With this in mind, we applied a metric that is used in ecology termed "Functional Richness" (*46*, *47*). This metric quantifies the area occupied by a species (languages in our study) in an abstract multidimensional space defined by a set of features and estimates the diversity the data represents. By computing this metric with all languages, and then only with those that are not endangered, we can estimate the potential loss in structural diversity (*48*). We calculated functional richness globally and for each region (*45*) (see Supplementary Material 1:13). This allows us to estimate what we will lose collectively if these languages disappear. We found that, although functional richness declines only moderately on a global scale with the loss of languages that are under threat, the consequences of language loss vary dramatically across regions (Fig. 4). Regions like Northeast South America, Alaska-Oregon and Northern Australia will be dramatically impacted because *all*

indigenous languages there are under threat, and so the functional richness that would remain is 0. The pronounced reduction of nearly half the functional space occupied by languages, even in regions with many non-threatened languages (e.g. Oceania, North Coast New Guinea, Greater Abyssinia, Greater Mesopotamia), will undermine our ability to investigate the basic structures of language and the diverse expressions used to encode them.

Conclusion

The adoption of standard linguistic data formats, such as CLDF (*49*), and the open availability of carefully curated global databases, such as Grambank, open up the possibility of quantitative cross-linguistic comparison on a scale that was not previously possible. Our analyses have demonstrated the importance of genealogy in shaping grammatical diversity, revealed the influence of both historical contingency and universal constraints in shaping grammatical design, and highlighted the imminent threat posed by language endangerment. Grambank data should facilitate more rigorous testing of claims about language universals, linguistic areas and the factors that drive the evolution of linguistic disparity. Because linguistic diversity has been found to be associated with a broad array of cultural and biological traits, ranging from religious beliefs to economic behavior, musical traditions and genetic lineages, the impact of these developments could extend beyond the field of linguistics. We hope that these links with other facets of human behavior will help make Grambank a key resource in the multidisciplinary endeavor that is understanding human diversity.

Figure 4. Decline of functional richness associated with language loss. At top, bars representing Functional Richness relative to the current diversity of the world's languages are shown in light green,

and Functional Richness of non-threatened languages in the same areas are shown in dark green. Functional Richness declines in all areas, with some regions showing dramatic decreases. At bottom, threatened (gray) and non-threatened (black) languages are plotted over a convex hull (green) that represents the overall area of functional space (x and y, representing two dimensions of a PCoA on the Grambank feature set) occupied by languages of the area.

References and Notes

1. H. Hammarström, R. Forkel, M. Haspelmath, S. Bank, *Glottolog 4.4* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2021; https://glottolog.org/).
2. M. H. Christiansen, N. Chater, Language as shaped by the brain. *Behav. Brain Sci.* **31**, 489–509 (2008).
3. B. Bickel, A. Witzlack-Makarevich, K. K. Choudhary, M. Schlesewsky, I. Bornkessel-Schlesewsky, The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLOS ONE*. **10**, e0132819 (2015).
4. J. Nichols, *Linguistic Diversity in Space and Time* (University of Chicago Press, Chicago, 1992).
5. P. Muysken, *From Linguistic Areas to Areal Linguistics* (John Benjamins, Amsterdam, 2008).
6. D. Dediu, S. C. Levinson, Abstract Profiles of Structural Stability Point to Universal Tendencies, Family-Specific Factors, and Ancient Connections between Languages. *PLOS ONE*. **7**, e45198 (2012).
7. D. Dediu, M. Cysouw, Some structural aspects of language are more stable than others: A comparison of seven methods. *PLOS ONE*. **8**, e55009 (2013).
8. M. S. Dryer, M. Haspelmath, Eds., *The World Atlas of Language Structures Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013; http://wals.info).
9. H. Hammarström, Sampling and genealogical coverage in WALS. *Linguist. Typology*. **13**, 105–119 (2009).
10. M. D. Auger, Cultural Continuity as a Determinant of Indigenous Peoples' Health: A Metasynthesis of Qualitative Research in Canada and the United States. *Int. Indig. Policy J.* **7**, 3 (2016).

11. M. Durie, H. Milroy, E. Hunter, "Mental Health and the Indigenous Peoples of Australia and New Zealand" in *Healing Traditions: The Mental Health of Aboriginal Peoples in Canada*, L. J. Kirmayer, G. G. Valaskakis, Eds. (UBC Press, Vancouver, 2009), pp. 36–55.
12. N. Evans, *Words of Wonder: Endangered Languages and What They Tell Us* (Wiley-Blackwell, Malden, Massachusetts, 2nd Edition., 2022).
13. W. J. Sutherland, Parallel extinction risk and global distribution of languages and species. *Nature*. **423**, 276–279 (2003).
14. L. Campbell, A. Belew, Eds., *Cataloging the World's Endangered Languages* (Routledge, London, 2018).
15. L. Bromham, R. Dinnage, H. Skirgård, A. Ritchie, M. Cardillo, F. Meakins, S. Greenhill, X. Hua, Global predictors of language endangerment and the future of linguistic diversity. *Nat. Ecol. Evol.* **6**, 163–173 (2022).
16. UNESCO, Global action plan of the International Decade of Indigenous Languages (IDIL 2022-2032) (2021), (available at https://unesdoc.unesco.org/ark:/48223/pf0000379851).
17. Q. D. Atkinson, R. D. Gray, Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics. *Syst. Biol.* **54**, 513–526 (2005).
18. J. Schmidt, *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen* (H. Böhlau, Weimar, 1872).
19. A. François, "Trees, waves and linkages: Models of language diversification" in *The Routledge Handbook of Historical Linguistics*, C. Bowern, B. Evans, Eds. (Routledge, New York, 2015), pp. 161–189.
20. J. Nichols, "Diachronically stable structural features" in *Historical Linguistics, 1993: Selected Papers from the 11th International Conference on Historical Linguistics*, H. Andersen, Ed. (John Benjamins, Amsterdam, 1995), pp. 337–356.

21. L. Campbell, *Historical Linguistics: An Introduction* (MIT Press, Cambridge, Massachusetts, Third Edition., 2013).

22. M. Dunn, A. Terrill, G. Reesink, R. A. Foley, S. C. Levinson, Structural phylogenetics and the reconstruction of ancient language history. *Science*. **309**, 2072–2075 (2005).

23. H. Matsumae, P. Ranacher, P. E. Savage, D. E. Blasi, T. E. Currie, K. Koganebuchi, N. Nishida, T. Sato, H. Tanabe, A. Tajima, S. Brown, M. Stoneking, K. K. Shimizu, H. Oota, B. Bickel, Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Sci. Adv.* **7**, eabd9223 (2021).

24. S. J. Greenhill, C.-H. Wu, X. Hua, M. Dunn, S. C. Levinson, R. D. Gray, Evolutionary dynamics of language systems. *Proc. Natl. Acad. Sci.* **114**, E8822–E8829 (2017).

25. R. Dinnage, A. Skeels, M. Cardillo, Spatiophylogenetic modelling of extinction risk reveals evolutionary distinctiveness and brief flowering period as threats in a hotspot plant genus. *Proc. R. Soc. B*. **287**, 20192817 (2020).

26. R. Bouckaert, D. Redding, O. Sheehan, T. Kyritsis, R. Gray, K. E. Jones, Q. Atkinson, Global language diversification is linked to socio-ecology and threat status (2022), *SocArXiv*, doi:10.31235/osf.io/f8tr6.

27. P. Muysken, "Three processes of borrowing: borrowability revisited" in *Bilingualism and Migration*, G. Extra, L. Verhoeven, Eds. (De Gruyter, Inc., Berlin/Boston, GERMANY, 1999), pp. 229–246.

28. F. Meakins, J. Stewart, "Mixed Languages" in *The Cambridge Handbook of Language Contact: Volume 2: Multilingualism in Population Structure*, A. M. Escobar, S. Mufwene, Eds. (Cambridge University Press, Cambridge, 2022), pp. 310–343.

29. D. C. Dennett, *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (Simon & Schuster, New York, 1995).

30. A. Meillet, *Introduction à L'étude Comparative des Langues Indo-européennes* (Hachette, Paris, 1903).

31. C.-T. J. Huang, I. Roberts, "Principles and Parameters of Universal Grammar" in *Oxford Handbook of Universal Grammar*, I. Roberts, Ed. (Oxford University Press, Oxford, 2016), pp. 306–354.

32. M. C. Baker, *The Atoms of Language: The Mind's Hidden Rules of Grammar* (Oxford University Press, Oxford, 2001).

33. J.-L. Mendívil-Giró, Why don't languages adapt to their environment? *Front. Commun.* **3**, 24 (2018).

34. N. J. Enfield, *Natural Causes of Language* (Language Science Press, Berlin, 2014).

35. G. Raîche, T. A. Walls, D. Magis, M. Riopel, J.-G. Blais, Non-graphical solutions for Cattell's scree test. *Methodology*. **9**, 23–29 (2013).

36. S. A. Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, N. Jacoby, A. A. Egner, E. J. Hopkins, R. M. Howard, J. K. Hartshorne, M. V. Jennings, J. Simson, C. M. Bainbridge, S. Pinker, T. J. O'Donnell, M. M. Krasnow, L. Glowacki, Universality and diversity in human song. *Science*. **366**, eaax0868 (2019).

37. J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, C. D. Bustamante, Genes mirror geography within Europe. *Nature*. **456**, 98–101 (2008).

38. J. H. Greenberg, "Some universals of grammar with particular reference to the order of meaningful elements" in *Universals of Language*, J. H. Greenberg, Ed. (MIT Press, Cambridge, Massachusetts, 1963), pp. 73–113.

39. M. S. Dryer, The Greenbergian word order correlations. *Language*. **68**, 81–138 (1992).

40. J. Nichols, Head-marking and dependent-marking grammar. *Language*. **62**, 56–119 (1986).

41. B. Bickel, J. Nichols, "Inflectional morphology" in *Language Typology and Syntactic Description: Volume 3: Grammatical Categories and the Lexicon*, T. Shopen, Ed. (Cambridge University Press, Cambridge, 2007), pp. 169–240.
42. E. Sapir, *Language: An introduction to the study of speech* (Harcourt, Brace and Co., New York, 1921).
43. S. J. Gould, *Wonderful Life: The Burgess Shale and the Nature of History* (W.W. Norton & Co., New York, 1990).
44. M. Muthukrishna, A. V. Bell, J. Henrich, C. M. Curtin, A. Gedranovich, J. McInerney, B. Thue, Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance. *Psychol. Sci.* **31**, 678–701 (2020).
45. B. Bickel, J. Nichols, T. Zakharko, A. Witzlack-Makarevich, K. Hildebrandt, M. Reißler, L. Bierkandt, F. Zúñiga, J. B. Lowe, *The AUTOTYP Typological Databases* (Version 0.1.2., 2017; https://github.com/autotyp/autotyp-data/tree/0.1.2).
46. N. W. H. Mason, D. Mouillot, W. G. Lee, J. B. Wilson, Functional richness, functional evenness and functional divergence: the primary components of functional diversity. *Oikos*. **111**, 112–118 (2005).
47. S. Villéger, N. W. H. Mason, D. Mouillot, New Multidimensional Functional Diversity Indices for a Multifaceted Framework in Functional Ecology. *Ecology*. **89**, 2290–2301 (2008).
48. C. Pimiento, F. Leprieur, D. Silvestro, J. S. Lefcheck, C. Albouy, D. B. Rasher, M. Davis, J.-C. Svenning, J. N. Griffin, Functional diversity of marine megafauna in the Anthropocene. *Sci. Adv.* **6**, eaay7650 (2020).
49. R. Forkel, J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström,

M. Haspelmath, G. A. Kaiping, R. D. Gray, Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Sci. Data*. **5**, 180205 (2018).

50. K. R. Kirby, R. D. Gray, S. J. Greenhill, F. M. Jordan, S. Gomes-Ng, H.-J. Bibiko, D. E. Blasi, C. A. Botero, C. Bowern, C. R. Ember, D. Leehr, B. S. Low, J. McCarter, W. Divale, M. C. Gavin, D-PLACE: A global database of cultural, linguistic and environmental diversity. *PLOS ONE*. **11**, e0158391 (2016).
51. S. Danielsen, M. Dunn, P. Muysken, "The spread of the Arawakan languages: A view from structural phylogenetics" in *Ethnicity in ancient Amazonia: Reconstructing past identities from archaeology, linguistics, and ethnohistory*, A. Hornborg, J. D. Hill, Eds. (University Press of Colorado, Boulder, 2011), pp. 173–196.
52. H. Hammarström, G. Reesink, M. Dunn, H. Skirgård, S. van der Meer, J. Lesage, J. Peacock, R. Singer, H. de Vos, *Nijmegen Typological Survey* (Max Planck Institute for Psycholinguistics, Nijmegen, 2017; https://hdl.handle.net/1839/935A5B75-9624-4C5E-AEB7-AB28C2D8C209).
53. J. L. Fleiss, Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378–382 (1971).
54. J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data. *Biometrics*. **33**, 159–174 (1977).
55. F. Plank, WALS values evaluated. *Linguist. Typology*. **13**, 41–75 (2009).
56. V. N. Polyakov, V. D. Solovyev, S. Wichmann, O. Belyaev, Using WALS and Jazyki Mira. *Linguist. Typology*. **13**, 137–167 (2009).
57. R. Forkel, S. Bank, C. Rzymski, H.-J. Bibiko, clld/clld: clld - a toolkit for cross-linguistic databases (2020), , doi:10.5281/zenodo.3968247.
58. G. Jäger, Global-scale phylogenetic linguistic inference from lexical resources. *Sci. Data*. **5**,

180189 (2018).

59. D. J. Stekhoven, missForest: Nonparametric missing value imputation using random forest (2013), (available at https://cran.r-project.org/web/packages/missForest/index.html).
60. D. J. Stekhoven, P. Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. **28**, 112–118 (2012).
61. T. G. Martins, D. Simpson, F. Lindgren, H. Rue, Bayesian computing with INLA: New features. *Comput. Stat. Data Anal.* **67**, 68–83 (2013).
62. E. Paradis, K. Schliep, ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. **35**, 526–528 (2019).
63. P. J. Ribeiro, P. J. Diggle, O. Christensen, M. Schlather, R. Bivand, B. Ripley, geoR: Analysis of geostatistical data (2020), (available at https://cran.r-project.org/web/packages/geoR/index.html).
64. M. W. Pennell, J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn, M. E. Alfaro, L. J. Harmon, geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*. **30**, 2216–2218 (2014).
65. S. A. Fritz, A. Purvis, Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conserv. Biol.* **24**, 1042–1051 (2010).
66. R Core Team, R: A language and environment for statistical computing (2021), (available at https://www.R-project.org/).
67. R. P. Freckleton, P. H. Harvey, M. Pagel, A. E. J. B. Losos, Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence. *Am. Nat.* **160**, 712–726 (2002).
68. A. V. Bell, P. J. Richerson, R. McElreath, Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *Proc. Natl. Acad. Sci.* **106**, 17671–17674

(2009).

69. G. Csardi, T. Nepusz, The igraph software package for complex network research. *InterJournal*. **Complex Systems**, 1695 (2006).
70. P.-C. Bürkner, brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80**, 1–28 (2017).
71. A. Gelman, B. Goodrich, J. Gabry, A. Vehtari, R-squared for Bayesian Regression Models. *Am. Stat.* **73**, 307–309 (2019).
72. M. Grenié, H. Gruson, fundiversity: a modular R package to compute functional diversity indices (2022), , doi:10.32942/osf.io/dg7hw.
73. H. Hammarström, T. Castermans, R. Forkel, K. Verbeek, B. Speckmann, Simultaneous Visualization of Language Endangerment and Language Description. *Lang. Doc. Conserv.* **12**, 359–392 (2018).

Acknowledgments:

We would like to thank the many language experts, linguists and speakers, who have enriched our dataset by sharing with us their expertise and knowledge of particular languages. These are:

Niina Aasmäe, Alfredo Acosta Blanco, Yvonne Agbetsoamedo, Cynthia Allen, Sunkulp Ananthanarayan, Victoria Apel, I Wayan Arka, Amadu Sajoh Bah, Danielle Barth, Rasmus Bernander, Rogier Blokland, Jeremy Bradley, Mitchell Browen, Yihan Chen, Jiaoyi Chen, Bernand Comrie, Denis Creissels, Mervi de Heer, Rebecca Defina, Cephas Delalorm, Anne Marie Diagne, Rebecca Dixon, Christian Döhler, Mark Donohue, Marie-France Duhamel, Ebikudo Ebitare, Niklas Edenmyr, Nicholas J. Enfield, Gisbert Fanselow, Anne-Marie Fehn, Simeon Floyd, Ulla-Maija Forsberg, Alexandre François, Paul Geraghty, Nikolett F. Gulyás, Roy Stephen Hagman, Hyun-Jong Hahm, Arja Hamari, Abbie Hantgan, Andrew Harvey, Torgny Hedström, Heinike Heinsoo, Caroline Hendy, Sulev Iva, Peggy Jacob, Ivan Kapitonov, Olle Kejonen, Maria Khachaturyan, Myjolynne Kim, Jinyoung Kim, Jacqueline van Kleef, Sjaak van Kleef, Gerson Klumpp, Elizaveta Kushnir, Olga Kuznetsova, Jorge Emilio Rosés Labrada, Kate Lynn Lindsey, Florian Lionnet, Constance Kutsch Lojenga, Carlos M. López Lacayo, Adela López Vargas, Hannah Lutzenberger, Antonio Magaña Macías, Andrej L. Malchukov, Alexandra Marley, Orkhan Mehraliev, Chenxi Meng, Amina Mettouchi, Alexis Michaud, Daria Mishchenko, Mirjan Möller, Zarina Molochieva, Steve Morelli, Maarten Mous, Åshild Næss, David Nash, Tatiana Nikitina, Ratih Oktarini, Bruno Olsson, David Osgarby, Sofia Oskolskaya, Sarah Parkinson, Becky Paterson, Andrew Pawley, Bron Peddington-Webb, John Peterson, Netra Prasad Paudyal, James Lee Pratchett, Saskia van Putten, Tihomir Rangelov, Luis Migel Rojas Berscia, Nicholas Rolle, Paulette Roulon-Doko, Alan Rumsey, Eva Saar, Sophie Salffner, Alexandr Savelyev, Jonathan Schlossberg, Stefan Schnell, Dineke Schokkin, Guillaume Segerer, Frank Seidel, Gunter Senft, Jeff Siegel, Jane Simpson, Yannick Staschull, Lana Grelyn Takau,

Angela Terrill, Jachueline Thomas, Bill Thurston, Yvonne Treis, Laura Trokhymenko, Martine Vanhove, Hein van der Voort, Valentin Vydrin, Alexandra Vydrina, Mary Walworth, Joshua Wilbur, Vera Wilhelmsen, Solace Yankson and Raoul Zamponi. We would also like to thank the speakers and signers who collaborated with linguists to make the descriptive works we rely on possible. We are also grateful to the spouse of the late Dr Jane Hill, Dr Kenneth C. Hill, who supported Dr Jane Hill being included as a co-author.

We would also like to thank David Orme - creator of the R-package caper - for technical assistance in analysis and Adrian Bell and Bret Beheim for guidance regarding the cultural fixation scores.

Funding:

Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology, Leipzig

Department of Language and Cognition at the Max Planck Institute for Psycholinguistics, Nijmegen

Royal Society of New Zealand Marsden grant (UOA1308) to QDA and RDG.

Australian Research Council Centre of Excellence Grant (CE140100041) for the ARC Centre of Excellence for the Dynamics of Language. to NE, SG, HS

NSF BCS-1423711, BCS-0844550 and HSD-0902114

University of Turku: Research Infrastucture Support Grant 2018-2020 to an interdisciplinary consortium (main PI Päivi Onkamo) to OV and MN.

European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 646612) to MR.

Author contributions:

Conceptualization: RDG, SG, QA

Formal analysis: SG, HS, SPA, HJH, LM, DEB, RD, AC

Software: SG, HS, SPA, HJH, LM, DEB, RD

Validation: SG, HH, RF

Coding: KA, DA, DEB, NB, GB, RB, IC, SD, LD, ED, GF, YGA, HG, HPG, JG, VG, AH, RHA, LH, RHE, NH, BHR, JI, MJ, EJ, EK, CK, JK, NK, KK, RK, OK, NL, ML, HL, TM, CMG, SVM, MM, SM, KN, JN, MN, CAO, JP, NP, SPE, SPI, DP, LR, AR, JR, SR, JRI, ER, KSA, JSA, LS, RSC, FS, AS, WDS, HDS, KS, DV, JV, JW, TWI, HW, SY, JY, MY, TY, HS, MD, GR, RSI, HJH, HH, JC, JLA, JLE, AWM, TWE, CB, PE, JH, JE, RZ, JVE, MP, JM

Visualization: SG, HS, SPA, RDG, DEB, RD

Data curation: HS, HH, RF, SG

Funding acquisition: RDG, QA, OV, NE, MR, SCM

Project administration (feature patrons): HS, HH, AWM, HJH, JLE, TWE, JLA, JC

Supervision: RDG, SG, HS, HH, AWM, HJH, JLE, TWE, JLA, JC, MH, MD, GR, RSI, CB, PE, JH, OV, MR, RF, CB, PE, JH

Writing – original draft: HS, HJH, RDG, SG, DEB, QA, HH, LM, SPA

Writing – review & editing: HS, HJH, RDG, SG, QA, DEB, NE, MH, SCL, SPA

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data in this analysis is available to the reader in supplementary material and publication of data and scripts on in the scientific archive Zenodo,

with the exception of the data on population, official status and writing status which is available from SIL International as part of the twenty-third edition of Ethnologue's language dataset: https://www.ethnologue.com/pricing.

Supplementary Materials

SM1: Material and methods
- SM1:1 Grambank structure and design
- SM1:2 Technical validation
- SM1:3 Web interface
- SM1:4 Accessing data
- SM1:5 License and referencing
- SM1:6 Data coverage
- SM1:7 Preparation of data for analysis (removal of dialects, binarisation, cropping and imputation)
- SM1:8 Spatiophylogenetic analysis
- SM1:9 Principal Component Analysis
- SM1:10 Cultural Fixation scores
- SM1:11 Unusualness analysis
- SM1:12 Calculation of Manhattan distance
- SM1:13 Functional richness

SM2: Supplementary figures

SM3: Supplementary tables

SM1 Materials and methods

SM1:1 Grambank structure and design

Over 80 contributors have participated in the coding of the Grambank features, and a team of seven feature experts has supported their work. Extensive descriptive and procedural documentation for each feature was used to ensure reliable coding. Formal testing of inter-coder reliability demonstrates a high degree of consistency across coders. Care was taken to remove strict logical dependencies between features to eliminate the problem of non-independent data-points. As is the nature of languages, other kinds of dependencies may remain and are possible to explore with the dataset and to control for given the extensive documentation.

Grambank is available in the Cross-Linguistic Linked Data framework via the Cross-Linguistic Data Format (*49*). The dataset uses Glottolog language codes to identify languages (*1*), ensuring clear identification of languages and compatibility with other linguistic and cultural datasets, such as D-PLACE (*50*).

Institutional history

The Grambank project began as a joint project in 2015 between departments in two Max Planck Institutes (MPI): the Language and Cognition department (L&C) of the MPI of Psycholinguistics in Nijmegen, Netherlands – led by Stephen C. Levinson – and the Department of Linguistic and Cultural Evolution (DLCE) now at the MPI for Evolutionary Anthropology (MPI-EVA) in Leipzig, Germany – led by Russell Gray. This collaboration took place within the larger international research consortium named Glottobank, which also involves the Centre of Excellence for the Dynamics of Language in Canberra, Australia, and the University of Auckland, New Zealand. The Australian National University, University of Kiel, Uppsala University and the School of Oriental and African Studies also take part in the organization of Grambank.

The Grambank database builds on the work by the Nijmegen Typological Survey from the L&C department at MPI-Nijmegen led by Stephen C. Levinson and Harald Hammarström, as well as on the works of the Pioneers of Island Melanesia project and the Sahul survey, led by Ger Reesink and Michael Dunn. Grambank has inherited features (see next section) from these surveys as well as data points. Coders who have contributed to these preceding databases are also attributed as coders in the Grambank dataset. In acknowledgment of the work that went into the Sahul survey design we would like to thank Angela Terrill, Eva Lindström, Gunter Senft, Nicholas Evans, Sjef Barbiers, Mily Crevels, Rob Goedemans, Pieter Muysken, Leon Stassen and Hein van der Voort for their contribution to that questionnaire.

Grambank contains some data points that were originally published elsewhere: Hunter-Gatherer Language Database, SAILS and the aforementioned NTS & Sahul surveys. The database contains imported data points from the typological section of the Hunter-Gatherer database (HG), led by Claire Bowern, Patience Epps and Jane Hill. The HG database does not contain a one-to-one match between its features and features in Grambank. Data-points for import were matched carefully by Harald Hammarström, Thiago Chacon, Hedvig Skirgård, Hannah Haynie, Judith Voss and Jakob Lesage. Grambank also contains imported data-points from the work of Swintha Danielsen on Arawakan languages (*51*). Danielsen's work was based on the Sahul

survey which also serves as the base of Grambank, therefore import was straightforward. Imported datapoints are attributed to the appropriate coders in the Grambank dataset.

Grambank feature selection

The set of features included in Grambank reflect a balance between several design principles and practical pressures. The principles guiding the construction of this database included obtaining maximal coverage of the sorts of typological information contained in source materials that describe the world's languages, constructing a simple data structure with clear and interpretable feature values, and preserving compatibility with legacy data. The Grambank questionnaire was created by a team of linguists in the Glottobank consortium, drawing on experience primarily from the Nijmegen Typological Survey (NTS, *52*), which in turn builds on the Sahul survey. The NTS constituted a core questionnaire upon which Grambank was built, with additional inspiration from the data and experiences of the Pioneers of Island Melanesia project (*22*). The influential typological database WALS (*8*) also inspired features of both the NTS and Grambank. 103 of the 195 features in Grambank are inherited from the questionnaire of the Pioneers of Island Melanesia and 40 from the NTS, making these features well tested and documented.

The questions describe a wide variety of morphosyntactic and lexical features likely to be discussed in a grammatical description, such as word order, the existence of prefixes and suffixes with particular functions, marking of grammatical categories, and agreement rules. Each feature can be coded using grammars and grammar sketches without necessarily requiring the coder to have a comprehensive knowledge of the entire language. Empirically, a randomly selected language that is described by at least a grammar sketch can be filled in for 68% of the features on average (see Supplementary Material 1:2).

Each feature in the questionnaire is structured in the form of a brief feature name, a feature description, feature ID, and a set of possible feature values. Feature names take the form of a question that typically probes the presence or absence of an individual grammatical element. Feature summaries provide a succinct description of the targeted phenomenon and the criteria that should be used to identify it. A source field is used to cite the resource and page number where the coded information was found. The comments field allows the coder to enter any additional information that may be useful for understanding their response.

Dependencies

Typological surveys that cover a large range of grammatical topics often contain data points that are not logically independent from one another. For example, in a database that has features for the number of case categories and the position of case marking, any language that is coded as having suffixed case marking will also necessarily be coded as having case categories. Such dependencies might complicate the analysis of comparative data. For this reason, the Grambank dataset largely eliminates strict logical dependencies between the features.

It is worth noting that the following Grambank features participate in a near strict logical loop.

- GB020 Are there definite or specific articles?
- GB021: Do indefinite/non-specific nominals commonly have indefinite/non-specific articles?

- GB022: Are there prenominal articles?
- GB023 Are there postnominal articles?

A "Yes" for GB020 and/or GB021 would seem to suggest a "Yes" for GB022 and/or GB023 and vice versa (the existence of articles presupposed they have a position, and if there are articles that have a position, it would suggest they are either definite/specific or indefinite). However, this is not a strict loop because there are articles that do not trigger a "Yes" for GB021 that can trigger a "Yes" for GB022 and/or GB023.

There are also two sets, outlined below, where it is impossible for a language to be coded as 0 for all features. It is not possible to have no word order whatsoever, and to not have at least one alignment system. For more on the specifics of this, see the feature documentation accompanying the dataset. Note that it is possible to have other value combinations, such as "1-1-1" or "0-0-?".

Transitive verb-order set

- GB131 Is a pragmatically unmarked constituent order verb-initial for transitive clauses?
- GB132 Is a pragmatically unmarked constituent order verb-medial for transitive clauses?
- GB133 Is a pragmatically unmarked constituent order verb-final for transitive clauses?

Alignment set

- GB408 Is there any accusative alignment of flagging?
- GB409 Is there any ergative alignment of flagging?
- GB410 Is there any neutral alignment of flagging?

Furthermore, besides the strict logical dependencies discussed so far there are other kinds of dependencies that are relevant for understanding languages. There is for example, as one of our anonymous reviewers pointed out, a likely historical connection between different elements all being pre-posed to the noun.

Given our extensive documentation of the features it is possible for users to identify such connections. One manner in which this can be addressed by users is by constructing new meta features that encompass and depend on our original features. For example, the three features below all concern marking of gender in the pronoun system:

- GB030 Is there a gender distinction in independent 3rd person pronouns?
- GB196 Is there a male/female distinction in 2nd person independent pronouns?
- GB197 Is there a male/female distinction in 1st person independent pronouns?

It may be interesting for a user to combine them to derive a feature asking "Is there a gender distinction in pronouns?".

Dependencies other than the strict logical dependencies and the thematic relationships between features described above are known to exist between grammatical features (e.g. a likely historical connection between different elements being pre-posed to the noun noted by an anonymous

reviewer). Dependencies arising from language use and history are topics of ongoing research that the Grambank dataset can facilitate.

Example feature documentation

For each feature, we provide documentation that aims to aid the coders in applying the questionnaire consistently over the entire language sample. The features are described by each patron at our shared wiki (https://github.com/grambank/grambank/wiki), and this information is then found in the CLDF dataset in the Parameters table. Below is an example of the documentation provided.

Feature ID: GB028
Name: Is there a distinction between inclusive and exclusive?
Patron: Hannah J. Haynie
Summary:
Is there a pronoun or other marker that explicitly marks the inclusion of an interlocutor? This feature is not restricted to the pronominal system but includes person indexing as well. If inclusive is marked overtly in either the pronominal system or through verbal marking this is sufficient to trigger a 1 for this feature, even if exclusive has no overt morphological marking.

Procedure:
1. Code 1 if there is a pronoun or other marker, such as a person index, that explicitly marks the inclusion of an interlocutor in the first person plural.
2. Code 0 if the sections of the grammar discussing pronoun systems and person indexing on verbs describe no distinctions between inclusive and exclusive persons, and no pronominal forms or indices are found in examples glossed with grammatical information including INCL/EXCL or meanings such as ‘you and I’ or ‘we all (not you)’. Pay close attention to the non-singular forms of first person pronouns and indices.
3. If you are uncertain whether some pronominal or index form(s) mark(s) a clusivity distinction (e.g. a form in a single example glossed ‘you and I’ that is known to encode dual number but is not clearly described regarding inclusivity, or multiple first person pronouns whose differences are not adequately described), code ? and provide a brief comment describing the forms or descriptions that were unclear.

Examples

Southern Sierra Miwok (ISO 639-3: skd, Glottolog: sout2985)

Personal Pronominal Suffixes:

	Series 1	**Series 2**	**Series 3**	**Series 4**
1DU.INCL			*-ti:*	*-ti:*
1PL	*-tti-/-mahhi:*	*-me-*		

1PL.INCL			*-ticci:*	*-ticci:*
1PL.EXCL			*-mahhi:*	*-mahhi:*

(Broadbent 1964: 43)

Southern Sierra Miwok would be coded as 1. The lack of a first person dual exclusive form does not affect this designation, nor does the fact that the language has first person plural markers in Series 1 and 2 that do not mark clusivity.

Chalcatongo Mixtec (ISO 639-3: mig, Glottolog: sanm1295)

Pronouns

PERS	**GENDER**	**FREE**	**CLITIC**
1	Familiar	*rùʔù*	*=rí*
	Polite	*naʔa*	*=na*
	Inclusive (pl)	*žóʔó*	*=žó*

(Macaulay 1996: 81)

Chalcatongo Mixtec would be coded as 1. A plural pronoun that is unmarked for clusivity can be derived from the polite or familiar first person pronouns with a prefix, but the inclusive first person is inherently plural. There is no first person plural pro-form that is marked for exclusivity. The existence of an inclusive form is sufficient to trigger a 1 and the lack of an exclusive form has no impact on this.

Yongbei Zhuang (ISO 639-3: zyb, Glottolog: yong1276)

First person

Singular	**Plural (excl.)**	**Plural (incl.)**
ku	*tuo*, *po tu*	*lau'*

(Luo 2008: 327)

Yongbei Zhuang is coded as 1.

Grambank feature values

Individual structural features were formulated to take mainly binary (yes/no) values. This ensures a simple data structure, maximal clarity and interpretability of each datapoint, and a standard data format for the majority of the data. Six features have multistate values, each of which describes a particular word order or set of word orders that are available in that language. This makes it possible to identify situations where multiple word orders are possible without creating a logical dependency between features. They can be binarised, as seen in Supplementary Material 1:7 and Table S5.

Grambank departs from the traditions of many typological databases, like many chapters in WALS, in encoding whether a particular strategy for expressing a specific function is possible in a language, rather than stating what the single most common or dominant strategy is for expressing that function. The approach that Grambank uses aims to preserve valuable information about the spectrum of expressive possibilities in a language.

There are two types of missing data represented in Grambank. First, a response marked with a '?' denotes a datapoint where the source materials contain insufficient information for the coder to determine the value. A '?' response is accompanied by a reference to the source(s) consulted by the coder. A missing (empty) value represents a data-point for which no coder has made an attempt to code that particular feature for that language. There is thus a distinction in the data between values that have been checked, but could not be coded definitively at that time ('?') and values that are entirely missing for that feature/language combination. These two types of missing data in Grambank are different still from the 'not applicable' values used in some typological databases which is used to indicate that a particular feature is not relevant to a particular language because of another feature value. The formulation of Grambank questions removes the 'not applicable' distinction and the absence of a phenomenon is simply coded as '0' (absent) in this dataset.

Grambank data collection

The primary sources used in Grambank are published descriptions of grammatical structures. There are over 7,000 languages found around the world, and of these, approximately 60% are described by a grammar or a grammar sketch (*1*). Data for Grambank were also obtained by consulting linguists with expertise on particular languages; see acknowledgements for a list of experts who have shared their knowledge.

The coding workflow and support structures employed by Grambank were designed to minimize any potential data compatibility and consistency issues that may arise from the diversity of source materials considered. The questionnaire is adapted to being answerable to a standard level given a grammar sketch, and coders were provided with continuous support for discussing and evaluating possible interpretations of the data. Differences in the quality of linguistic descriptions across languages and the existence of competing analyses impacts the completeness of data for individual languages, but should have minimal impact on coding decisions.

Data were entered by research assistants and language specialists who filled in the Grambank questionnaire using available grammars and provided references for each datapoint, as well as comments if appropriate. Coders were trained to fill in the questionnaire by local supervisors who were involved in the design and ongoing curation of Grambank features. Training included

coding a previously coded language, detailed supervisor-led discussion of each questionnaire feature, introduction to the project's documentation and discussion forum, and examination of previous discussions and complicated coding decisions. A key feature of the Grambank coding process was that each feature had one or a pair of feature experts – known project-internally as "patrons" – who adjudicated complicated coding situations where agreement cannot be reached in discussions between the local supervisor and individual coders. In cases where there was doubt or disagreement about specific coding decisions, the patron made the final judgment. Documentation of each feature can be found in our GitHub repository's wiki (https://github.com/grambank/grambank/wiki). In this way we ensure consistency across coders and provide a rich documentation of the decisions required to convert the complexity of a grammatical description into a large-scale digital database in a transparent and reproducible manner.

Grammars often do not explicitly state whether a particular phenomenon is absent. Coders therefore have to inspect not only the text, but also the available language examples in order to make informed judgments about the values of features. In some cases it is difficult to judge whether no mention of a feature in the available grammar(s) is evidence that the phenomenon itself is absent in the language, or simply an oversight or omission by the author. The coder judges this by how extensive the description of that grammatical domain is in the grammar (e.g. it can typically be assumed that definite articles are absent if they are not mentioned in a section on the noun phrase). In cases where there was uncertainty and it could not be resolved with more examination of the sources and discussion, the relevant feature was coded as '?' for that language.

SM1:2 Technical validation

An inter-coder reliability study was conducted to assess the quality of the curated Grambank data. 20 languages were randomly selected from the set of 4,338 languages with a grammar or grammar sketch. For each of the 20 languages, three out of six members of the Grambank design team were randomly selected to code the language independently of each other. They were each given the same instructions, the same deadline, the same preparatory and auxiliary materials and the same source documents describing the language in question. In this way, a total of 8,311 data-points were collected, which allowed for 7,876 pairwise comparisons.

Coders disagreed most often on the basic issue of whether there is enough information to assign a specific value for a particular feature: in 25% (1996/7876) of the comparisons one of the coder assigned a '?' and the other a specific value. In 20% (1557/7876) of the comparisons both coders agree on a '?', i.e., that there is insufficient information for concluding a specific value. When both coders assigned a specific value for the language, however, they agreed on the value 87% of the time (3753/4323). This number rises to 90% if only datapoints based on the same grammatical description are compared. While pairwise comparisons are simple to interpret, they are not controlled for number of raters and chance agreement. Fleiss' Kappa (*53*) calculates the measure of agreement over chance, which in this study is 0.72. While there are no widely established standards of significance for Fleiss' Kappa, guidelines (*54*) classify this score as "substantial agreement".

As the bulk of the coded data in Grambank has been collected by research assistants and the above inter-rater reliability study involved members of the design team rather than these research assistant coders, one may legitimately ask whether the results generalize from experts to research assistants. While no controlled study was used to answer this question, there were cases of unplanned double-coding. Among these double-coded languages, there were two languages that also featured in the inter-coder reliability study above. These can provide a general measure of how research assistant coding compares to expert coding. The levels of agreement when comparing research assistants with other research assistants (78%, 79%, 87%, 91%, 91%), research assistant vs. expert (87%, 89%, 95%, 96%), and expert vs. expert (87% as above) do not differ appreciably. The reason for this may be that time and devotion to the task makes up for the difference in expertise. Few other figures on reliability of typological databases are available for comparison. However, an accuracy rate of 87% is similar to rates for a select few well known languages in WALS (*55*) and Jazyki Mira (*56*). Hence, this may be the natural margin of error associated with human factors and the abstraction level of typological features.

SM1:3 Web interface

The latest released version of the Grambank database is available for interactive browsing at https://grambank.clld.org under a Creative Commons 4.0 Attribution license. It is served by a web application built with the toolkit developed for the Cross-Linguistic Linked Data project (*57*). Consequently it inherits the core database schema common to all CLLD applications, which includes standard data types for common entities such as:

- *contribution:* a citable sub-unit of a dataset
- *language:* an instance of the main subject of study
- *parameter:* a measurable factor which can be compared across languages -- a *feature* in Grambank
- *value:* a measurement, i.e. a value determined for a particular language and a given parameter
- *source:* a bibliographical record describing the source of a value

The CLLD framework also provides tools for basic analysis and visualization of underlying data. The Grambank website integrates these tools into interfaces for accessing data by feature or by language, with further pages that summarize data by other fields (e.g. language family, source). The Languages page also presents an interactive mapping tool, as well as a table of coded languages that can be searched by ID, language name, or latitude/longitude. The Features page of the website presents a list of features in tabular form, and can be filtered by ID, name, morphosyntactic unit, form, or grammatical function. Linked pages for individual features provide further information about the feature, data values in tabular format, and an interactive tool that enables map visualization of feature value distributions. Additional filters allow users to sort languages by families and macroareas.

SM1:4 Accessing Data

The Grambank data are archived with Zenodo as a Cross Linguistic Data Format (CLDF) structure dataset (*49*). Because the CLDF format is essentially a set of CSV files, it is simple to access the data from a wide variety of computing environments. Unzipping a download of the whole of Grambank CLDF dataset will result in a directory with the following contents:

- StructureDataset-metadata.json: The machine readable description of the dataset
- values.csv: The main data file, containing all codings
- languages.csv: A CSV file with additional metadata about the coded languages
- parameters.csv: A CSV file with metadata about the coded features.
- sources.bib: A BibTeX file containing bibliographic metadata about the sources used for Grambank coding.

Methods for accessing and using this data in environments such as Python, SQL, R, and with off-the-shelf CSV tools are described in detail at the GitHub repository of the CLDF dataset.

SM1:5 License and referencing

Grambank is released under a Creative Commons 4.0 (CC-BY) license. Any user may share and adapt the data, as long as they give appropriate credit by citing this paper and the relevant version of the database. Languages are still being added to Grambank and the project welcomes feedback from experts, which may result in additions or changes in the coding of languages. The web publication of Grambank will be updated regularly with new releases; therefore users should reference the Grambank data they use by its specific release version and download date. The first version is 1.0 and should hence be referenced as “Grambank 1.0”, this is the dataset that is presented in this paper and consists of 2,467 languoids (languages, dialects and proto-languages).

SM1:6 Data coverage

The Grambank dataset contains 2,467 language varieties and 195 features. For analysis in this paper, we chose to remove all but one dialect per language, which leaves us with 2,430 languages (see Supplementary Material 1:7). The dataset contains 24% missing data and spans all continents and major language families.

The Grambank data gathering procedure progresses per language, i.e. the entire questionnaire is filled in as much as possible for one language at a time. This leads to high data coverage. Grambank contains 24% missing data, which can for example be compared to 84% in WALS (*8*). Fig. S1 shows this comparison.

The Grambank questionnaire is filled in primarily based on published grammatical descriptions (typically sources classified as "grammar" or "grammar sketch" in Glottolog (*1*)). Fig. S2 shows the Grambank coverage per Glottolog macroarea.

SM1:7 Preparation of data for analysis (removal of dialects, binarisation, cropping and imputation)

For the analysis in this paper it was necessary to merge dialects, binarize features with multi-state values, prune away features and languages with large amounts of missing data and/or impute the remainder of the missing data. The resulting subsets of the data were used in the analysis, it is specified for each analysis if the imputed dataset was used. For analysis that involves the global phylogeny, only languages which are represented by a tip in that phylogeny were included.

There are 2,467 language varieties in Grambank. This includes 70 dialects. In order to maximize the overlap with other data sources used in the analysis (e.g. WALS (*8*), AUTOTYP (*45*), and the global tree (*26*)), we chose to drop all but one dialect per language. The dialect that was kept was the one with the least amount of missing data. The remaining language variety is assigned the glottocode of its parent language variety that is classified as "language" in Glottolog (i.e. not "dialect").

For the comparison of coverage between WALS (*8*) and Grambank (see Fig. S1), we also reduce dialects in WALS by keeping the one with the least amount of missing data in the same fashion. This leaves 2,430 languages in Grambank and 2,435 languages in WALS. There were 35 languoids in WALS that were not mapped to a glottocode and therefore not possible to include in the comparison at all.

We did the same procedure to the tips of the global language tree (*26*), dropping all but one tip per language (at random) if there were multiple dialects included and assigned it the glottocode of its parent language. We also dropped tips in the global tree that did not correspond to languages in our pruned and imputed dataset (see below for imputation procedure). This left 1,404 tips in the global tree and languages in the Grambank dataset for the spatiophylogenetic analysis (see Supplementary Material 1:8).

There are six features in the Grambank dataset that have multi-state values; all others are binary. Multi-state features are all of the type: "what is the order of element X and Y?" with the alternatives "XY, "YX" or "both". They were all split into two features each, of the format "Is the order XY?" and Is the order YX?" with the "both" values triggering a 1 (yes) for both features. This process gives 201 binary features out of the original 195.

The full dataset contains 24% missing data. In order to avoid problems of excessive imputation, we first crop the dataset such that we remove features and languages with more than 25% missing data leaving 1,509 languages and 113 binarised features.

There remains 4% missing data in the cropped dataset. This missing data is imputed using a random forest trained on the observed values, as implemented in the R-package 'missForest' v. 1.4 (*59*, *60*). The Out of Bag error rate is estimated at 14%. The random forest technique is entirely naive as to language genealogy or geography; it imputes missing data based on languages with a similar profile regardless of relatedness or spatial distance.

All the code associated with this paper is published alongside the paper, including data wrangling from CLDF to the scripts generating each plot in this paper. The code is found on GitHub and Zenodo. The scripts that prepare the data according to the above procedure are:

make_wide.R
make_wide_binarized.R
impute_missing_values.R
compare_coverage_WALS.R
spatiophylogenetic_modelling/processing/pruning_EDGE_tree.R

SM1:8 Spatiophylogenetic analysis

The estimation of spatial and phylogenetic effects for each feature of Grambank was calculated using a binomial spatiophylogenetic model following the procedure laid out in (*25*). This model is implemented using Integrated Nested Laplace Approximations (INLA) of a Bayesian model using the *R* package *INLA* v20.03.17 (*61*).

The model contains two structured random effects: one representing the phylogenetic relationships between languages, and one representing the spatial distances. A key departure from the procedure laid out in prior research (*25*) is that the spatial relationships are represented as spatial coordinates, unlike in the procedural paper where spatial relationships are represented within a spatial mesh. We use coordinates to ensure spatial and phylogenetic variation are compared on an equal footing, with one phylogenetic taxon and one location per language.

Phylogenetic relationships are drawn from a recently released Bayesian posterior distribution of phylogenetic trees capturing genealogical relationships between the world's languages (*26*). We use the maximum clade credibility tree derived from this posterior distribution, which incorporates prior information on established genealogical classifications within families (*1*), conservative confidence intervals on the timing of internal diversification and origin of families, a phylogeographic model of language diversification in space, and archaeological and genetic evidence of human expansion around the globe.

Spatial relationships are built from the latitude and longitude of language metadata, collected by Glottolog (*1*). We can only include languages from Grambank that are also represented in the phylogeny. There are 1,404 languages that appear in both the dialect-dropped, cropped and binarised Grambank dataset (see Supplementary Material 1:7) and the phylogeny. In order to maximize overlap, the global tree was also dropped for dialects (dropping all but one tip at random out of sets of tips which are dialects of the same language). The dataset used in this analysis contains 4% missing values, we did not impute them. We followed the same principles for cropping for missing data as outlined in Supplementary Material 1:7, leaving us with 113 features.

The spatiophylogenetic model uses precision matrices to represent the phylogenetic and spatial relationships, which are calculated from covariance matrices. Phylogenetic covariance is estimated through a model of Brownian motion, and spatial covariance is determined through a Matérn covariance function. The phylogenetic covariance matrix is built using the vcv.phylo function from the *R* package *ape* v 5.4-1 (*62*), and the spatial covariance matrix is built from the varcov.spatial function in the package *geoR* v1.8-1 (*63*), using the Matérn covariance function with the parameters: sigma = 1.15 & kappa = 2. Covariance matrices are standardized to have a variance of approximately 1 by dividing the matrix by its typical variances, before being inverted to become precision matrices.

Penalizing-complexity priors are set for each random effect, which offer a 10% chance of variance being >1, although prior choice has little influence on the results (*see below*).

Spatial parameterization: In addition to the Matérn parameters described above, we test two additional Matérn parameters (kappa and sigma), which iteratively expand the influence of

spatial relationships (Fig. S10). Increasing the reach of spatial relationships had little influence on our general conclusions (Fig. S11).

Prior choice: Following earlier research (*25*), priors for both the phylogenetic and spatial effects used the 'pcprior' (penalized complexity prior) distribution with parameters 1 & 0.1, which correspond to an exponential distribution with ~10% of of its probability above 1. To test the sensitivity of the results to these priors, we range the probability above 1 to vary from 1% (very strict), 10%, 50%, and 99% (effectively uniform). The choice of prior had negligible effects on parameter estimates and did not change the mode comparison results (see Fig. S12). We used 10% (pcprior = 0.1) for the main analysis.

Simulations: To ensure the spatiophylogenetic model will return statistically valid results, we ran a series of simulations using the phylogenetic and geographic location of the Grambank sample. Simulated binary variables varied across two conditions: the amount of phylogenetic signal (Pagel's Lambda of 0.01, 0.3, 0.6, & 0.9), and the proportion at which traits occur (0.1, 0.25, & 0.4) – a total of 12 conditions. Variables were simulated using the *geiger* v2.0.9 (*64*) function sim.char(). Variables were simulated 15 times per condition. Phylogenetic signal is varied using geiger and the function rescale(), which rescales the phylogeny branch lengths according to the desired parameter. The proportions were gathered by randomly generating the Q matrix and repeating the simulation until the desired proportion and signal was retrieved. Fig. S13 shows the results of the simulations. In all conditions, the dual process model correctly identifies the phylogenetic signal over spatial signal. Both the "phylogeny only" and "dual process" models estimates of phylogenetic signal in the correct rank order. The error around phylogenetic estimates aligns with existing simulation results for estimating signal in binary traits (*65*). As traits become equally common (there are as many 1's as there are 0's), the precision of the phylogenetic estimate decreases, although phylogenetic signal is still observed in the correct rank order, and does not confuse phylogenetic signal with spatial relationships.

Ancestral State Reconstruction: To illustrate more clearly the structure of the phylogenetic signal in the three features with the strongest phylogenetic signal, we used the INLA approach to reconstruct ancestral states of proto-languages for each feature respectively. The analysis is the same as for the main spatiophylogenetic analysis (kappa = 2, sigma = 1.15, dual model with both phylogeny and spatial precision matrices). The key difference lies in the phylogenetic precision matrix, which in this analysis also includes positions for the ancestral language - internal nodes in the tree. These nodes are not associated with feature values, those values are missing. The INLA-model estimates predictions for missing values, based on the fitted posterior distribution, thus producing predicted feature values of the ancestral states. Note that these internal node positions are not associated with any spatial information, i.e. we have not inferred any longitude or latitude of proto-languages. Spatial information is however included in the overall model as information about the tip-values (this means predictions for internal nodes are made with the spatial field set to zero, that is, with spatial variation estimated from the tips 'removed'). See Figures S3-S5 for tree plots of the result of this analysis. These figures show the three features with the strongest spatial signal out of the whole set and their distribution across the world.

Testing the association between domain and spatial & phylogenetic effect: The features of grambank can be divided into four different domains: clause, verbal, nominal and pronominal. You can see the mean phylogenetic and spatial effects per feature as grouped by these domains in

Fig. 1 in the main text. In order to test whether domain membership predicts phylogenetic and spatial effects we ran BRMS models with and without the domains as a predictor and compared their model fit scores. We used a beta distribution since the values are bound between 0 and 1 and compared WAIC scores. The response variable is the mean spatial and phylogenetic effect per feature respectively, with the default INLA model parameters (kappa = 2, sigma = 1.15 and pc prior = 0.1). Specifically we ran four BRMS models for the 113 grambank features:

- a null model where the intercept predicts the spatial effect for features
- a model where the domain predicts the spatial effect for features
- a null model where the intercept predicts the phylogenetic effect for features
- a model where the domain predicts the phylogenetic effect for features

The difference in WAIC values between the null and domain models for the effects was smaller than the SE of this difference, from which we conclude that there is no improvement in predictive accuracy from taking feature domain into account.

SM1:9 Principal Component Analysis and theoretical scores

We carried out a traditional non-weighted Principal Component Analysis to derive the dimensions along which data primarily varied. We used the function *prcomp* in the statistical programming language *R* v4.1.0 (*66*).

The data was binarised, cropped and imputed for the PCA (Supplementary Material 1:7). It is necessary that the data is binarised because the PCA relies on the mean of each variable, which in the case of the multi-state features is not meaningful. It is also necessary to remove and/or impute missing data as PCA requires a complete dataset. The variables were scaled to have unit variance and centered.

We examined the rotations/loadings of the components for each feature (Fig. S14). In order to evaluate what phenomena most contributed to each component, we also examined the rotated data per language and compared to other aggregate scores capturing known linguistic theoretic concepts.

We compared the rotated data to concepts used in linguistic typology to characterize language variation. For each of these concepts we created an index that measures, for each language, the occurrence of Grambank feature values that might be expected in a language that perfectly exemplifies the relevant theoretical properties. The concepts we encoded with typological indices are:

- *word order* (the degree to which a language uses structures hypothesized to correlate with verb-object or object-verb word order in (*38*, *39*))
- *locus of marking* (the degree to which a language mainly features head or dependent marking, as described by (*40*)),
- *fusion* (degree to which a language encodes meanings and functions with bound morphology as opposed to phonologically free-standing markers (*41*))
- *flexivity* (degree of allomorphic variation (*41*))

The nature of the questions in the Grambank questionnaire prevents us from exploring other typological concepts like Bickel and Nichols' "exponence", which expresses the degree to which individual morphemes encode multiple functions/meanings.

Each of the above metrics were calculated by assigning values to each Grambank feature that express information about the phenomenon captured by that metric (0, 0.5 and 1), according to the extent to which the feature is consistent or inconsistent with the typological phenomenon. For *word order* our feature-wise metric values reflect consistency with proposed verb-initial word order patterns, and for *locus of marking* the feature-wise metric values reflect consistency with proposed head marking patterns. We used these values to calculate per-feature indices of consistency with the metric's theoretical concept and then expressed a language's overall score for any metric as the mean of that language's consistency indices. A value of 0 assigned to a feature indicates that the feature contradicts the pattern or phenomenon measured by a metric. For these features that oppose the patterns captured by the theoretical metrics we reverse the values of language-specific coding in the consistency index, i.e. 0 becomes 1 and 1 becomes 0. For example, for the word order metric features related to verb-final orders such as GB022: "Are there prenominal articles?" and GB133: "Is a pragmatically unmarked constituent order

verb-final for transitive clauses?" have a "word-order-point" value of 0. Features associated with verb-initial order such as GB023: "Are there postnominal articles" and GB262: "Is there a clause-initial polar interrogative particle?" are awarded a "word-order-point" value of 1. If the language value is 1 and the word-order-point value of that feature is 1, the word-order metric consistency index for that feature in the language is 1. Each language will thus be assigned a consistency index of either 0 or 1 for each feature. The assignment of per-feature *word order* consistency indices based on the interaction of "word-order-point" feature values and language-specific feature coding and the calculation of mean *word order* score per language are illustrated in table S9:1 for four features and three languages.

Table S9:1. Example of theoretical metric calculation

Feature	**word-order-point**	**poko1263**		**hind1269**		**khak1248**	
		Language-value	word-order-value	Language-value	word-order-value	Language-value	word-order-value
GB022 Are there prenominal articles	**0**	0	1	0	1	1	0
GB133 Is a pragmatically unmarked constituent order verb-final for transitive clauses?	**0**	0	1	1	0	1	0
GB023Are there postnominal articles	**1**	1	1	0	0	0	0
GB262 Is there a clause-initial polar interrogative particle?	**1**	1	1	1	1	0	0
mean word order score			**1**		**0.5**		**0**

For our *fusion* metric we assigned a value of 0.5 to features that are consistent with the typological pattern of expressing information through phonologically bound morphs but which do not necessarily indicate that grammatical information is expressed by phonologically fused elements. For example, GB075: "Are there postpositions?" encodes whether languages use an element that follows a noun to express adpositional meanings. Both postposition words (which are phonologically independent) and postpositional enclitics (which are phonologically fused)

can trigger a 1 value for this feature. Because a 1 value for this feature is not inconsistent with the concept of typological fusion but does not necessarily mean that the language uses phonologically fused enclitics for this function, we assign a value of 0.5 for this feature. This value is multiplied by a language's feature value to obtain a feature-level index of consistency with *fusion* (i.e. in languages where this feature is coded 1 the feature index for *fusion* will be 0.5, while in languages where the feature is coded 0 the feature index for *fusion* will be 0).

A high score for our word order metric indicates that a language has relatively more order features that have been hypothesized to correlate with verb-initial order than features associated with verb-final order. A high score for the locus of marking metric reflects greater use of head-marking strategies than dependent-marking strategies. A high score for the fusion metric indicates that a language tends to express grammatical meanings through phonologically bound morphemes (e.g. affixes) rather than freestanding words. Finally, a high score for the flexivity metric indicates that a language has lexically conditioned allomorphy in multiple grammatical or lexical categories (e.g. noun classes, suppletion in lexical forms).

To test whether the patterns captured by component loadings were best described by these specific typological concepts, rather than broader or more narrowly defined phenomena, we created two additional metrics:

- informativity
- noun class/gender

The first of these measures is *informativity*, or the degree to which basic grammatical meanings/functions are obligatorily encoded in the grammar (regardless of how, exactly, these meanings are encoded). This captures how much information needs to be specified when making an utterance in a language. For example, does the language have a rule that tense needs to be marked (regardless of how it is marked)?

The second of these additional metrics encodes *noun class/gender* (i.e. the degree to which a language categorizes nouns into classes/genders, excluding classifiers). The informativity score allows us to ascertain whether our fusion metric is actually capturing a more general tendency for languages to require more types of information to be obligatorily encoded in grammar. The noun class/gender metric allows us to assess the degree to which any latent pattern we observe is driven by flexivity in general versus the more specific phenomenon of noun class/gender, which makes up a large proportion of the features that contribute to the flexivity score. As expected, we find that flexivity is highly correlated with noun class/gender ($r = 0.77$, $p < 0.05$). More importantly, we find that noun class/gender is more strongly associated with PC2 ($r = 0.73$, $p < 0.05$) than the more general flexivity metric ($r = 0.64$, $p < 0.05$), suggesting that the pattern captured by that component relates to the more specific concept of noun class/gender.

The noun class/gender score was calculated in the same manner as the others, but the informativity score was computed in a different way. It was calculated by grouping features which pertain to the same grammatical function (reflexive, passive voice, singular number etc.) and counting that function as present if a language has a positive value for any member of that set. An average was then taken across all available sets for a language, indicating how many of these functions are expressed, either by bound marking or free marking. A language with a low

score for this index encodes fewer types of information obligatorily in grammar, and may express these meanings optionally or lexically. A language with a high informativity score requires non-optional expression of many different grammatical functions.

The code for calculating the theoretical scores is published alongside all other code for this study. The relevant scripts are:

- R_grambank/make_theo_scores.R
- R_gramban/make_theo_score_fusion.R

Wordhood (i.e. what constitutes a word) is a concept that is difficult to converge on globally, and there may be biases among grammar writers that create unnecessary connections between grammaticality and phonological fusion of morphemes in some grammars. To evaluate whether our fusion index truly measures the phonological dependence/independence of grammatical material, rather than a more general tendency to express many types of grammatical meanings, we compared our fusion index to the informativity index. The weak correlation between the informativity score and the fusion score ($r = 0.40$, $p < 0.05$) suggests that the fusion index is not merely a measure of informativity but is actually capturing something interesting about the structure of language (i.e. not the bias of authors).

We take all 6 theoretical scores and compare the score per language to the PCA positions (see Fig. S15). PC1 is strongly correlated with the fusion score, PC2 to noun class/gender and PC3 is not correlated strongly with any score. To test this more robustly we also ran an analysis that controls for phylogeny (see Table S3).

We ran a Phylogenetic Generalized Least Squares-analysis (PGLS, *67*) on each of the first three Principal Components and each theoretical score. This allows us to assess the correlation of each pairing while controlling for shared ancestry as represented by the global language tree (*26*), which is not the case with the simple Pearson correlation matrix in Fig. S15. The Principal Components and theoretical scores were each divided by their standard deviations to make the coefficients easier to compare. Table S3 shows the results. PC1 correlates most strongly with the fusion score. PC2 correlates most strongly with gender/Noun class. PC3 is not strongly correlated with any theoretical score.

Figures S16 and S17 show the position of specific languages within the Indo-European and Austroasiatic language families respectively.

SM1:10 Cultural Fixation Scores

Fixation scores (often abbreviated F_{ST}) are a way of measuring similarity between groups of data in a dataset. It is commonly used in genetics to study how close different groups of individuals are, how the structure compares to what would have happened if everyone mated randomly. The outcome of the analysis is a score for each pairing of groups in your data. A low score indicates that members of those two groups are similar, whereas a high score indicates that they are dissimilar. The value is dependent on both the between-group and within-group variation in the data, as well as the overall frequency of the variable in the entire dataset.

There are several different approaches to fixation scores in the literature. For this study, we used the method proposed by (*68*) which is developed specifically for cultural data. For more on the details of the Cultural Fixation Score and how it differs from other fixation scores, see (*68*).

For the Grambank dataset, we use the groups from the AUTOTYP project (25 cultural-historic areas like "Andean" and "Indic") and the macroareas from Glottolog 4.0. Each language is associated with one of each of these regions, and the pairwise cultural fixation scores indicate how likely it is that two areas should be merged or kept separate. This analysis uses the dialect merged, cropped and binarised dataset (i.e. 1,509 languages and 113 features) – but not imputed data. To illustrate the scores, Fig. S18 shows a barplot of cultural fixation scores over macroareas and Fig. S19 the cultural fixation scores over AUTOTYP areas

To investigate if the AUTOTYP areas that are found in the Americas do indeed form a distinct cluster, we rendered a network based on the cultural fixation scores and computed the modularity score if we group the nodes into Americas vs not Americas (see Fig. S20). We used the function modularity from the *R* package *igraph* (*69*) and the score was -0.061. This indicates that a division Americas vs not Americas is not a neat way of dividing up the relationship between languages of AUTOTYP areas given their pairwise cultural fixation scores.

SM1:11 Unusualness analysis

We define unusualness based on the information-theoretic notion of surprisal. According to this measure, a language is considered to be more unusual the rarer its features and/or combinations of features are cross-linguistically. Concretely, we compute the surprisal associated with each language *i*,

$$U_i = -\log(P_i)$$

where

$$P_i = \Pr(X^1 = x_i^1, X^2 = x_i^2, \ldots, X^{195} = x_i^{195})$$

is the probability of the Grambank description of language *i*. Estimating P_i is complicated by the fact that our sample size is much smaller than the number of possible grammars (i.e. what is referred to as a n<<p scenario in machine learning). We overcome this obstacle by constructing a model-based estimator based on different assumptions about the structure of grammars.

Probability density estimation

For this analysis we used the dialect-merged, cropped, imputed and binarised dataset (see Supplementary Material 1:7), which contains 1,509 languages over 113 features. The possibility space (the number of possible distinct languages in the Grambank description) is 2^{113}. However, our goal is to approximate the probability distribution of the Grambank description of the languages that exist today – and not some theoretical distribution of "possible" or "frequent" languages independent from the finite sample we were able to observe. In this regard, our sample is not negligible, specifically when contrasted to the number of languages for which a comprehensive grammar exists (~4,000). Nevertheless, a direct estimation of the probability distribution is unfeasible as all Grambank descriptions are unique (and we do not want to assume that all the languages not described in Grambank have to be identical to some other Grambank language.) In order to overcome this limitation, we use our understanding of linguistic diversity in order to develop two estimators for this target probability distribution.

Bayesian Latent Class Analysis

Our first estimation model is based on the idea that some of the strongest regularities in grammar are likely to be confined within bundles of features (e.g. word order of the nominal phrase, locus of marking, etc.). The probability of the Grambank description of an unobserved language will thus depend on whether it displays patterns and traits that are regularly found in other languages. Rather than using pre-built categories for the features, we induce hierarchical clustering. The gap statistic indicates an optimal choice of 9 clusters. For each of those clusters we can then identify a discrete and small number of latent classes that more efficiently capture the variation in the data. We implement this through Bayesian Latent Class Analysis. For each bundle of features we find the optimal number of clusters (between 1-6) based on the BIC criterion. For all 9 bundles, a

single cluster turns out to be privileged - which reveals how skewed the representation of different language types is.

Local kernel density estimation

As an alternative to the method developed above, we implemented a method based on locally smoothing the space of attested grammars. The motivation is that a high density of similar Grambank descriptions points to what is probably a *smooth* high probability density region – so that Grambank descriptions of unattested languages which are close to many attested ones will get a high probability. We parametrize this approach by constructing an approximation to the probability distribution with an exponential kernel based on Gower's distance (i.e. the fraction of overall differences between two Grambank descriptions), so that the probability of any specific description is:

$$P_i^k \propto \sum_l \exp(-k \cdot d_{il})$$

Where the summation is carried over all languages of Grambank (parametrized with l), k is the kernel parameter, and d_{il} is Gower's distance between the target Grambank description i and language l. It should be noted that we do not calculate the exact probabilities in this case (as this would require estimating this probability on all possible Grambank descriptions), but just a number that is proportional to it – which is sufficient for the purpose of our analyses.

We studied k = 1, 5, 10, 15, 20, 25, 30 and 40, covering widely differing scales of locality.
In order to gain an intuition of the effect of this parameter choice it is instructive to consider how much the presence of a specific Grambank description contributes to the probability distribution near it. To start with, consider that observing one specific Grambank description contributes to its probability a number proportional to exp(0)=1. Let us use this contribution as the scale of measurement in these following examples. In the broader case (k=1), observing a Grambank description makes even distant languages substantially more likely: languages that are 10%, 20% and even 50% different get a boost of 0.9, 0.8, and 0.6, respectively. So even languages that are as similar as they are different from a given language will still receive a large boost from them. On the other hand, the most local case (k=40), contributes to languages that are 10%, 20% and 50% different (0.02, 0.0003 and 0.000000002 correspondingly). In this scenario, only very similar languages are taken into account when determining the probability of any Grambank description.

Comparison between methods
We compare the Bayesian LCA and the kernel approaches (see Fig. S21).

As it can be appreciated, the Bayesian LCA approach yields almost identical results to those of the least local kernel approaches, suggesting our derived latent classes are not particularly effective at capturing the complexities of the probability distribution at a small scale. The distributions reflect clearly the scale of smoothing: models that learn locally (i.e. have large kernel values) result in a heavy concentration around the highest value of the metric such that most languages are unusual. The opposite pattern holds for the LCA and the models with small kernel values: most languages are concentrated on the lower values of unusualness. Given these findings, for further analyses we pick the estimator yielding the distribution with the least skewness – in other words, the one that does not concentrate languages in either extreme of the scale (which is Kernel 15). Fig. S22 shows the distribution of Unusualness scores (Kernel 15) per language in the world and Fig. S23 shows it as grouped by AUTOTYP areas.

Unusualness model

We deploy a Bayesian regression model of unusualness. The spatial and phylogenetic effects are both variance covariance (vcv) matrices based on a Brownian motion approach. The spatial data is taken from Glottolog (*1*) and the phylogeny is the global language tree (*26*). This is the same method of generating the vcv:s as the INLA modeling (see Supplementary Material 1:8), with the same kappa and sigma values (2 and 1.15 resp) for the spatial vcv. The rest of the analysis is different in that it uses Bayesian Regression Models using 'Stan' (BRMS) rather than Bayesian inference for Latent Gaussian Models (INLA). We use default (uninformative) priors for all coefficients as implemented in the Stan wrapper *brms R* package (*70*). We ran 4 independent chains for 6,000 iterations, and all parameters of the model showed convergence quickly into the run of each chain. A summary of the model parameters can be found in Table S8. The Bayesian R^2 of this model is 0.75 (est. error = 0.02 (*71*)). The posterior predictive distributions of this model (arranged according to cultural-historical areas) can be found in Fig. S24.

SM1:12 Calculation of Manhattan distances

Manhattan distances show the sum total of the number of differences between two records of data, in our case between pairs of languages. For this metric we used the binarized version of the dataset, i.e. each language for each feature had a value of 0 or 1 (or missing). If there are 10 binary features then a Manhattan distance of 4 for a pair of languages would mean that for 4 features they had different values (0 when the other had 1 or vice versa). This measurement is not relative to how many complete pairs of data points there are. If for one feature and one language pair there is at least 1 missing datapoint, that feature is ignored. A Manhattan distance of 0 means for all features the language pair has exactly the same values.

For the calculation in our dataset we used the dialect-merged and cropped, but not imputed version (see Supplementary Material 1:7). There are 113 features in the dataset that is cropped for missing data, meaning that the maximal possible Manhattan distance between any two languages is 113. The highest value found was 74; the pair consisted of the Sino-Tibetan language Wambule [wamb1257] and the Atlantic-Congo language Bobangi [bang1354]. There were 6 language pairs with a distance of 0. In each of these cases, the two languages were from the same language family (see Table S9). The mean distance was 39. A plot of the distribution is found in Fig. S25.

SM1:13 Functional richness

We followed the approach used in ecology where Functional Richness analyses are commonly based on Principal Coordinates Analysis (PCoA, also known as Classical Metric Multi-dimensional Scaling), as this maximizes the amount of the total variation in the dataset that can be captured in two dimensions (here 33%). We calculate this using the *R* package *fundiversity* (*72*). To model endangerment, we use the Agglomerated Endangerment Scale (AES, *73*) and categorize languages as either non-threatened or threatened (the latter of which includes all AES categories associated with endangerment or recent dormancy). Of the languages in Grambank, seven languages had no AES value recorded. To avoid overestimating the effects of endangerment we excluded these languages from the analysis.

SM2 Supplementary figures

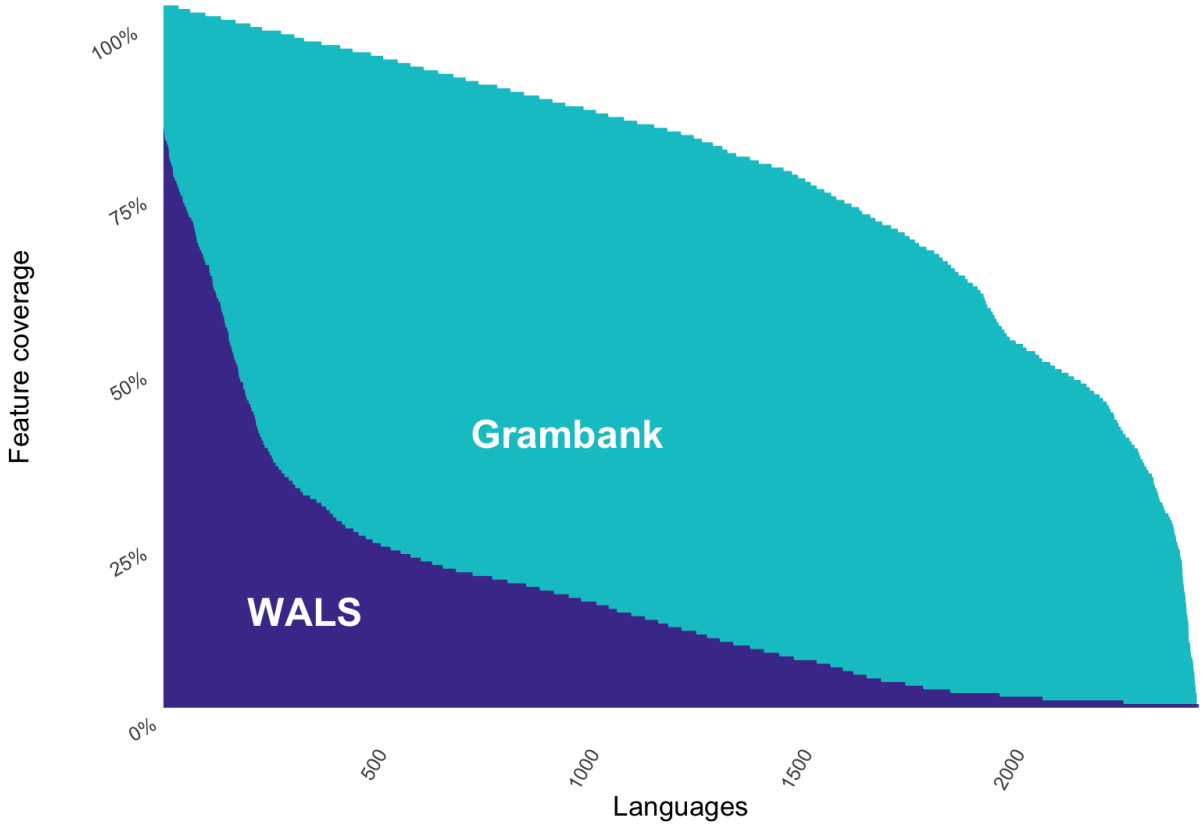

Figure S1. Comparison of coverage per language and feature in WALS and Grambank. This plot shows that the amount of missing data per language is much lower in Grambank compared to WALS. The total number of languages is 2,430 for Grambank and 2,435 for WALS. The numbers are derived on the dialect-aggregated dataset, see Supplementary Material 1:7.

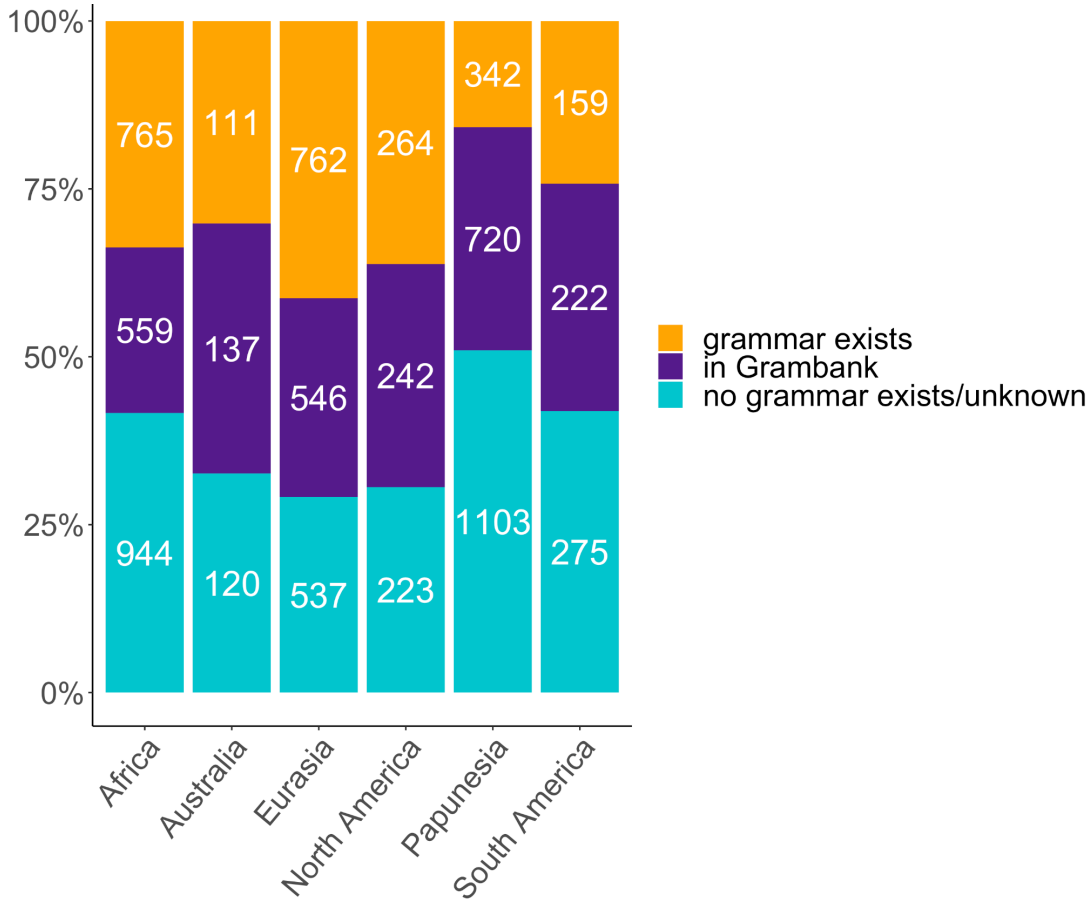

Figure S2. Barplot showing the coverage of languages per Glottolog macroarea. Light blue represents languages which do not yet have a grammar as indexed by Glottolog, dark blue indicates languages that are already in the Grambank database and orange denotes languages which have a grammar indexed in Glottolog but which are not (yet) in the Grambank dataset. Languages in the light blue category are most likely not possible to include in Grambank, whereas the orange category could be included in future. The numbers are derived from the dialect-aggregated dataset, see Supplementary Material 1:7.

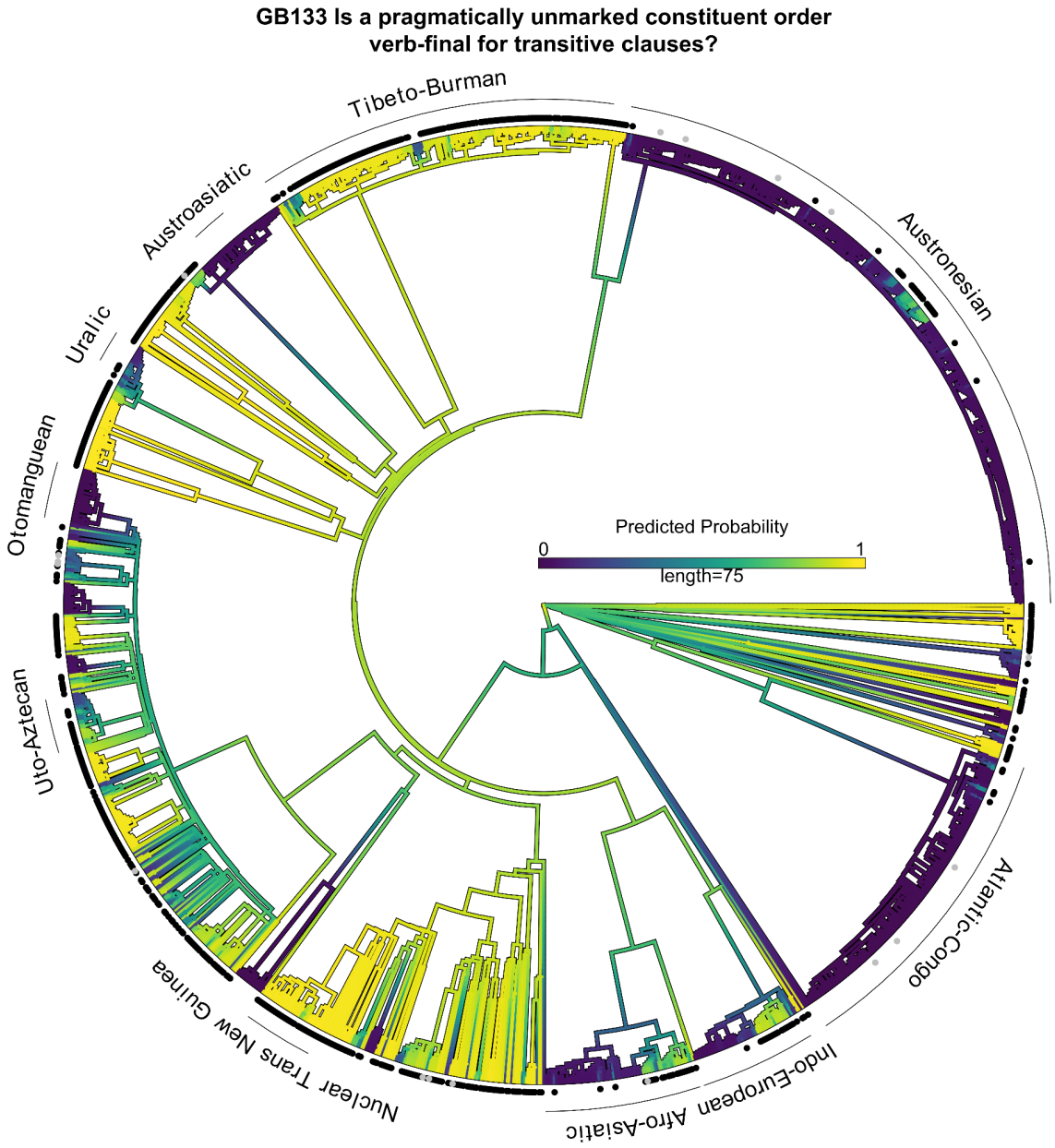

Figure S3. Tree plot of GB133, the Grambank feature with the highest phylogenetic effect in the INLA (dual) model. Tip point colors represent observed values: black = yes (verb-final is a pragmatically unmarked constituent order for transitive clauses), uncolored = no (verb-final is *not* a pragmatically unmarked constituent order for transitive clauses), gray = missing data. Branch colors represent probability estimates: yellow = higher probability that verb-final is a pragmatically unmarked constituent order for transitive clauses, purple = lower probability that verb-final is a pragmatically unmarked constituent order for transitive clauses.

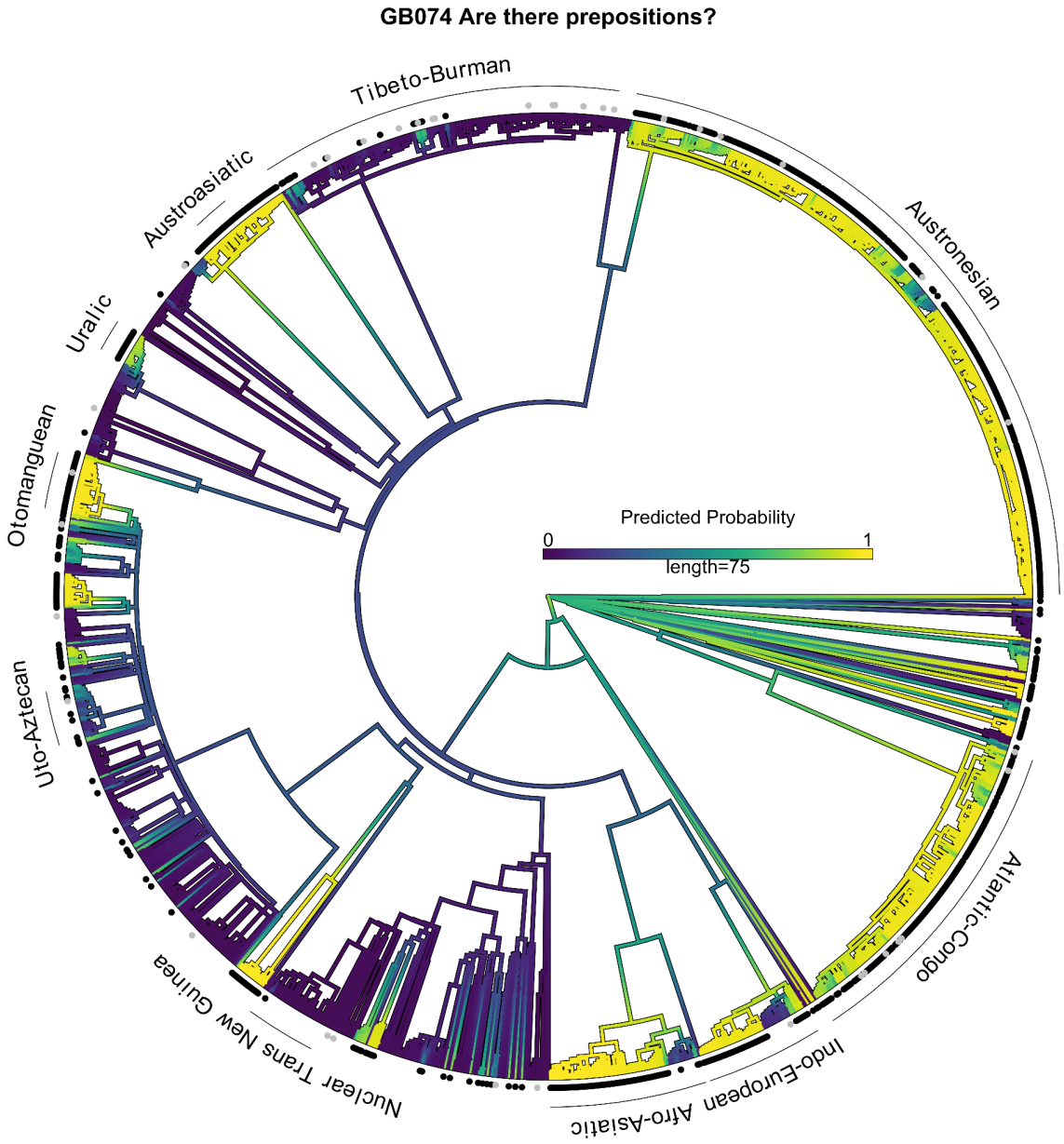

Figure S4. Tree plot of GB074, the Grambank feature with the second highest phylogenetic effect in the INLA (dual) model. Tip point colors represent observed values: black = yes (there are prepositions), uncolored = no (there are *not* prepositions), gray = missing data. Branch colors represent probability estimates: yellow = higher probability that there are prepositions, purple = lower probability that there are prepositions.

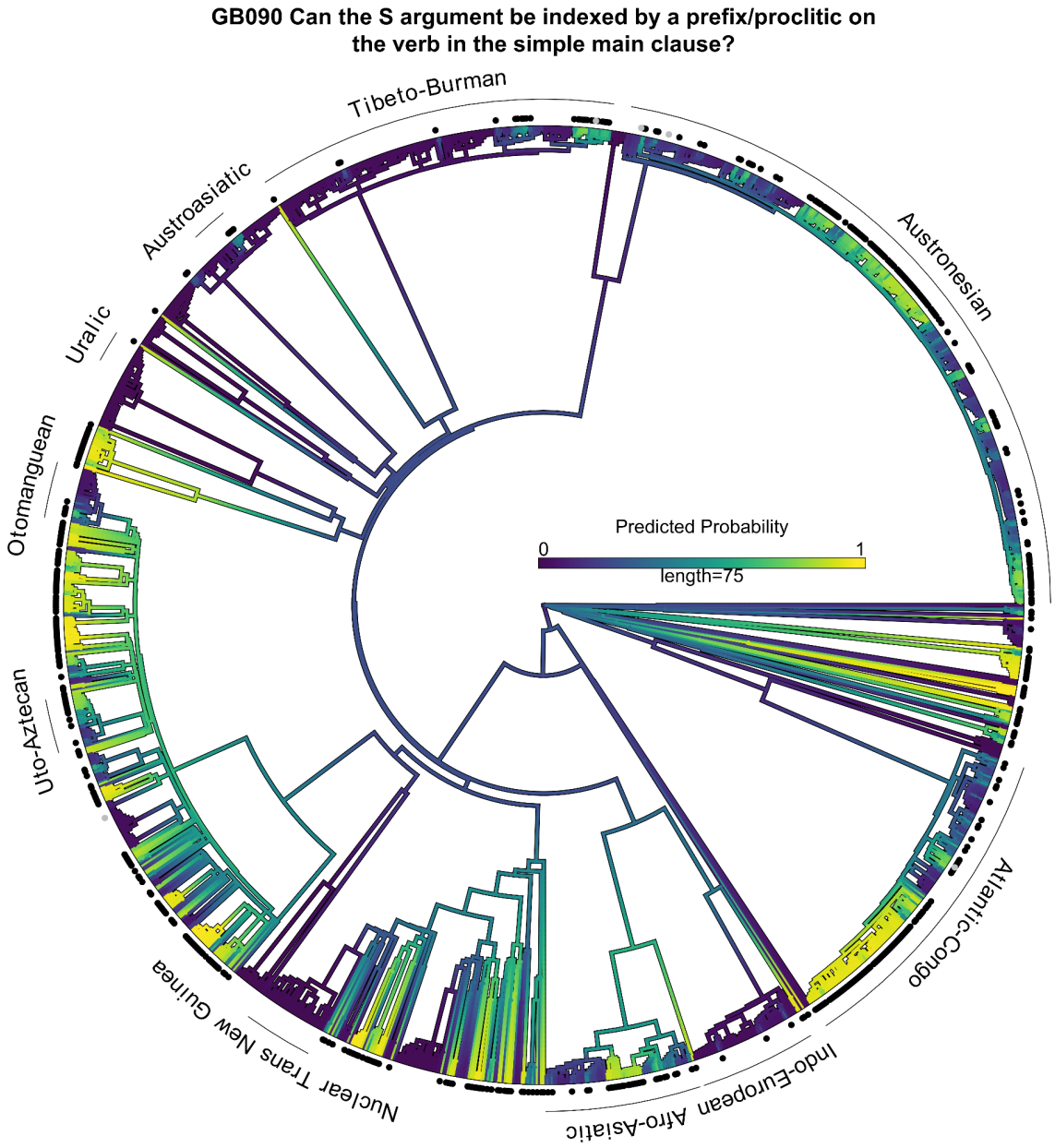

Figure S5. Tree plot of GB090, the Grambank feature with the third highest phylogenetic effect in the INLA (dual) model. Tip point colors represent observed values: black = yes (the S argument can be indexed by a prefix or proclitic on the verb in simple main clauses), uncolored = no (the S argument can *not* be indexed by a prefix or proclitic on the verb in simple main clauses), gray = missing data. Branch colors represent probability estimates: yellow = greater probability that the S argument can be indexed by a prefix or proclitic on the verb in simple main clauses, purple = lower probability that the S argument can be indexed in this way.

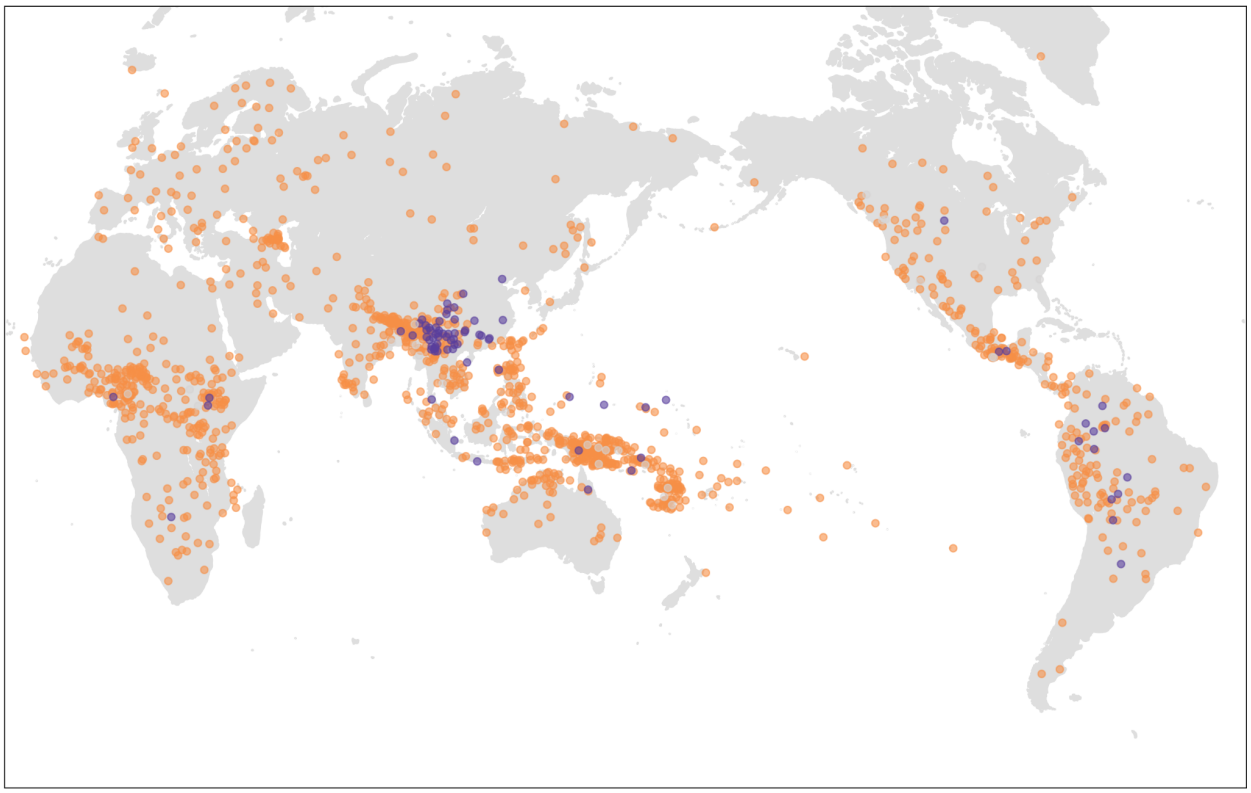

Figure S6. Map of GB038, the Grambank feature with the highest spatial effect in the INLA (dual) model. Purple indicates languages that have demonstrative classifiers; Orange indicates languages that do *not* have demonstrative classifiers.

GB080 Do verbs have suffixes/enclitics, other than those that only mark A, S or P (do include portmanteau: A & S + TAM)?

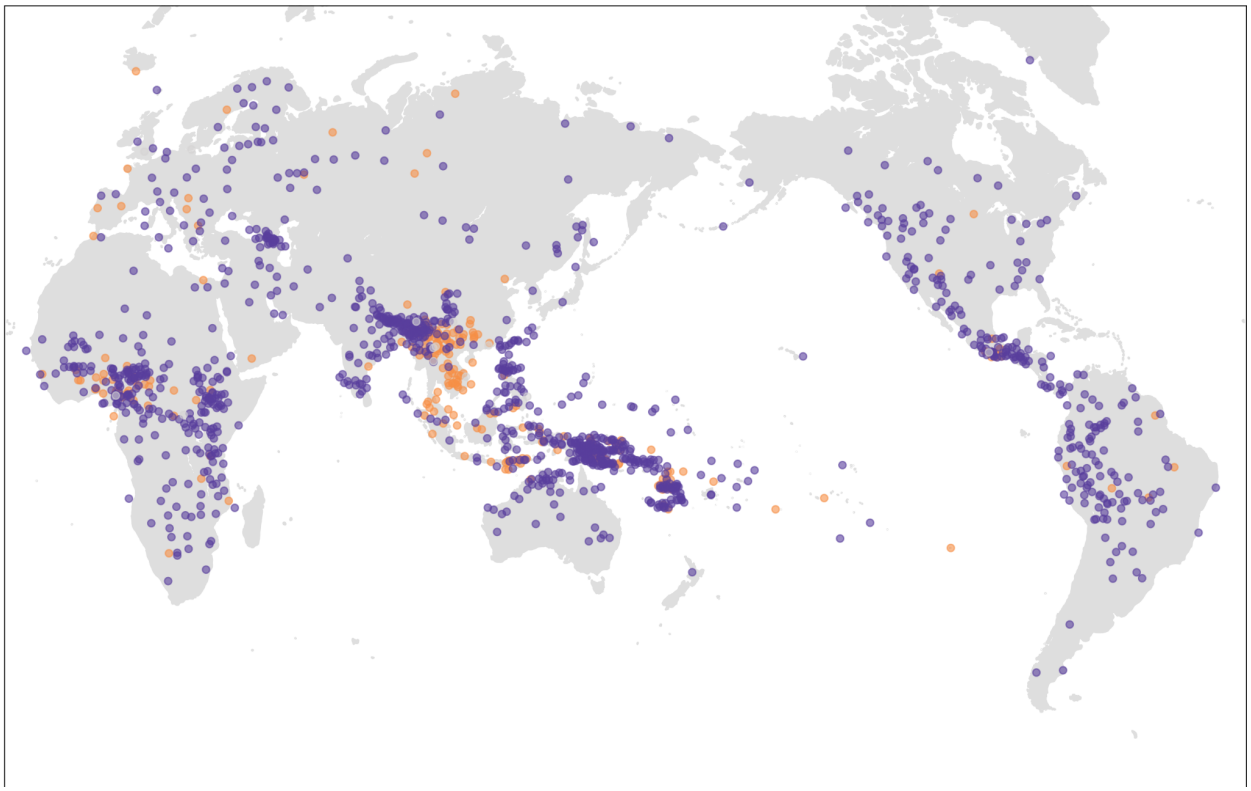

Figure S7. Map of GB080, the Grambank feature with the second highest spatial effect in the INLA (dual) model. Purple indicates languages that have suffixes or enclitics that encode information other than the categories listed in the feature; Orange indicates languages that do *not* have such suffixes or enclitics.

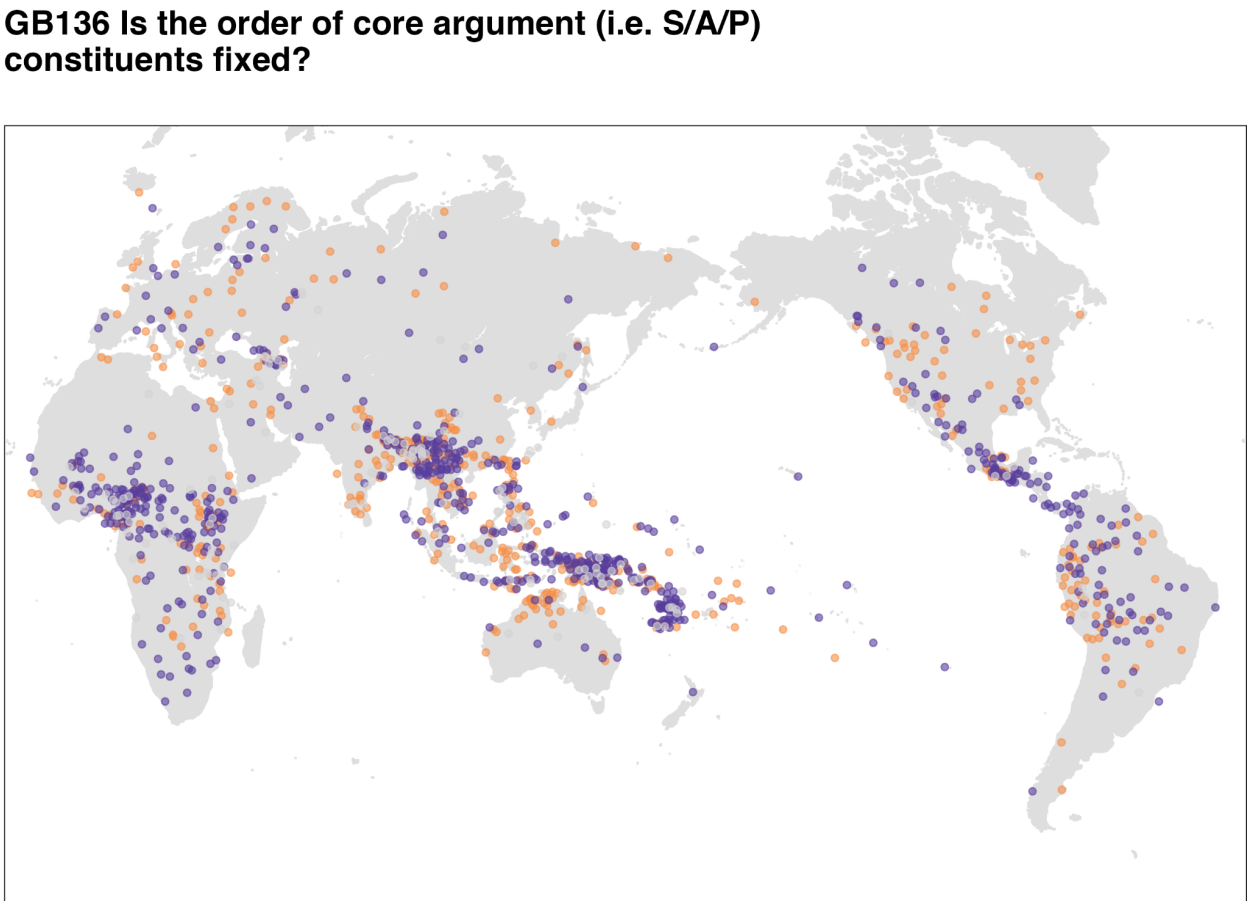

Figure S8. Map of GB136, the Grambank feature with the third highest spatial effect in the INLA (dual) model. Purple indicates that fixed word order occurs in the language; Orange indicates that fixed word order does *not* occur in the language.

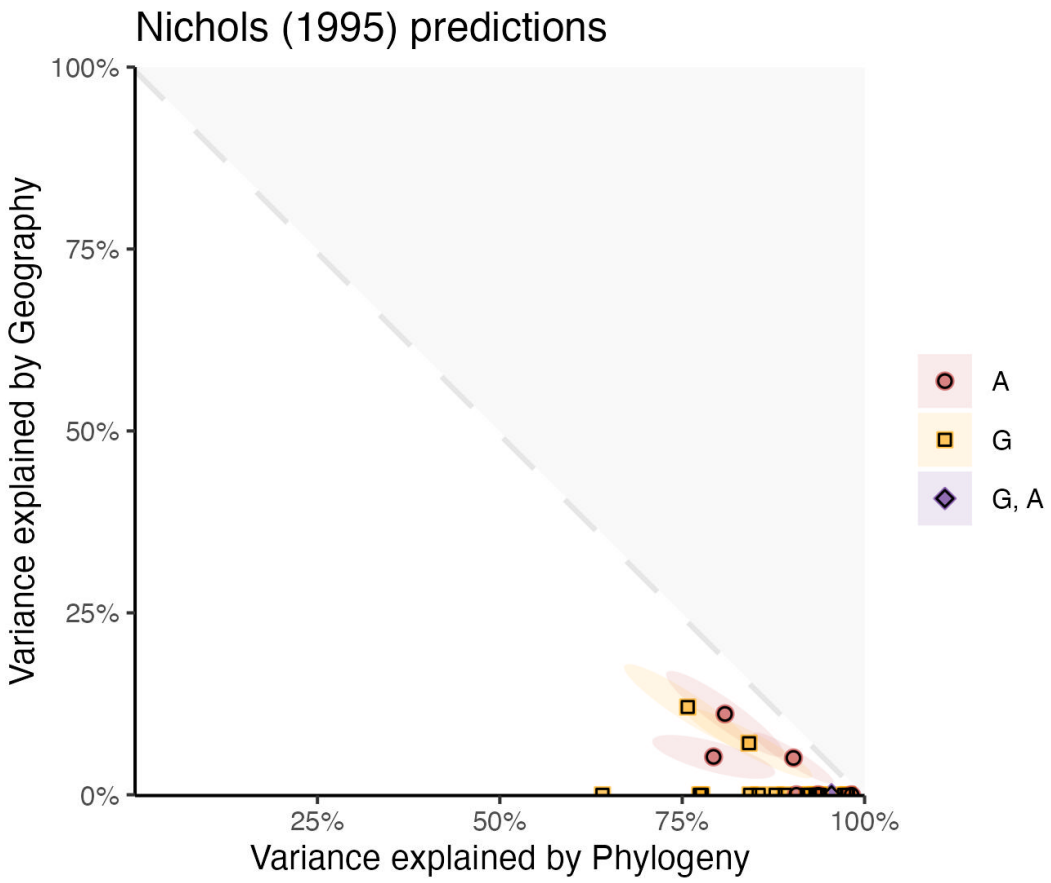

Figure S9. Scatterplot of the phylogenetic (x-axis) and spatial effects (y-axis) for features included in Nichols (1995). The points are colored for the prediction by Nichols: A = Areal, G = Genetic and G, A = Both. The term *genetic* here is used by Nichols (*20*) in a similar/identical fashion to how we have used *phylogenetic* in this paper.

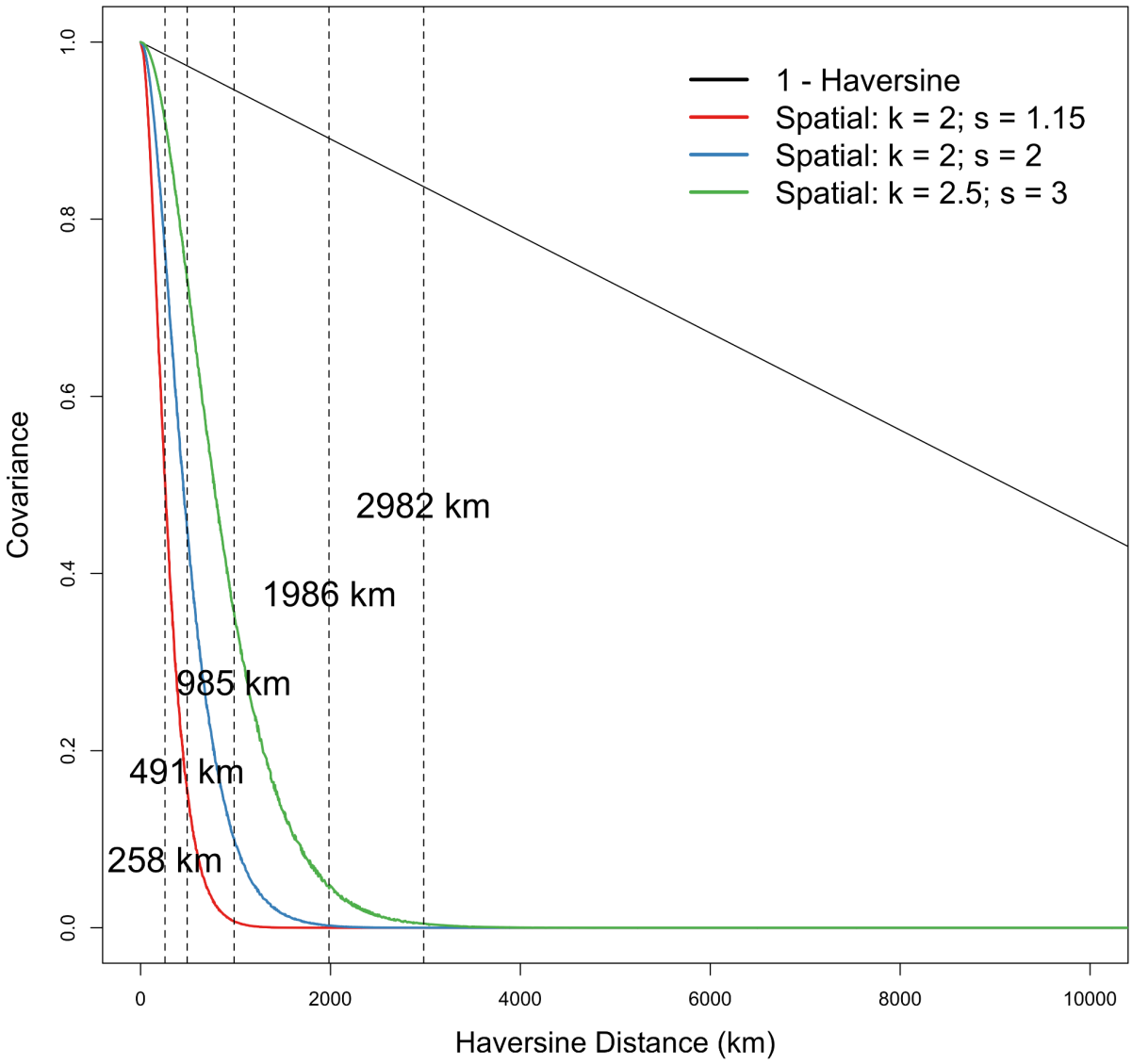

Figure S10 Spatial decay in precision matrices for spatiophylogenetic analysis. This figure shows the relative decay in covariance based on the various parameterisations of the Matérn function. The x-axis shows Haversine distance ("as-the-crow-flies" distances, taking into account the curvature of the earth), and is shown on the y-axis with the black line for reference. The red line indicates the parameterization of spatial covariance used in the main text. Blue, and green lines show parameterizations that iteratively increase the relationship of geography between languages in the model. Vertical dotted lines ground the covariance functions in real-world distances to give a sense of at what point geographic relationships are no longer statistically relevant in this model.

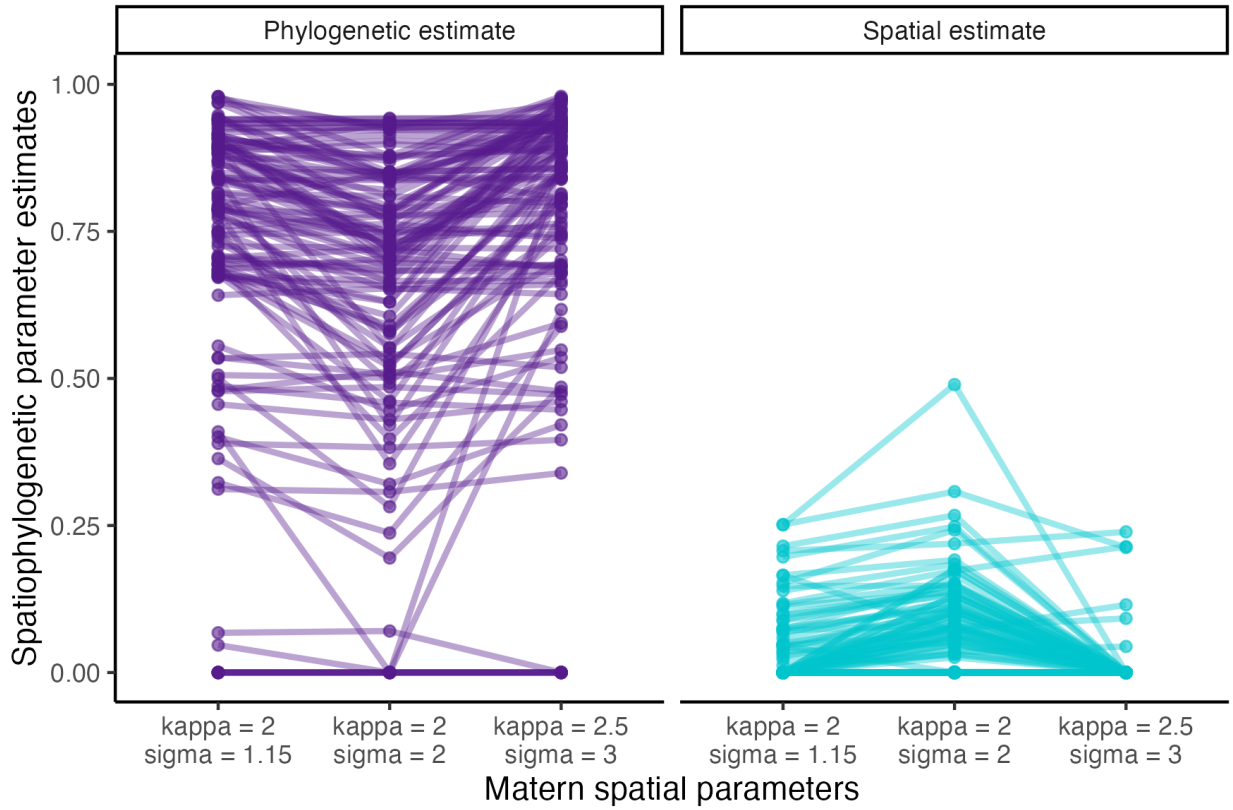

Figure S11: Spatiophylogenetic parameter estimates for the effect of language (left) or geography (right) when varying Matéern spatial decay parameter. Decay functions cause the spatial influence of languages to be effectively zero at approximately 1000km, 2000km, and 3000km moving from left to right on the x axis. Increasing the influence of spatial effect generally has little influence on the conclusions drawn.

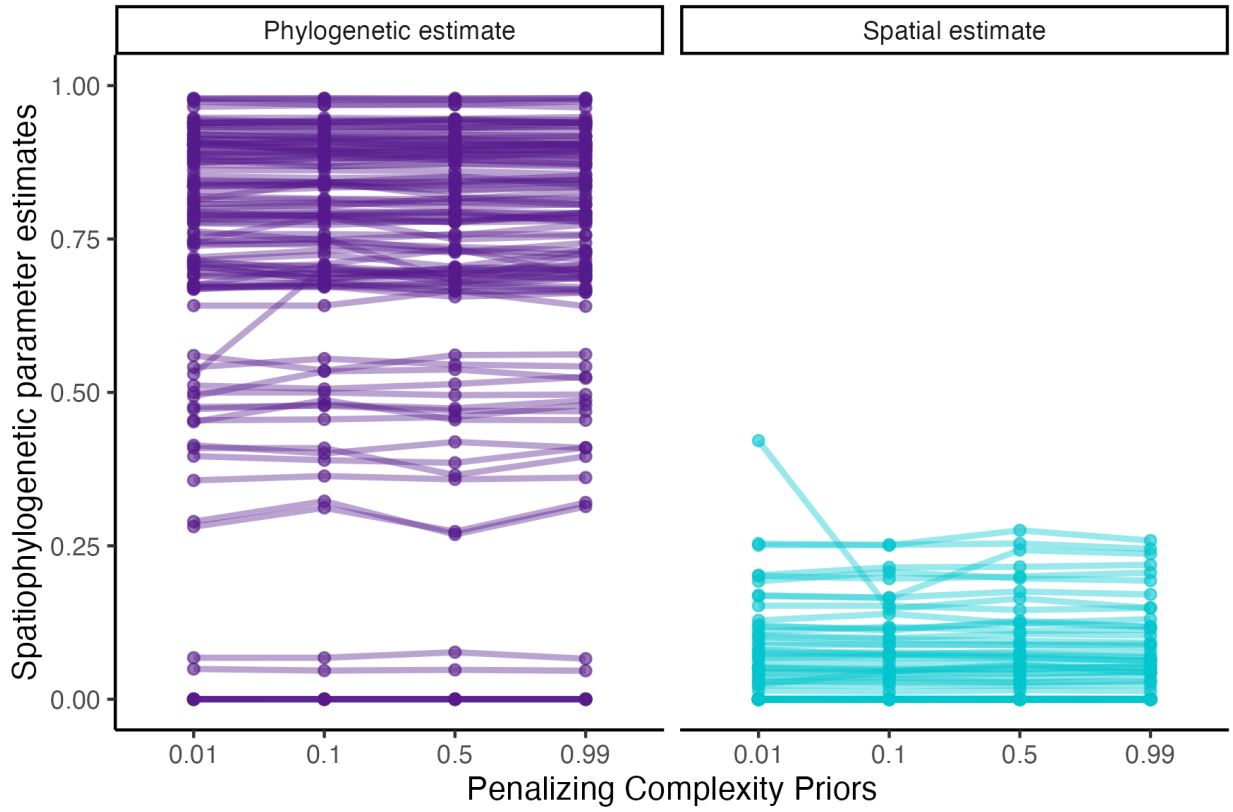

Figure S12: Varying Priors for Penalizing Complexity in the INLA-analysis.

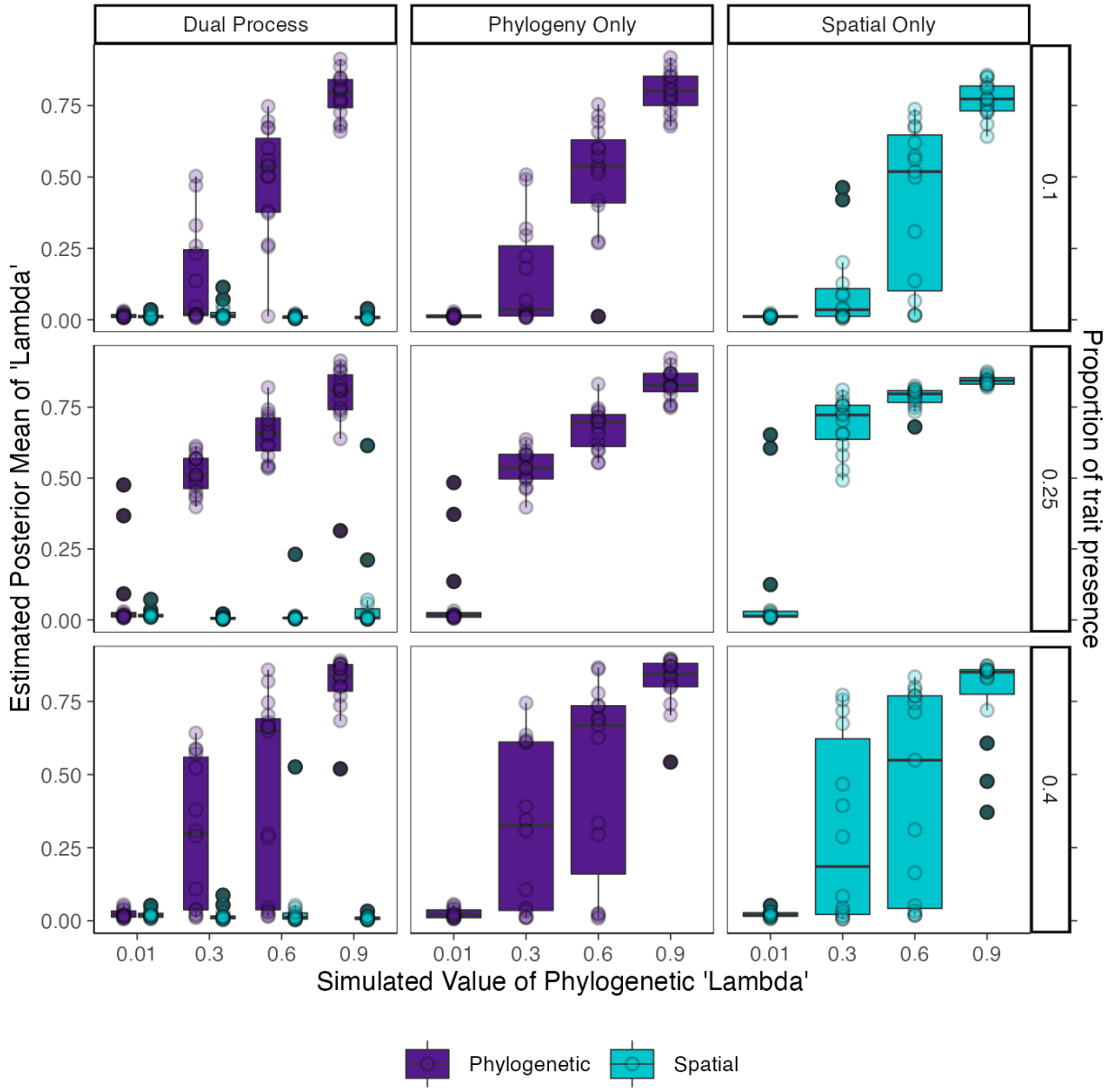

Figure S13: Simulation results for the 12 conditions (four levels of phylogenetic signal, for three different proportions of traits). Each column of graphs contains the results for a particular model structure, each row of graphs contains the results for a particular proportion of traits, and within each graph shows the results across the four levels of phylogenetic signal. The dual process model contains two boxplots per level of phylogenetic signal, one representing the posterior mean for the phylogenetic effect, and one for the posterior mean of the spatial effect.

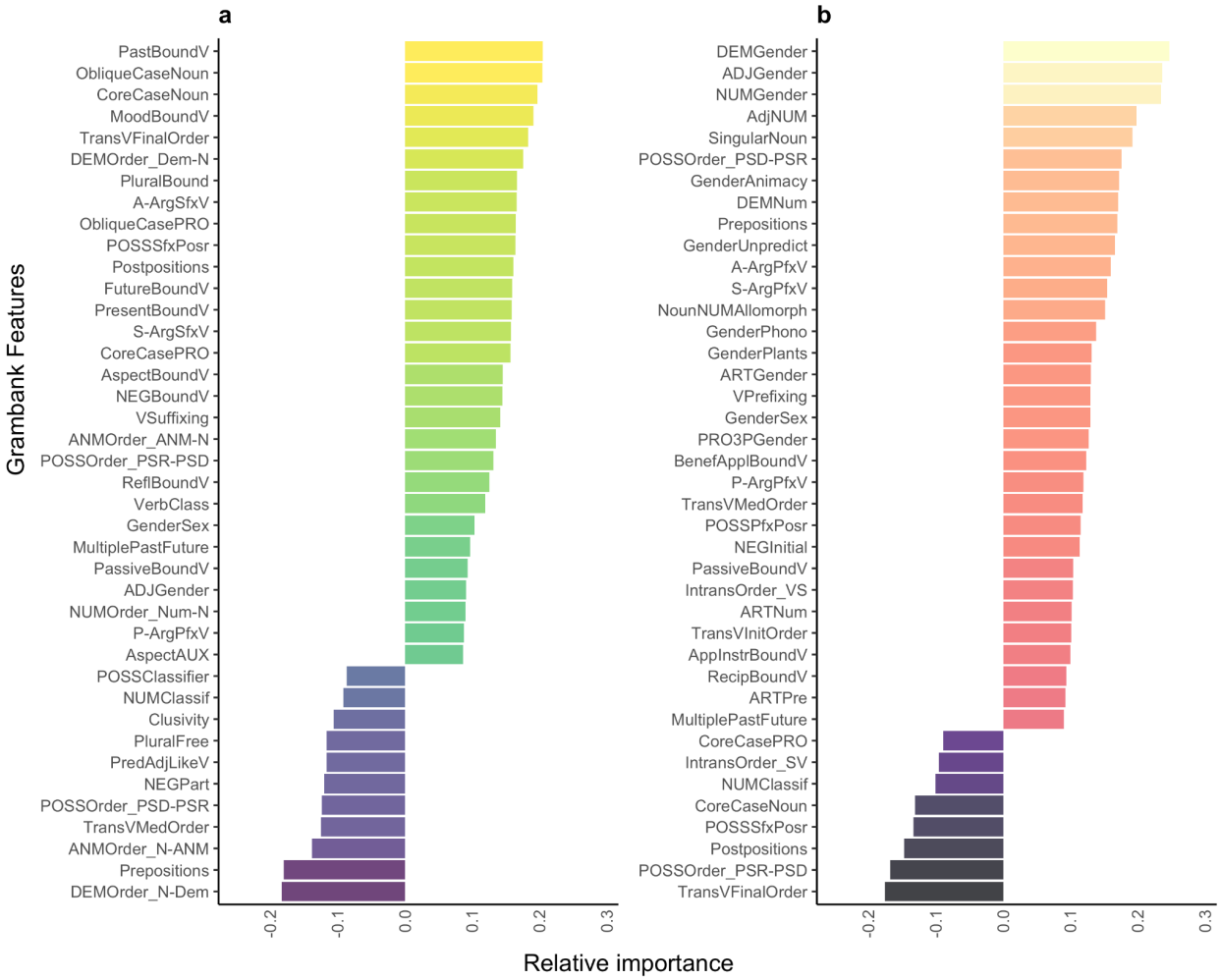

Figure S14. Feature loadings onto PC1 and PC2, including only the top 40 most contributing features.

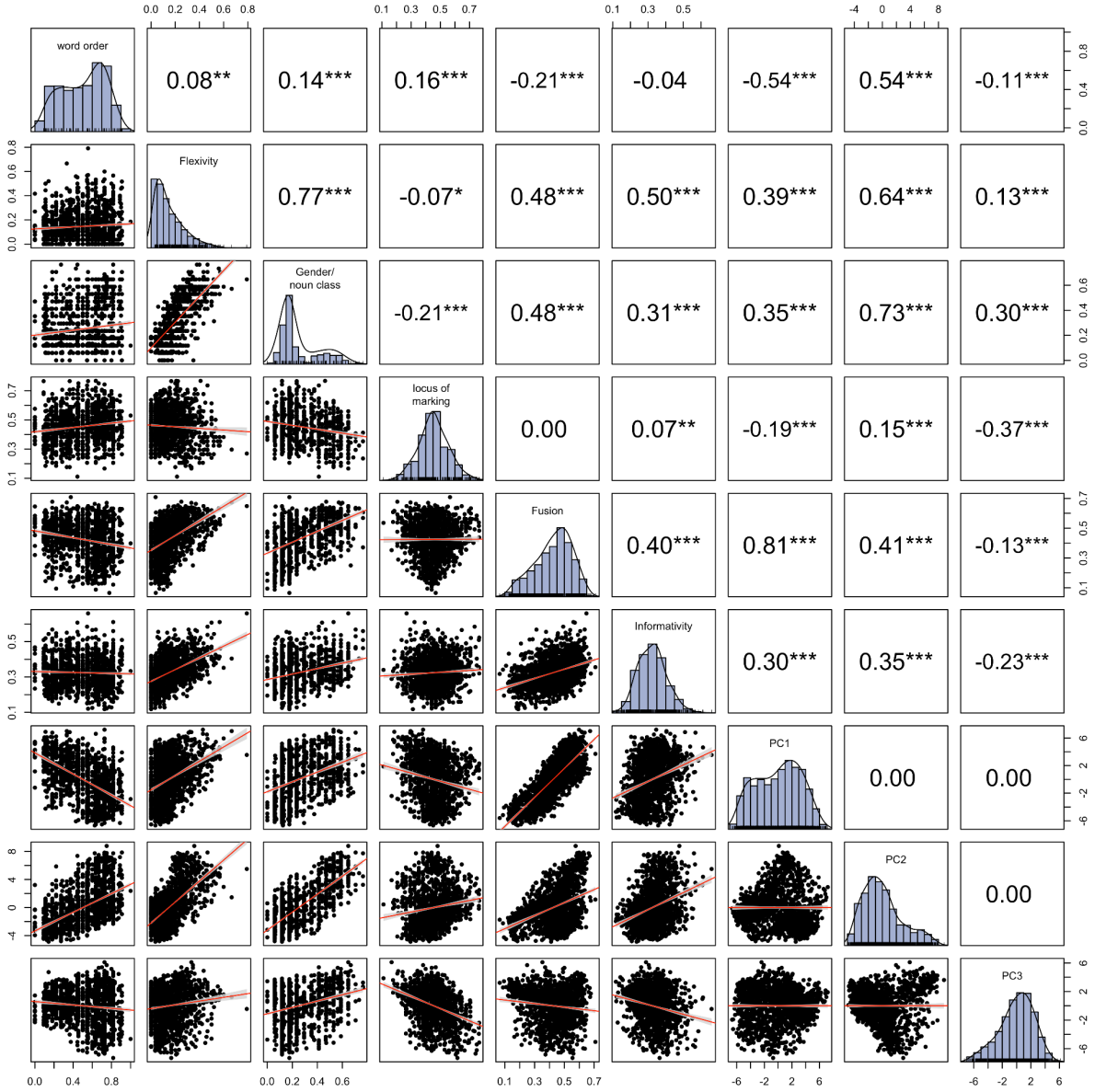

Figure S15: Scatterplot matrix showing the Pearson correlations between the first three principal components of the data and the theoretical metrics.

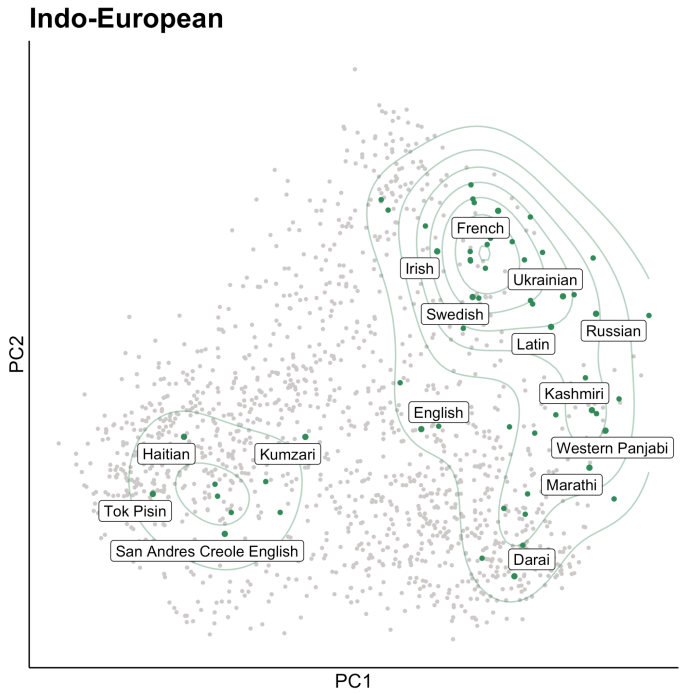

Figure S16. Scatterplot of Indo-European languages (green) among all other languages (grey) and their position given PC1 and PC2 with specific languages highlighted with names.

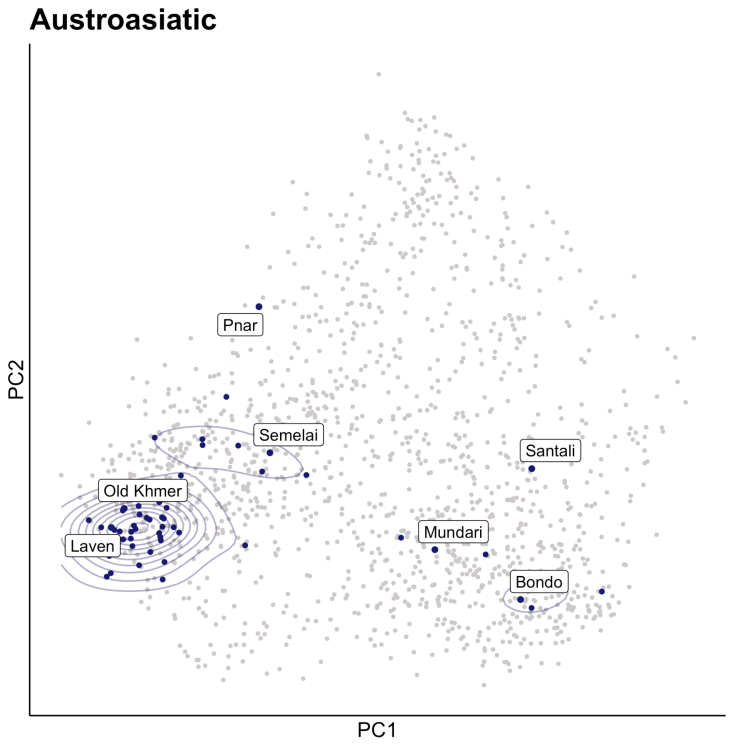

Figure S17. Scatterplot of Austroasiatic languages (blue) among all other languages (gray) and their position given PC1 and PC2 with specific languages highlighted with names. The two major clusters in the Austroasiatic family correspond to languages inside and outside of the Indian subcontinent.

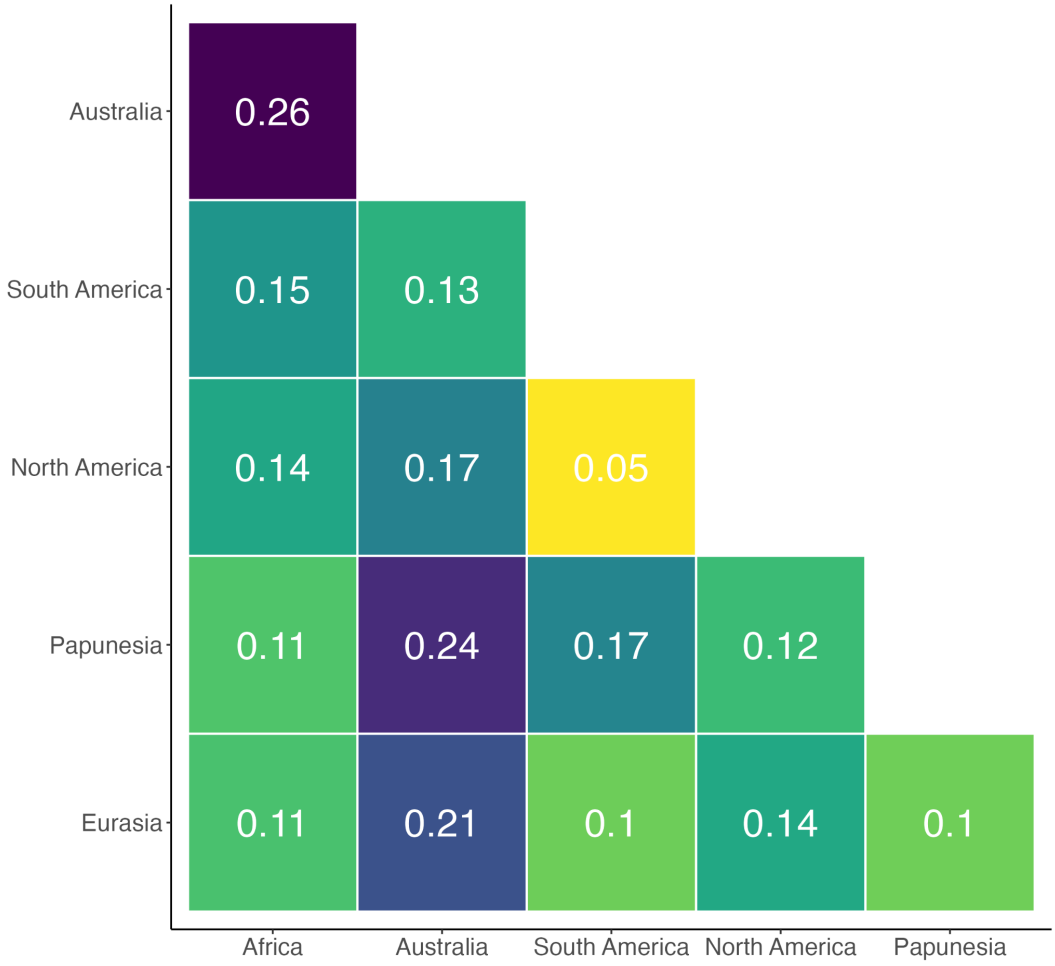

Figure S18. Pairwise Cultural Fixation scores over macroareas in the Grambank dataset. The pair with the lowest score, and therefore most likely to be similar, is North and South America.

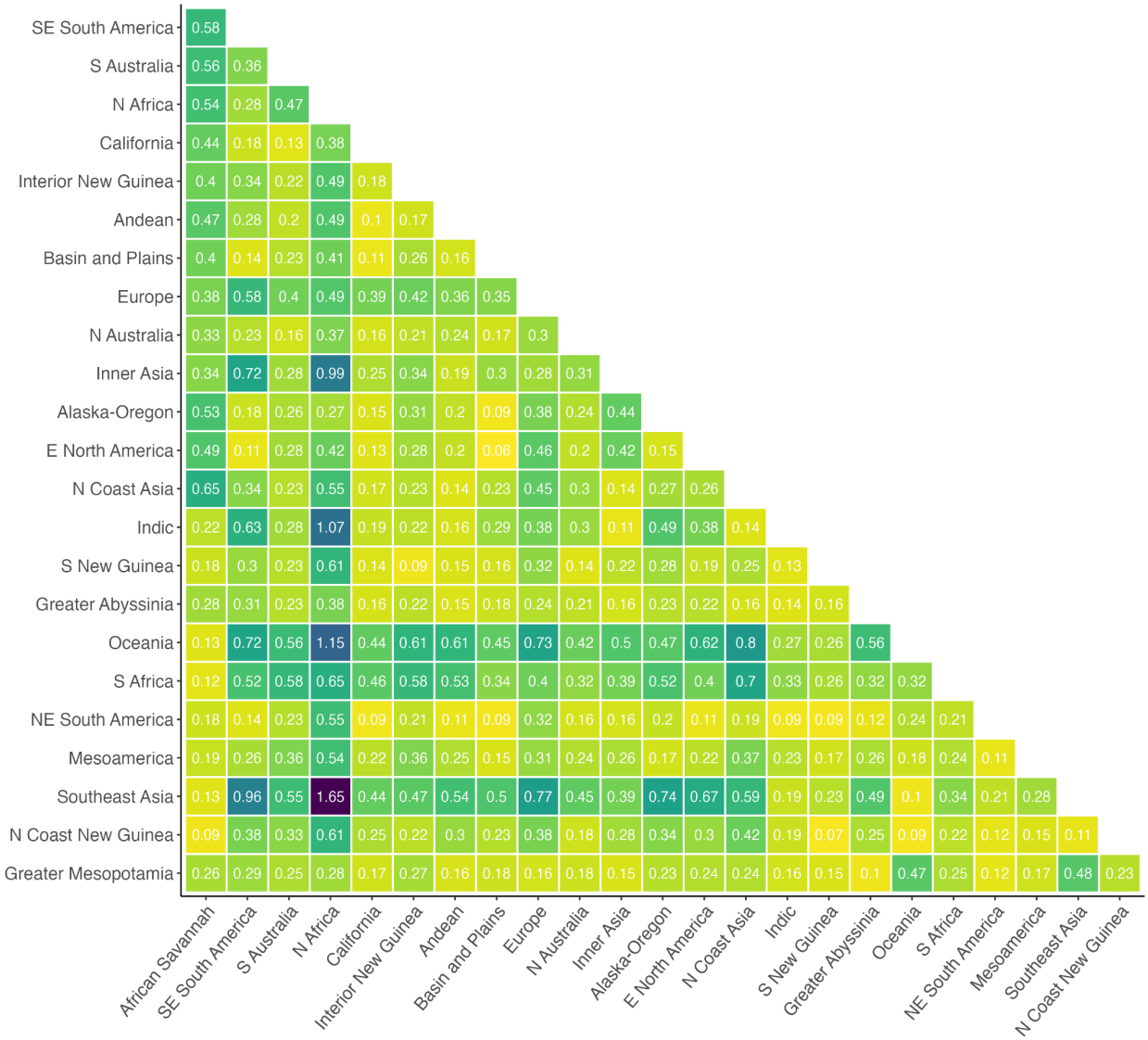

Figure S19. Pairwise Cultural Fixation scores over AUTOTYP-areas in the Grambank dataset.

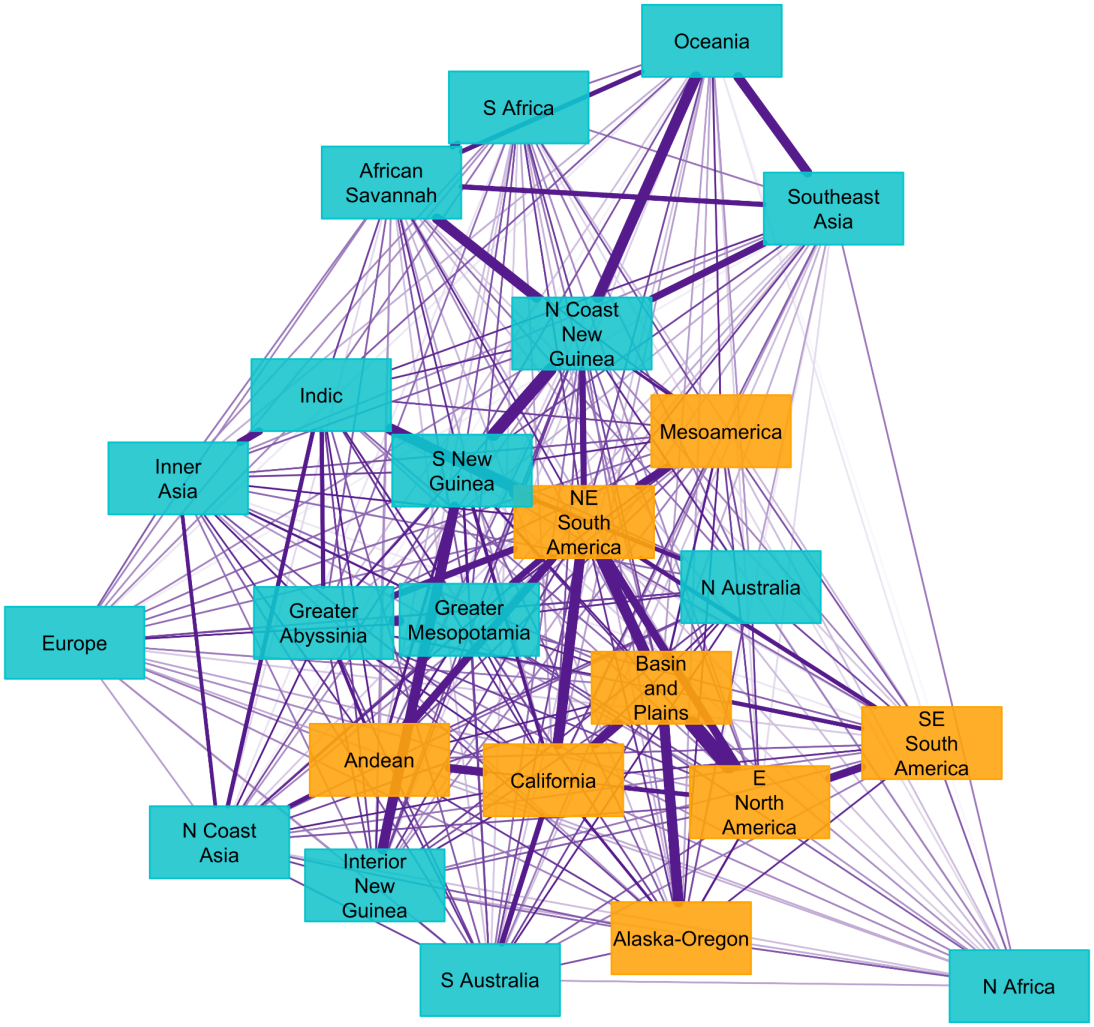

Figure S20. Network visualization of grammatical affinity between linguistic regions of the world. Languages are grouped by AUTOTYP areas, with areas in the Americas (orange) and areas elsewhere in the world (turquoise) represented in boxes. The thickness of lines between nodes indicates the strength of the affinity between areas, i.e. a thicker line indicates a lower Cultural Fixation score.

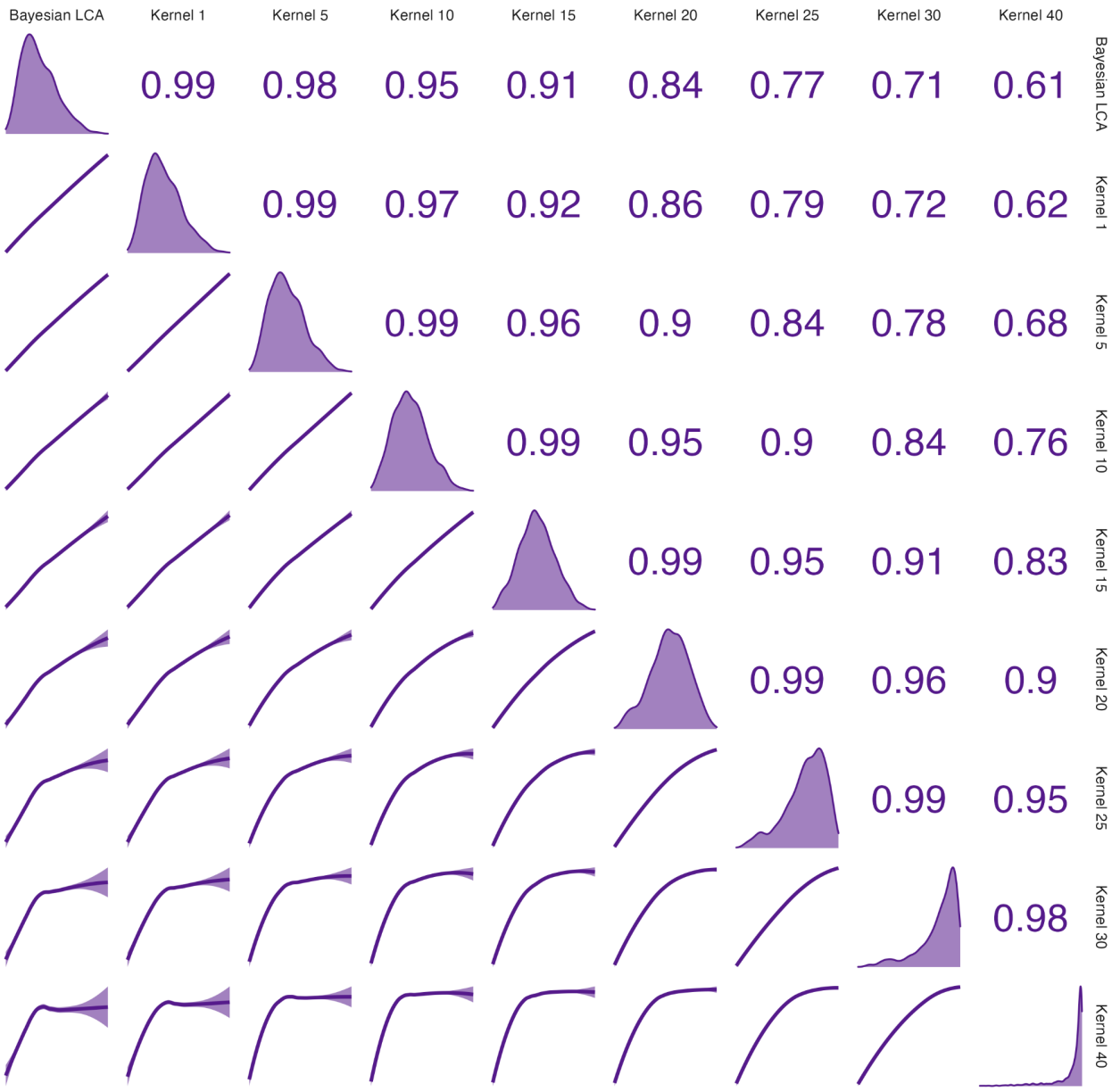

Figure S21. Comparison between different unusualness probability density estimation approaches. Each column/row corresponds to individual estimators. Lower triangle panels show smooth loess curves. Panels on the diagonal show probability densities. Upper triangle panels show Spearman correlation values.

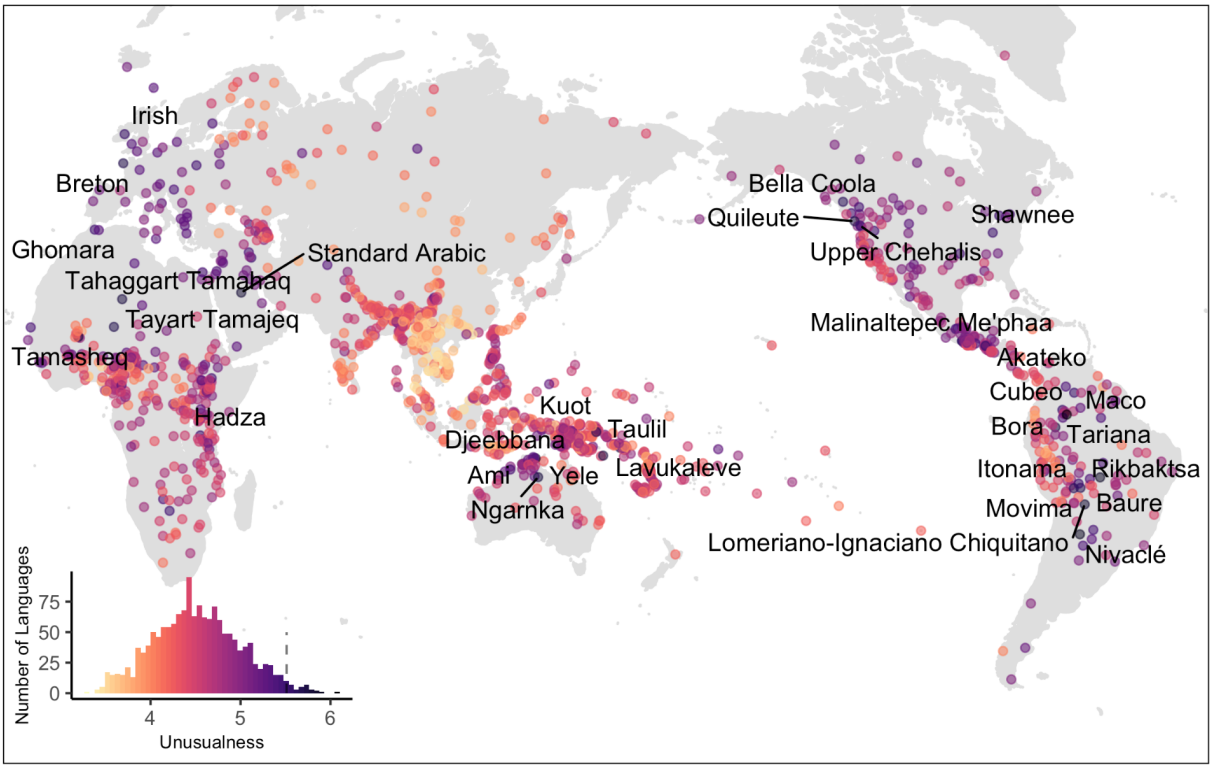

Figure S22. Map displaying the languages with the most unusual feature values. Languages are colored by how unusual their feature values are, and extreme languages are labeled. The inset histogram shows the overall distribution of unusualness scores across all the languages in Grambank, with the dashed line representing the cut-off limit to the top 2% used to identify the most unusual languages (labeled). This analysis uses Kernel 15.

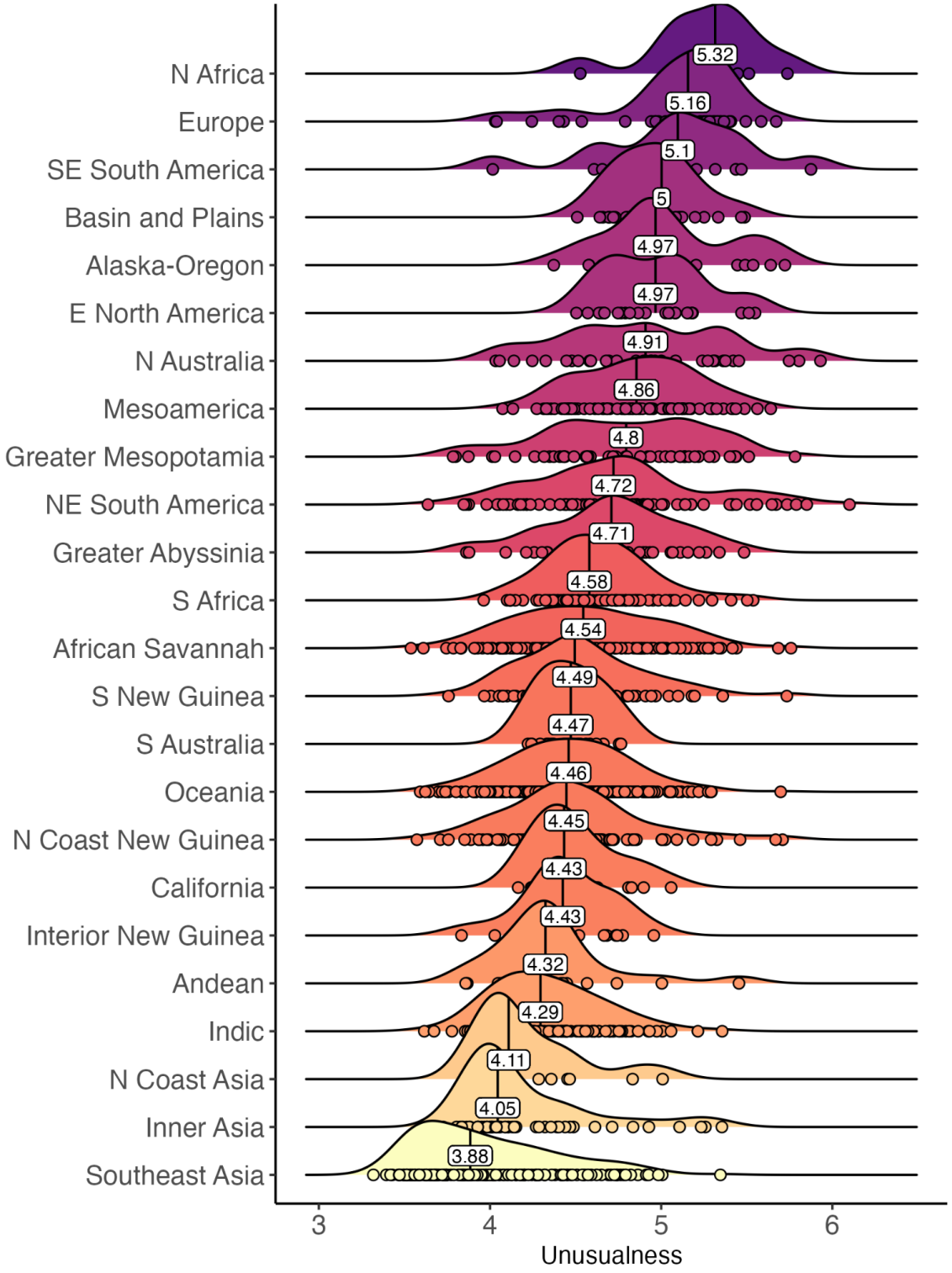

Figure S23. Distributions of unusualness scores (Kernel 15) per language as grouped by AUTOTYP area. The points represent each language and a value far to the right is more unusual. The line in each distribution and the label represents the median value per group.

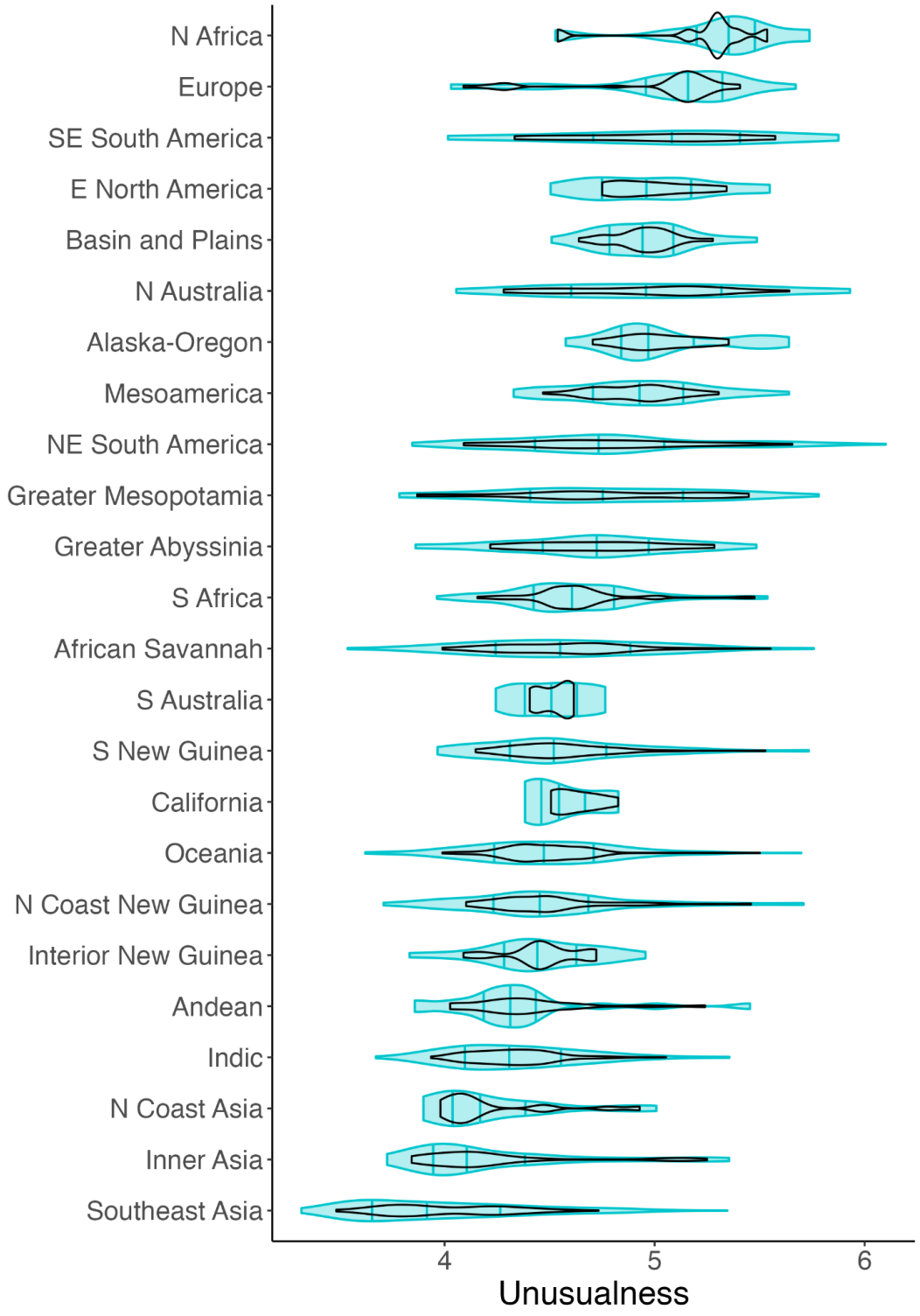

Figure S24. Performance of the model for unusualness, displayed over cultural-historical areas. Light blue violin plots correspond to the unusualness score that the model predicts (samples from the posterior predictive distribution of the model), whereas the black-countour violin plots represent the known unusualness scores - the response variable.

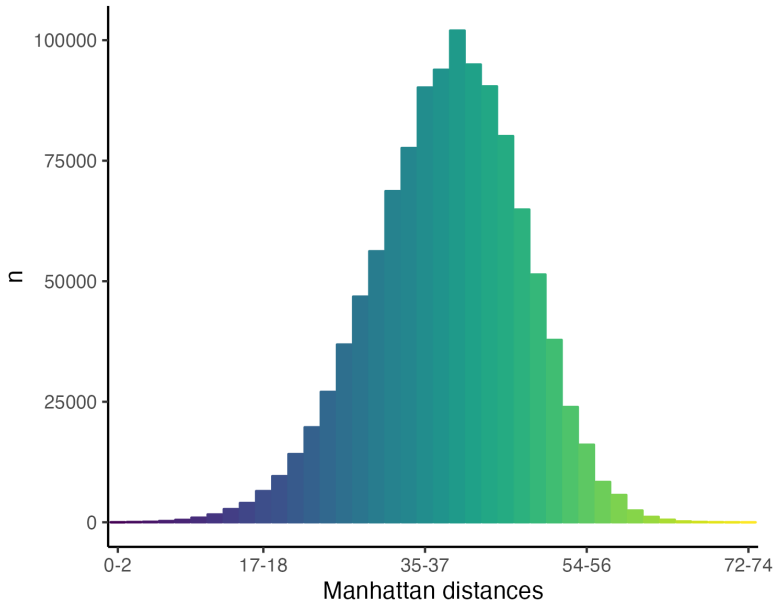

Figure S25. Histogram of Manhattan distances between languages in Grambank. Pairwise Manhattan distances show for each pair of languages in the dataset how many times they had different values, in absolute terms. The height of the bars show how many pairs of languages had that particular Manhattan distance. A Manhattan distance of 0 means that there were no features for which that language pair had different values. The mean Manhattan distance in the entire dataset is 39.

SM3 Supplementary tables

Table S1 Model fit scores (WAIC) of BRMS models with a beta-distribution prediction the mean spatial and phylogenetic effects of Grambank features.

Model	WAIC	SE (WAIC)
null model (spatial)	-1424.07	85.86
domain model (spatial)	-1424.52	83.23
null model (phylogenetic)	-31.07	8.43
domain model (phylogenetic)	-25.64	8.48

Table S2. Phylogenetic and spatial effect in INLA model per feature.

Feature_ID	Phylogenetic effect (mean)	Phylogenetic effect (Standard Deviation)	Spatial effect (mean)	Spatial effect (Standard Deviation)
GB133	0.982	0.004	0	0
GB074	0.977	0.006	0	0
GB090	0.976	0.005	0	0
GB092	0.972	0.004	0	0
GB065a	0.962	0.008	0	0
GB057	0.955	0.018	0	0
GB031	0.948	0.01	0	0
GB043	0.941	0.024	0	0
GB094	0.941	0.017	0	0
GB075	0.939	0.01	0	0
GB171	0.937	0.02	0	0
GB431	0.936	0.019	0	0
GB089	0.933	0.016	0	0
GB091	0.933	0.015	0	0
GB081	0.93	0.022	0	0
GB058	0.926	0.026	0	0
GB196	0.926	0.039	0	0
GB198	0.925	0.027	0	0.001
GB170	0.921	0.025	0	0
GB025b	0.919	0.024	0	0
GB083	0.917	0.019	0	0
GB079	0.916	0.018	0	0

GB070	0.915	0.02	0	0
GB109	0.915	0.023	0	0
GB104	0.913	0.027	0	0
GB433	0.906	0.022	0	0
GB036	0.904	0.05	0	0
GB093	0.902	0.026	0.015	0.01
GB103	0.902	0.021	0	0
GB131	0.902	0.037	0.051	0.023
GB030	0.9	0.023	0	0
GB072	0.9	0.019	0	0
GB193b	0.893	0.052	0.032	0.027
GB051	0.891	0.039	0	0
GB059	0.89	0.024	0	0
GB022	0.881	0.05	0.028	0.016
GB172	0.881	0.054	0	0
GB108	0.879	0.033	0.001	0.002
GB028	0.878	0.035	0	0
GB114	0.869	0.039	0	0
GB086	0.861	0.029	0	0
GB053	0.855	0.038	0	0
GB193a	0.852	0.048	0.043	0.02
GB042	0.85	0.049	0	0
GB116	0.846	0.06	0	0

GB024b	0.843	0.045	0.095	0.032
GB044	0.843	0.037	0	0
GB155	0.842	0.058	0.071	0.032
GB318	0.838	0.052	0	0
GB130a	0.826	0.067	0.102	0.048
GB111	0.809	0.045	0	0
GB132	0.809	0.053	0.111	0.039
GB082	0.805	0.043	0	0
GB115	0.805	0.042	0.043	0.015
GB107	0.803	0.048	0.04	0.025
GB185	0.803	0.045	0	0
GB110	0.801	0.06	0	0
GB186	0.801	0.075	0	0
GB312	0.796	0.065	0.067	0.03
GB020	0.794	0.052	0.031	0.037
GB113	0.793	0.056	0.052	0.02
GB071	0.786	0.046	0.039	0.016
GB149	0.785	0.075	0	0
GB065b	0.784	0.044	0.093	0.029
GB192	0.777	0.087	0	0
GB054	0.774	0.09	0	0
GB147	0.758	0.058	0.121	0.039
GB096	0.753	0.087	0	0

GB068	0.745	0.056	0	0
GB117	0.743	0.066	0	0
GB309	0.735	0.062	0	0
GB024a	0.729	0.087	0.192	0.07
GB035	0.727	0.07	0	0
GB105	0.726	0.064	0	0
GB299	0.714	0.082	0.054	0.025
GB177	0.701	0.096	0	0
GB317	0.7	0.177	0	0
GB432	0.7	0.071	0.095	0.036
GB120	0.697	0.066	0	0
GB130b	0.696	0.099	0.154	0.061
GB025a	0.69	0.084	0.15	0.057
GB184	0.689	0.069	0	0
GB099	0.682	0.122	0	0
GB021	0.676	0.082	0	0
GB073	0.674	0.075	0.071	0.026
GB039	0.657	0.079	0	0
GB138	0.655	0.098	0.08	0.04
GB321	0.65	0.11	0	0
GB052	0.641	0.134	0	0
GB084	0.62	0.077	0.081	0.027
GB137	0.555	0.093	0.194	0.053

GB298	0.542	0.093	0	0
GB023	0.522	0.145	0.144	0.07
GB158	0.511	0.082	0	0
GB095	0.508	0.115	0	0
GB121	0.475	0.095	0	0
GB119	0.473	0.1	0	0
GB098	0.468	0.16	0	0
GB430	0.458	0.155	0	0
GB313	0.412	0.118	0	0
GB038	0.401	0.177	0.265	0.111
GB069	0.396	0.13	0	0
GB080	0.359	0.108	0.252	0.075
GB139	0.292	0.104	0.13	0.046
GB316	0.282	0.153	0	0
GB037	0.068	0.051	0	0
GB136	0.035	0.046	0.205	0.044
GB129	0	0	0	0
GB165	0	0	0	0
GB166	0	0	0	0
GB197	0	0	0	0
GB319	0	0	0	0
GB320	0	0	0	0

Table S3: Correlation coefficients of association between Principal Components and Theoretical scores, as calculated by PGLS.

PC	Theoretical score	coef	t-value	p-value (of t)
PC1	Word order	-0.09014	-4.77918	0
PC1	Flexivity	0.14755	9.42063	0
PC1	Noun class/gender	0.16118	8.06301	0
PC1	Locus of marking	-0.02264	-1.78043	0.07522
PC1	Fusion	0.45011	35.77013	0
PC1	Informativity	0.0778	7.02691	0
PC2	Word order	0.09187	-4.77918	0.00002
PC2	Flexivity	0.34598	9.42063	0
PC2	Noun class/gender	0.47968	8.06301	0
PC2	Locus of marking	0.08509	-1.78043	0
PC2	Fusion	0.35256	35.77013	0
PC2	Informativity	0.18245	7.02691	0
PC3	Word order	-0.03864	-4.77918	0.12904
PC3	Flexivity	-0.05418	9.42063	0.01215
PC3	Noun class/gender	0.16187	8.06301	0
PC3	Locus of marking	-0.15352	-1.78043	0
PC3	Fusion	-0.20571	35.77013	0
PC3	Informativity	-0.14179	7.02691	0

Table S4 Table of Grambank features.

ID	Name	Patrons
GB020	Are there definite or specific articles?	JLA JC
GB021	Do indefinite nominals commonly have indefinite articles?	JLA JC
GB022	Are there prenominal articles?	JLA JC
GB023	Are there postnominal articles?	JLA JC
GB024	What is the order of numeral and noun in the NP?	HJH
GB025	What is the order of adnominal demonstrative and noun?	JLA JC
GB026	Can adnominal property words occur discontinuously?	HJH
GB027	Are nominal conjunction and comitative expressed by different elements?	HS
GB028	Is there a distinction between inclusive and exclusive?	HJH
GB030	Is there a gender distinction in independent 3rd person pronouns?	HJH
GB031	Is there a dual or unit augmented form (in addition to plural or augmented) for all person categories in the pronoun system?	HJH
GB035	Are there three or more distance contrasts in demonstratives?	JLA JC
GB036	Do demonstratives show an elevation distinction?	JLA JC
GB037	Do demonstratives show a visible-nonvisible distinction?	JLA JC
GB038	Are there demonstrative classifiers?	JLA JC
GB039	Is there nonphonological allomorphy of noun number markers?	JLA JC
GB041	Are there several nouns (more than three) which are suppletive for number?	HS
GB042	Is there productive overt morphological singular marking on nouns?	HS
GB043	Is there productive morphological dual marking on nouns?	HS
GB044	Is there productive morphological plural marking on nouns?	HS
GB046	Is there an associative plural marker for nouns?	HS
GB047	Is there a productive morphological pattern for deriving an action/state noun from a verb?	HS
GB048	Is there a productive morphological pattern for deriving an agent noun from a verb?	HS
GB049	Is there a productive morphological pattern for deriving an object noun from a verb?	HS
GB051	Is there a gender/noun class system where sex is a factor in class assignment?	HJH
GB052	Is there a gender/noun class system where shape is a factor in class assignment?	HJH

GB053	Is there a gender/noun class system where animacy is a factor in class assignment?	HJH
GB054	Is there a gender/noun class system where plant status is a factor in class assignment?	HJH
GB057	Are there numeral classifiers?	JLA JC
GB058	Are there possessive classifiers?	JLA JC
GB059	Is the adnominal possessive construction different for alienable and inalienable nouns?	HJH
GB065	What is the pragmatically unmarked order of adnominal possessor noun and possessed noun?	HJH
GB068	Do core adjectives (defined semantically as property concepts such as value, shape, age, dimension) act like verbs in predicative position?	JLA JC
GB069	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) used attributively require the same morphological treatment as verbs?	JLA JC
GB070	Are there morphological cases for non-pronominal core arguments (i.e. S/A/P)?	JLE
GB071	Are there morphological cases for pronominal core arguments (i.e. S/A/P)?	JLE
GB072	Are there morphological cases for oblique non-pronominal NPs (i.e. not S/A/P)?	JLE
GB073	Are there morphological cases for independent oblique personal pronominal arguments (i.e. not S/A/P)?	JLE
GB074	Are there prepositions?	JLE
GB075	Are there postpositions?	JLE
GB079	Do verbs have prefixes/proclitics, other than those that only mark A, S or P (do include portmanteau: A & S + TAM)?	JLE
GB080	Do verbs have suffixes/enclitics, other than those that only mark A, S or P (do include portmanteau: A & S + TAM)?	JLE
GB081	Is there productive infixation in verbs?	HJH
GB082	Is there overt morphological marking of present tense on verbs?	HS
GB083	Is there overt morphological marking on the verb dedicated to past tense?	HS
GB084	Is there overt morphological marking on the verb dedicated to future tense?	HS
GB086	Is a morphological distinction between perfective and imperfective aspect available on verbs?	HS
GB089	Can the S argument be indexed by a suffix/enclitic on the verb in the simple main clause?	AWM
GB090	Can the S argument be indexed by a prefix/proclitic on the verb in the simple main clause?	AWM
GB091	Can the A argument be indexed by a suffix/enclitic on the verb in the simple main clause?	AWM
GB092	Can the A argument be indexed by a prefix/proclitic on the verb in the simple main clause?	AWM
GB093	Can the P argument be indexed by a suffix/enclitic on the verb in the simple main clause?	AWM

GB094	Can the P argument be indexed by a prefix/proclitic on the verb in the simple main clause?	AWM
GB095	Are variations in marking strategies of core participants based on TAM distinctions?	AWM
GB096	Are variations in marking strategies of core participants based on verb classes?	AWM
GB098	Are variations in marking strategies of core participants based on person distinctions?	AWM
GB099	Can verb stems alter according to the person of a core participant?	AWM
GB103	Is there a benefactive applicative marker on the verb (including indexing)?	JLE
GB104	Is there an instrumental applicative marker on the verb (including indexing)?	JLE
GB105	Can the recipient in a ditransitive construction be marked like the monotransitive patient?	AWM
GB107	Can standard negation be marked by an affix, clitic or modification of the verb?	HS
GB108	Is there directional or locative morphological marking on verbs?	JLE
GB109	Is there verb suppletion for participant number?	HS
GB110	Is there verb suppletion for tense or aspect?	HS
GB111	Are there conjugation classes?	JLA JC
GB113	Are there verbal affixes or clitics that turn intransitive verbs into transitive ones?	JLE
GB114	Is there a phonologically bound reflexive marker on the verb?	JLE
GB115	Is there a phonologically bound reciprocal marker on the verb?	JLE
GB116	Do verbs classify the shape, size or consistency of absolutive arguments by means of incorporated nouns, verbal affixes or suppletive verb stems?	JLA JC
GB117	Is there a copula for predicate nominals?	JLA JC
GB118	Are there serial verb constructions?	JLA JC
GB119	Can mood be marked by an inflecting word ('auxiliary verb')?	HS
GB120	Can aspect be marked by an inflecting word ('auxiliary verb')?	HS
GB121	Can tense be marked by an inflecting word ('auxiliary verb')?	HS
GB122	Is verb compounding a regular process?	JLA JC
GB123	Are there verb-adjunct (aka light-verb) constructions?	JLA JC
GB124	Is incorporation of nouns into verbs a productive intransitivizing process?	HJH
GB126	Is there an existential verb?	HS
GB127	Are different posture verbs used obligatorily depending on an inanimate locatum's shape or position (e.g. 'to lie' vs. 'to stand')?	JLE
GB129	Is there a notably small number, i.e. about 100 or less, of verb roots in the language?	HS

GB130	What is the pragmatically unmarked order of S and V in intransitive clauses?	HJH
GB131	Is a pragmatically unmarked constituent order verb-initial for transitive clauses?	HJH
GB132	Is a pragmatically unmarked constituent order verb-medial for transitive clauses?	HJH
GB133	Is a pragmatically unmarked constituent order verb-final for transitive clauses?	HJH
GB134	Is the order of constituents the same in main and subordinate clauses?	HJH
GB135	Do clausal objects usually occur in the same position as nominal objects?	HJH
GB136	Is the order of core argument (i.e. S/A/P) constituents fixed?	HJH
GB137	Can standard negation be marked clause-finally?	HJH
GB138	Can standard negation be marked clause-initially?	HJH
GB139	Is there a difference between imperative (prohibitive) and declarative negation constructions?	HS
GB140	Is verbal predication marked by the same negator as all of the following types of predication: locational, existential and nominal?	HS
GB146	Is there a morpho-syntactic distinction between predicates expressing controlled versus uncontrolled events or states?	JLE
GB147	Is there a morphological passive marked on the lexical verb?	JLE
GB148	Is there a morphological antipassive marked on the lexical verb?	JLE
GB149	Is there a morphologically marked inverse on verbs?	JLE
GB150	Is there clause chaining?	HJH
GB151	Is there an overt verb marker dedicated to signalling coreference or noncoreference between the subject of one clause and an argument of an adjacent clause ('switch reference')?	HJH
GB152	Is there a morphologically marked distinction between simultaneous and sequential clauses?	HJH
GB155	Are causatives formed by affixes or clitics on verbs?	JLE
GB156	Is there a causative construction involving an element that is unmistakably grammaticalized from a verb for 'to say'?	JLE
GB158	Are verbs reduplicated?	JLE
GB159	Are nouns reduplicated?	JLE
GB160	Are elements apart from verbs or nouns reduplicated?	JLE
GB165	Is there productive morphological trial marking on nouns?	HS
GB166	Is there productive morphological paucal marking on nouns?	HS
GB167	Is there a logophoric pronoun?	HJH
GB170	Can an adnominal property word agree with the noun in gender/noun class?	JLA JC

GB171	Can an adnominal demonstrative agree with the noun in gender/noun class?	JLA JC
GB172	Can an article agree with the noun in gender/noun class?	JLA JC
GB177	Can the verb carry a marker of animacy of argument, unrelated to any gender/noun class of the argument visible in the NP domain?	AWM
GB184	Can an adnominal property word agree with the noun in number?	JLA JC
GB185	Can an adnominal demonstrative agree with the noun in number?	JLA JC
GB186	Can an article agree with the noun in number?	JLA JC
GB187	Is there any productive diminutive marking on the noun (exclude marking by system of nominal classification only)?	JLA JC
GB188	Is there any productive augmentative marking on the noun (exclude marking by system of nominal classification only)?	JLA JC
GB192	Is there a gender system where a noun's phonological properties are a factor in class assignment?	HJH
GB193	What is the order of adnominal property word and noun?	JLA JC
GB196	Is there a male/female distinction in 2nd person independent pronouns?	HJH
GB197	Is there a male/female distinction in 1st person independent pronouns?	HJH
GB198	Can an adnominal numeral agree with the noun in gender/noun class?	JLA JC
GB203	What is the order of the adnominal collective universal quantifier ('all') and the noun?	HJH
GB204	Do collective ('all') and distributive ('every') universal quantifiers differ in their forms or their syntactic positions?	HJH
GB250	Can predicative possession be expressed with a transitive 'habeo' verb?	HS
GB252	Can predicative possession be expressed with an S-like possessum and a locative-coded possessor?	HS
GB253	Can predicative possession be expressed with an S-like possessum and a dative-coded possessor?	HS
GB254	Can predicative possession be expressed with an S-like possessum and a possessor that is coded like an adnominal possessor?	HS
GB256	Can predicative possession be expressed with an S-like possessor and a possessum that is coded like a comitative argument?	HS
GB257	Can polar interrogation be marked by intonation only?	JLA JC
GB260	Can polar interrogation be indicated by a special word order?	JLA JC
GB262	Is there a clause-initial polar interrogative particle?	JLA JC
GB263	Is there a clause-final polar interrogative particle?	JLA JC

GB264	Is there a polar interrogative particle that most commonly occurs neither clause-initially nor clause-finally?	JLA JC
GB265	Is there a comparative construction that includes a form that elsewhere means 'surpass, exceed'?	HJH
GB266	Is there a comparative construction that employs a marker of the standard which elsewhere has a locational meaning?	HJH
GB270	Can comparatives be expressed using two conjoined clauses?	HJH
GB273	Is there a comparative construction with a standard marker that elsewhere has neither a locational meaning nor a 'surpass/exceed' meaning?	HJH
GB275	Is there a bound comparative degree marker on the property word in a comparative construction?	HJH
GB276	Is there a non-bound comparative degree marker modifying the property word in a comparative construction?	HJH
GB285	Can polar interrogation be marked by a question particle and verbal morphology?	JLA JC
GB286	Can polar interrogation be indicated by overt verbal morphology only?	JLA JC
GB291	Can polar interrogation be marked by tone?	JLA JC
GB296	Is there a phonologically or morphosyntactically definable class of ideophones that includes ideophones depicting imagery beyond sound?	JLE
GB297	Can polar interrogation be indicated by a V-not-V construction?	JLA JC
GB298	Can standard negation be marked by an inflecting word ('auxiliary verb')?	HS
GB299	Can standard negation be marked by a non-inflecting word ('auxiliary particle')?	HS
GB300	Does the verb for 'give' have suppletive verb forms?	HS
GB301	Is there an inclusory construction?	JLA JC
GB302	Is there a phonologically free passive marker ('particle' or 'auxiliary')?	JLE
GB303	Is there a phonologically free antipassive marker ('particle' or 'auxiliary')?	JLE
GB304	Can the agent be expressed overtly in a passive clause?	JLE
GB305	Is there a phonologically independent reflexive pronoun?	JLE
GB306	Is there a phonologically independent non-bipartite reciprocal pronoun?	JLE
GB309	Are there multiple past or multiple future tenses, distinguishing distance from Time of Reference?	HS
GB312	Is there overt morphological marking on the verb dedicated to mood?	HS
GB313	Are there special adnominal possessive pronouns that are not formed by an otherwise regular process?	HJH

GB314	Can augmentative meaning be expressed productively by a shift of gender/noun class?	JLA JC
GB315	Can diminutive meaning be expressed productively by a shift of gender/noun class?	JLA JC
GB316	Is singular number regularly marked in the noun phrase by a dedicated phonologically free element?	HS
GB317	Is dual number regularly marked in the noun phrase by a dedicated phonologically free element?	HS
GB318	Is plural number regularly marked in the noun phrase by a dedicated phonologically free element?	HS
GB319	Is trial number regularly marked in the noun phrase by a dedicated phonologically free element?	HS
GB320	Is paucal number regularly marked in the noun phrase by a dedicated phonologically free element?	HS
GB321	Is there a large class of nouns whose gender/noun class is not phonologically or semantically predictable?	HJH
GB322	Is there grammatical marking of direct evidence (perceived with the senses)?	HJH
GB323	Is there grammatical marking of indirect evidence (hearsay, inference, etc.)?	HJH
GB324	Is there an interrogative verb for content interrogatives (who?, what?, etc.)?	HJH
GB325	Is there a count/mass distinction in interrogative quantifiers?	HJH
GB326	Do (nominal) content interrogatives normally or frequently occur in situ?	HJH
GB327	Can the relative clause follow the noun?	JLE
GB328	Can the relative clause precede the noun?	JLE
GB329	Are there internally-headed relative clauses?	JLE
GB330	Are there correlative relative clauses?	JLE
GB331	Are there non-adjacent relative clauses?	JLE
GB333	Is there a decimal numeral system?	JLE
GB334	Is there synchronic evidence for any element of a quinary numeral system?	JLE
GB335	Is there synchronic evidence for any element of a vigesimal numeral system?	JLE
GB336	Is there a body-part tallying system?	JLE
GB400	Are all person categories neutralized in some voice, tense, aspect, mood and/or negation?	AWM
GB401	Is there a class of patient-labile verbs?	AWM
GB402	Does the verb for 'see' have suppletive verb forms?	HS
GB403	Does the verb for 'come' have suppletive verb forms?	HS

GB408	Is there any accusative alignment of flagging?	AWM
GB409	Is there any ergative alignment of flagging?	AWM
GB410	Is there any neutral alignment of flagging?	AWM
GB415	Is there a politeness distinction in 2nd person forms?	HJH
GB421	Is there a preposed complementizer in complements of verbs of thinking and/or knowing?	HS
GB422	Is there a postposed complementizer in complements of verbs of thinking and/or knowing?	HS
GB430	Can adnominal possession be marked by a prefix on the possessor?	HJH
GB431	Can adnominal possession be marked by a prefix on the possessed noun?	HJH
GB432	Can adnominal possession be marked by a suffix on the possessor?	HJH
GB433	Can adnominal possession be marked by a suffix on the possessed noun?	HJH
GB519	Can mood be marked by a non-inflecting word ('auxiliary particle')?	HS
GB520	Can aspect be marked by a non-inflecting word ('auxiliary particle')?	HS
GB521	Can tense be marked by a non-inflecting word ('auxiliary particle')?	HS
GB522	Can the S or A argument be omitted from a pragmatically unmarked clause when the referent is inferrable from context ('pro-drop' or 'null anaphora')?	HJH

Table S5. Table of binarised Grambank features

ID	Abbreviation
GB024a	GB024a NUMOrder_Num-N
GB024b	GB024b NUMOrder_N-Num
GB025a	GB025a DEMOrder_Dem-N
GB025b	GB025b DEMOrder_N-Dem
GB065a	GB065a POSSOrder_PSR-PSD
GB065b	GB065b POSSOrder_PSD-PSR
GB130a	GB130a IntransOrder_SV
GB130b	GB130b IntransOrder_VS
GB193a	GB193a ANMOrder_ANM-N
GB193b	GB193b ANMOrder_N-ANM
GB203a	GB203a UQOrder_UQ-N
GB203b	GB203b UQOrder_N-UQ

Table S6. Grambank features with information on theoretical scores and predictions from Nichols (1995).

Feature_ID	Fusion	Flexivity	Gender/noun class	locus of marking	word order	informativity	Main_domain	Nichols_1995_label	Nichols_1995_prediction
GB303						antipassive	clause		
GB149	1					inverse	verbal domain		
GB070	1			0			nominal domain		
GB071	0.5			0			pronoun		
GB408				0			nominal domain	Dom alignment	G
GB409				0			nominal domain	Dom alignment	G
GB410				0			nominal domain	Dom alignment	G
GB074					1		nominal domain	Adposition place	G
GB075					0		nominal domain	Adposition place	G
GB080	1						verbal domain		
GB081	1						verbal domain		
GB079	1						verbal domain		
GB092	1			1			verbal domain	1 agreement	G
GB093	1			1			verbal domain	2 agreement	
GB089	1			1			verbal domain	1 agreement	G
GB090	1			1			verbal	1 agreement	G

							domain		
GB091	1			1			verbal domain	1 agreement	G
GB094	1			1			verbal domain	2 agreement	
GB098		1					verbal domain		
GB095		1					verbal domain		
GB096		1					verbal domain		
GB105							clause		
GB072	1			0			nominal domain		
GB073	0.5			0			pronoun		
GB108	1					directional	verbal domain		
GB027						comitative	clause		
GB103	1					benefactive	verbal domain		
GB104	1					instrumental	verbal domain		
GB026							nominal domain		
GB193							nominal domain		
GB069							nominal domain		
GB065							nominal domain		A
GB059						alienability	nominal domain		

GB430	1			0			nominal domain		
GB431	1			1			nominal domain	Noun Poss. Place	A
GB432	1			0			nominal domain		
GB433	1			1			nominal domain	Noun Poss. Place	A
GB313							pronoun		
GB058		1	0			possessive classifiers	nominal domain		
GB155	1						verbal domain	+A	G
GB156							clause		
GB028						clusivity	pronoun	Incl/excl	G
GB301							clause		
GB265							clause		
GB270							clause		
GB273							clause		
GB275	1						clause		
GB276							clause		
GB266							clause		
GB146	0.5					control	nominal domain		
GB020						definiteart icles	nominal domain		
GB022					0		nominal domain		
GB021						indef	nominal domain		

GB023					1		nominal domain		
GB035						demonstrativedistance	nominal domain		
GB037						demonstrativevisibility	nominal domain		
GB036						demonstrativeelevation	nominal domain		
GB151	1					switch reference	verbal domain		
GB025							nominal domain		
GB038		1	0			demonstrative classifiers	nominal domain		
GB159							nominal domain		
GB160							nominal domain		
GB158							verbal domain		
GB048	0.5						nominal domain		
GB049	0.5						nominal domain		
GB047	0.5						nominal domain		
GB321		1	1				nominal domain		
GB051		1	1			gendersex	nominal domain	Genders	G
GB052		1	1			gendershape	nominal domain	Genders	G

GB054		1	1			genderplant	nominal domain	Genders	G
GB192		1	1				nominal domain	Genders	G
GB196			1			pronoungender2	pronoun		
GB197			1			pronoungender1	pronoun		
GB053		1	1			genderanimacy	nominal domain	Genders	G
GB170	1	1	1	0			nominal domain	Genders	G
GB171	1	1	1	0			nominal domain	Genders	G
GB172	1	1	1	0			nominal domain		
GB314						augmentative	nominal domain		
GB315						diminutive	nominal domain		
GB296							nominal domain		
GB167						pronounlog	pronoun		
GB257							clause		
GB260							clause		
GB262					1		clause		
GB263							clause		
GB264							clause		
GB285	1						clause		
GB286	1						clause		

GB291							clause		
GB324							clause		
GB326							clause		
GB325						count_mass	nominal domain		
GB116		1				verbclassify	verbal domain		
GB177	1	1	1	1			verbal domain		
GB057		1	0			numera classifers	nominal domain	Numeral Classifier	G, A
GB188	1	1				augmentative	nominal domain		
GB187	1	1				diminutive	nominal domain		
GB046						assocplural	nominal domain		
GB316						singular	nominal domain		
GB317						dual	nominal domain		
GB318						plural	nominal domain		
GB319						trial	nominal domain		
GB320						paucal	nominal domain		
GB039		1					nominal domain		
GB165	1			1		trial	nominal domain		
GB166	1			1		paucal	nominal domain		

GB041		1					nominal domain		
GB043	1			1		dual	nominal domain		
GB109		1		1			verbal domain		
GB184	1			0			nominal domain		
GB185	1			0			nominal domain		
GB186	1			0			nominal domain		
GB044	1			1		plural	nominal domain	Noun Sg/Pl	G
GB042	1			1		singular	nominal domain		
GB302						passive	clause	-A	G
GB304							clause		
GB099		1		1			verbal domain		
GB031						pronoundualaug	pronoun		
GB030		1	1			pronoungender3	pronoun		
GB400							verbal domain		
GB415						politeness	pronoun		
GB132							clause	Word order	A
GB118							verbal domain		
GB131					1		clause	Word order	A
GB136							clause		

GB130							clause	Word order	A
GB522							clause		
GB133					0		clause	Word order	A
GB150							clause		
GB122							verbal domain		
GB123							verbal domain		
GB140						differentneg	clause		
GB256							clause		
GB253							clause		
GB254							clause		
GB252							clause		
GB135							clause		
GB134							clause		
GB068							nominal domain		
GB117						copulaprednom	verbal domain		
GB333							numeral		
GB334							numeral		
GB335							numeral		
GB336							numeral		
GB024							nominal domain		
GB203							nominal domain		

GB204							nominal domain		
GB198	1	1	1	0			nominal domain	Genders	G
GB115	1			1		reciprocity	verbal domain		
GB114	1			1		reflexivity	verbal domain		
GB327					1		nominal domain		
GB328					0		clause		
GB329							clause		
GB330							clause		
GB331							clause		
GB421					1		clause		
GB422					0		clause		
GB086	1					aspect	verbal domain		
GB120	1					aspect	verbal domain		
GB520						aspect	verbal domain		
GB322						evidentiality_direct	verbal domain		
GB323						evidentiality_indirect	verbal domain		
GB139						prohibitive	clause		
GB297							clause		
GB119	1					mood	verbal domain		

GB312	1					mood	verbal domain		
GB519						mood	verbal domain		
GB138							clause		
GB107	1						verbal domain		
GB137							clause		
GB298	1						clause		
GB299							clause		
GB152	1					simultanseq	clause		
GB084	1					tense	verbal domain		
GB309						multiplete nse	verbal domain		
GB521						tense	verbal domain		
GB082	1					tense	verbal domain		
GB083	1					tense	verbal domain		
GB121	1					tense	verbal domain		
GB110		1					verbal domain		
GB111		1					verbal domain		
GB148	1					antipassive	verbal domain		
GB113	1						verbal domain	+A	A

GB147	1					passive	verbal domain	-A	G
GB305						reflexivity	pronoun		
GB306						reciprocity	pronoun		
GB124							verbal domain	-A	G
GB401							verbal domain		
GB129							verbal domain		
GB127						posturever bs	verbal domain		
GB126						existential verb	verbal domain		
GB250							nominal domain		
GB402		1					verbal domain		
GB403		1					verbal domain		
GB300		1					verbal domain		
GB024a							nominal domain		
GB024b							nominal domain		
GB025a							nominal domain		
GB025b							nominal domain		
GB065a							nominal domain		
GB065b							nominal domain		

GB130a							clause		
GB130b							clause		
GB193a							nominal domain		
GB193b							nominal domain		
GB203a							nominal domain		
GB203b							nominal domain		

Table S7: Cultural Fixation Scores between AUTOTYP-areas

Group_Var1	Group_Var2	Cultural Fixation Score	Americas_Var1	Americas_Var2
Basin and Plains	E North America	0.058	americas	americas
S New Guinea	N Coast New Guinea	0.0746	not americas	not americas
S New Guinea	NE South America	0.0851	not americas	americas
Oceania	N Coast New Guinea	0.0863	not americas	not americas
Basin and Plains	NE South America	0.0876	americas	americas
California	NE South America	0.0897	americas	americas
Indic	NE South America	0.0898	not americas	americas
Basin and Plains	Alaska-Oregon	0.0903	americas	americas
Interior New Guinea	S New Guinea	0.092	not americas	not americas
N Coast New Guinea	African Savannah	0.0923	not americas	not americas
Oceania	Southeast Asia	0.0961	not americas	not americas
Greater Abyssinia	Greater Mesopotamia	0.0979	not americas	not americas

California	Andean	0.1044	americas	americas
California	Basin and Plains	0.1068	americas	americas
E North America	NE South America	0.1082	americas	americas
Inner Asia	Indic	0.1086	not americas	not americas
Andean	NE South America	0.109	americas	americas
SE South America	E North America	0.1092	americas	americas
NE South America	Mesoamerica	0.111	americas	americas
Southeast Asia	N Coast New Guinea	0.1148	not americas	not americas
Greater Abyssinia	NE South America	0.1177	not americas	americas
NE South America	N Coast New Guinea	0.1221	americas	not americas
S Africa	African Savannah	0.1241	not americas	not americas
NE South America	Greater Mesopotamia	0.1246	americas	not americas
S Australia	California	0.1284	not americas	americas
Southeast Asia	African Savannah	0.1306	not americas	not americas
Indic	S New Guinea	0.1307	not americas	not americas
Oceania	African Savannah	0.1337	not americas	not americas
California	E North America	0.1343	americas	americas
N Coast Asia	Indic	0.1346	not americas	not americas
Andean	N Coast Asia	0.135	americas	not americas
N Australia	S New Guinea	0.1367	not americas	not americas
SE South America	NE South America	0.1373	americas	americas
SE South America	Basin and Plains	0.1376	americas	americas
Inner Asia	N Coast Asia	0.1378	not americas	not americas

Indic	Greater Abyssinia	0.1381	not americas	not americas
California	S New Guinea	0.1445	americas	not americas
Andean	Greater Abyssinia	0.1463	americas	not americas
Mesoamerica	N Coast New Guinea	0.1489	americas	not americas
Andean	S New Guinea	0.1494	americas	not americas
California	Alaska-Oregon	0.15	americas	americas
Basin and Plains	Mesoamerica	0.1539	americas	americas
S New Guinea	Greater Mesopotamia	0.1542	not americas	not americas
Alaska-Oregon	E North America	0.1547	americas	americas
Andean	Indic	0.1553	americas	not americas
Inner Asia	Greater Abyssinia	0.1562	not americas	not americas
Andean	Greater Mesopotamia	0.157	americas	not americas
California	Greater Abyssinia	0.157	americas	not americas
N Australia	NE South America	0.1572	not americas	americas
Inner Asia	NE South America	0.1573	not americas	americas
N Coast Asia	Greater Abyssinia	0.1573	not americas	not americas
Inner Asia	Greater Mesopotamia	0.1578	not americas	not americas
Europe	Greater Mesopotamia	0.1585	not americas	not americas
California	N Australia	0.1611	americas	not americas
Basin and Plains	S New Guinea	0.1625	americas	not americas
S New Guinea	Greater Abyssinia	0.1631	not americas	not americas
Andean	Basin and Plains	0.1636	americas	americas
S Australia	N Australia	0.1648	not americas	not americas

Alaska-Oregon	Mesoamerica	0.1657	americas	americas
California	N Coast Asia	0.1674	americas	not americas
S New Guinea	Mesoamerica	0.1693	not americas	americas
Basin and Plains	N Australia	0.1696	americas	not americas
Indic	Greater Mesopotamia	0.1704	not americas	not americas
California	Greater Mesopotamia	0.1711	americas	not americas
Interior New Guinea	Andean	0.1729	not americas	americas
N Australia	N Coast New Guinea	0.176	not americas	not americas
California	Interior New Guinea	0.1762	americas	not americas
Oceania	Mesoamerica	0.1779	not americas	americas
NE South America	African Savannah	0.1793	americas	not americas
Mesoamerica	Greater Mesopotamia	0.1803	americas	not americas
N Australia	Greater Mesopotamia	0.1803	not americas	not americas
Basin and Plains	Greater Mesopotamia	0.181	americas	not americas
S New Guinea	African Savannah	0.1819	not americas	not americas
Andean	Inner Asia	0.1821	americas	not americas
Basin and Plains	Greater Abyssinia	0.1824	americas	not americas
SE South America	Alaska-Oregon	0.1825	americas	americas
SE South America	California	0.1837	americas	americas
N Coast Asia	NE South America	0.1856	not americas	americas
Indic	N Coast New Guinea	0.1873	not americas	not americas
Indic	Southeast Asia	0.1892	not americas	not americas
Mesoamerica	African Savannah	0.1898	americas	not americas

California	Indic	0.1919	americas	not americas
E North America	S New Guinea	0.1925	americas	not americas
Alaska-Oregon	NE South America	0.1971	americas	americas
Andean	E North America	0.1999	americas	americas
N Australia	E North America	0.2005	not americas	americas
S Australia	Andean	0.2029	not americas	americas
Andean	Alaska-Oregon	0.2036	americas	americas
N Australia	Greater Abyssinia	0.2098	not americas	not americas
Interior New Guinea	NE South America	0.21	not americas	americas
S Africa	NE South America	0.2109	not americas	americas
NE South America	Southeast Asia	0.2112	americas	not americas
Interior New Guinea	N Australia	0.2123	not americas	not americas
Inner Asia	S New Guinea	0.215	not americas	not americas
S Africa	N Coast New Guinea	0.217	not americas	not americas
Indic	African Savannah	0.2171	not americas	not americas
Interior New Guinea	N Coast New Guinea	0.2181	not americas	not americas
E North America	Mesoamerica	0.2181	americas	americas
Interior New Guinea	Greater Abyssinia	0.2189	not americas	not americas
Interior New Guinea	Indic	0.2192	not americas	not americas
California	Mesoamerica	0.2196	americas	americas
S Australia	Interior New Guinea	0.2207	not americas	not americas
E North America	Greater Abyssinia	0.2238	americas	not americas
S Australia	N Coast Asia	0.2283	not americas	not americas

Alaska-Oregon	Greater Abyssinia	0.2289	americas	not americas
S Australia	S New Guinea	0.2292	not americas	not americas
Indic	Mesoamerica	0.2296	not americas	americas
Europe	Greater Abyssinia	0.2296	not americas	not americas
S Australia	Basin and Plains	0.2301	not americas	americas
N Coast New Guinea	Greater Mesopotamia	0.2311	not americas	not americas
S Australia	Greater Abyssinia	0.2311	not americas	not americas
S New Guinea	Southeast Asia	0.2313	not americas	not americas
SE South America	N Australia	0.2313	americas	not americas
Alaska-Oregon	Greater Mesopotamia	0.232	americas	not americas
Interior New Guinea	N Coast Asia	0.2328	not americas	not americas
Basin and Plains	N Coast Asia	0.233	americas	not americas
Basin and Plains	N Coast New Guinea	0.2336	americas	not americas
S Australia	NE South America	0.2346	not americas	americas
Andean	N Australia	0.2357	americas	not americas
E North America	Greater Mesopotamia	0.236	americas	not americas
Oceania	NE South America	0.2361	not americas	americas
N Australia	Mesoamerica	0.238	not americas	americas
N Coast Asia	Greater Mesopotamia	0.2423	not americas	not americas
N Australia	Alaska-Oregon	0.2427	not americas	americas
California	Inner Asia	0.2443	americas	not americas
S Africa	Mesoamerica	0.2446	not americas	americas
Andean	Mesoamerica	0.2461	americas	americas

N Africa	Alaska-Oregon	0.2479	not americas	americas
S Australia	Greater Mesopotamia	0.2517	not americas	not americas
California	N Coast New Guinea	0.2544	americas	not americas
Greater Abyssinia	N Coast New Guinea	0.2546	not americas	not americas
N Coast Asia	S New Guinea	0.255	not americas	not americas
Inner Asia	Mesoamerica	0.2561	not americas	americas
S Africa	Greater Mesopotamia	0.2571	not americas	not americas
S Australia	Alaska-Oregon	0.2587	not americas	americas
Greater Abyssinia	Mesoamerica	0.2589	not americas	americas
Interior New Guinea	Basin and Plains	0.2607	not americas	americas
E North America	N Coast Asia	0.2612	americas	not americas
S New Guinea	S Africa	0.2617	not americas	not americas
SE South America	Mesoamerica	0.2625	americas	americas
S New Guinea	Oceania	0.2627	not americas	not americas
Europe	Inner Asia	0.2628	not americas	not americas
N Africa	Greater Mesopotamia	0.2648	not americas	not americas
Greater Mesopotamia	African Savannah	0.2665	not americas	not americas
Interior New Guinea	Greater Mesopotamia	0.2688	not americas	not americas
SE South America	N Africa	0.271	americas	not americas
Indic	Oceania	0.2726	not americas	not americas
Alaska-Oregon	N Coast Asia	0.2745	americas	not americas
Interior New Guinea	E North America	0.2756	not americas	americas
Inner Asia	N Coast New Guinea	0.2773	not americas	not americas

S Australia	Inner Asia	0.2775	not americas	not americas
S Australia	Indic	0.2797	not americas	not americas
Greater Abyssinia	African Savannah	0.2803	not americas	not americas
Mesoamerica	Southeast Asia	0.2804	americas	not americas
S Australia	E North America	0.2804	not americas	americas
Alaska-Oregon	S New Guinea	0.2843	americas	not americas
SE South America	Andean	0.2847	americas	americas
SE South America	Greater Mesopotamia	0.2882	americas	not americas
Europe	N Australia	0.2944	not americas	not americas
Basin and Plains	Indic	0.2962	americas	not americas
N Australia	N Coast Asia	0.2968	not americas	not americas
Europe	Mesoamerica	0.2974	not americas	americas
SE South America	S New Guinea	0.2987	americas	not americas
Basin and Plains	Inner Asia	0.2988	americas	not americas
E North America	N Coast New Guinea	0.3002	americas	not americas
N Australia	Indic	0.3039	not americas	not americas
Andean	N Coast New Guinea	0.304	americas	not americas
Interior New Guinea	Alaska-Oregon	0.3062	not americas	americas
SE South America	Greater Abyssinia	0.3069	americas	not americas
N Australia	Inner Asia	0.3094	not americas	not americas
Europe	NE South America	0.312	not americas	americas
Oceania	S Africa	0.3155	not americas	not americas
Europe	S New Guinea	0.317	not americas	not americas

Greater Abyssinia	S Africa	0.3204	not americas	not americas
N Australia	S Africa	0.3214	not americas	not americas
N Australia	African Savannah	0.3227	not americas	not americas
Indic	S Africa	0.328	not americas	not americas
S Australia	N Coast New Guinea	0.3323	not americas	not americas
Interior New Guinea	Inner Asia	0.3356	not americas	not americas
Basin and Plains	S Africa	0.3394	americas	not americas
S Africa	Southeast Asia	0.3398	not americas	not americas
SE South America	Interior New Guinea	0.3428	americas	not americas
SE South America	N Coast Asia	0.3439	americas	not americas
Alaska-Oregon	N Coast New Guinea	0.3445	americas	not americas
Basin and Plains	Europe	0.3457	americas	not americas
Inner Asia	African Savannah	0.3464	not americas	not americas
Andean	Europe	0.3507	americas	not americas
S Australia	Mesoamerica	0.3566	not americas	americas
SE South America	S Australia	0.3583	americas	not americas
N Africa	N Australia	0.3605	not americas	not americas
Interior New Guinea	Mesoamerica	0.3616	not americas	americas
N Africa	Greater Abyssinia	0.3633	not americas	not americas
Europe	African Savannah	0.3634	not americas	not americas
Europe	N Coast New Guinea	0.3689	not americas	not americas
N Africa	California	0.3695	not americas	americas
Europe	Indic	0.3739	not americas	not americas

N Coast Asia	Mesoamerica	0.3742	not americas	americas
SE South America	N Coast New Guinea	0.3778	americas	not americas
Europe	Alaska-Oregon	0.3779	not americas	americas
N Africa	Basin and Plains	0.3814	not americas	americas
California	Europe	0.3835	americas	not americas
E North America	Indic	0.3872	americas	not americas
Inner Asia	S Africa	0.3907	not americas	not americas
S Australia	Europe	0.3946	not americas	not americas
N Africa	E North America	0.3956	not americas	americas
Europe	S Africa	0.3965	not americas	not americas
Basin and Plains	African Savannah	0.3972	americas	not americas
Inner Asia	Southeast Asia	0.3975	not americas	not americas
E North America	S Africa	0.3998	americas	not americas
Interior New Guinea	African Savannah	0.4038	not americas	not americas
Interior New Guinea	Europe	0.4124	not americas	not americas
Inner Asia	E North America	0.4128	not americas	americas
N Coast Asia	N Coast New Guinea	0.4161	not americas	not americas
N Australia	Oceania	0.4183	not americas	not americas
Inner Asia	Alaska-Oregon	0.4324	not americas	americas
California	African Savannah	0.4363	americas	not americas
California	Southeast Asia	0.4401	americas	not americas
Europe	N Coast Asia	0.4411	not americas	not americas
California	Oceania	0.4449	americas	not americas

Basin and Plains	Oceania	0.4477	americas	not americas
N Australia	Southeast Asia	0.4478	not americas	not americas
S Australia	N Africa	0.4504	not americas	not americas
Europe	E North America	0.4604	not americas	americas
California	S Africa	0.4635	americas	not americas
Andean	African Savannah	0.4684	americas	not americas
N Africa	Europe	0.4703	not americas	not americas
Interior New Guinea	Southeast Asia	0.4707	not americas	not americas
Alaska-Oregon	Oceania	0.4728	americas	not americas
N Africa	Andean	0.4772	not americas	americas
Oceania	Greater Mesopotamia	0.4873	not americas	not americas
N Africa	Interior New Guinea	0.4901	not americas	not americas
Southeast Asia	Greater Mesopotamia	0.4909	not americas	not americas
E North America	African Savannah	0.4926	americas	not americas
Alaska-Oregon	Indic	0.4938	americas	not americas
Greater Abyssinia	Southeast Asia	0.4941	not americas	not americas
N Africa	Mesoamerica	0.4988	not americas	americas
Inner Asia	Oceania	0.4991	not americas	not americas
N Africa	African Savannah	0.5031	not americas	not americas
Basin and Plains	Southeast Asia	0.5034	americas	not americas
Alaska-Oregon	S Africa	0.5158	americas	not americas
N Africa	NE South America	0.516	not americas	americas
Alaska-Oregon	African Savannah	0.5221	americas	not americas

SE South America	S Africa	0.5229	americas	not americas
N Africa	N Coast Asia	0.5243	not americas	not americas
Andean	S Africa	0.5312	americas	not americas
Andean	Southeast Asia	0.5445	americas	not americas
S Australia	African Savannah	0.5522	not americas	not americas
Greater Abyssinia	Oceania	0.5553	not americas	not americas
S Australia	Southeast Asia	0.5574	not americas	not americas
S Australia	Oceania	0.5617	not americas	not americas
N Africa	N Coast New Guinea	0.5711	not americas	not americas
N Africa	S New Guinea	0.5728	not americas	not americas
SE South America	African Savannah	0.5771	americas	not americas
Interior New Guinea	S Africa	0.5828	not americas	not americas
S Australia	S Africa	0.585	not americas	not americas
SE South America	Europe	0.585	americas	not americas
N Coast Asia	Southeast Asia	0.5952	not americas	not americas
Andean	Oceania	0.6095	americas	not americas
Interior New Guinea	Oceania	0.6145	not americas	not americas
E North America	Oceania	0.6242	americas	not americas
SE South America	Indic	0.6359	americas	not americas
N Africa	S Africa	0.6406	not americas	not americas
N Coast Asia	African Savannah	0.6509	not americas	not americas
E North America	Southeast Asia	0.6699	americas	not americas
N Coast Asia	S Africa	0.7047	not americas	not americas

SE South America	Inner Asia	0.706	americas	not americas
Europe	Oceania	0.7061	not americas	not americas
SE South America	Oceania	0.7213	americas	not americas
Alaska-Oregon	Southeast Asia	0.7451	americas	not americas
Europe	Southeast Asia	0.747	not americas	not americas
N Coast Asia	Oceania	0.7992	not americas	not americas
N Africa	Inner Asia	0.9289	not americas	not americas
SE South America	Southeast Asia	0.9627	americas	not americas
N Africa	Indic	1.0179	not americas	not americas
N Africa	Oceania	1.0647	not americas	not americas
N Africa	Southeast Asia	1.5549	not americas	not americas

Table S8. Coefficients and associated error estimates for the spatiophylogenetic Bayesian regression model predicting Unsualness scores.

Coefficient	**Estimate**	**Estimated error**
Intercept	4.73	0.21
SD	0.24	0.01
SD (phylogeny)	0.08	0.01
SD (spatial)	0.15	0.02

Table S9: Language pairs with a Manhattan distance of 0

Glottocodes	**Names**	**Family name**
pahn1237-biao1256	Pa-Hng-Biao Mon	Hmong-Mien
xish1235-cosa1234	Xishanba Lalo-Cosao	Sino-Tibetan
kusa1251-hoav1238	Kusaghe-Njela-Hoava	Austronesian

kare1335-ingr1248	Karelian-Ingrian	Uralic
sout2959-nort2942	South Slavey-North Slavey	Athabaskan-Eyak-Tlingit
puni1241-phoe1239	Punic-Phoenician	Afro-Asiatic