Evolutionary dynamics of language systems

Simon J. Greenhill[a,b,1], Chieh-Hsi Wu[c], Xia Hua[d], Michael Dunn[e], Stephen C. Levinson[f,g], and Russell D. Gray[b,h]

[a]ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, ACT 0200, Australia; [b]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, 07745 Jena, Germany; [c]Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom; [d]Macroevolution and Macroecology, Division of Ecology, Evolution, and Genetics, Research School of Biology, Australian National University, Canberra, ACT 0200 Australia; [e]Department of Linguistics and Philology, Uppsala University, 75238 Uppsala, Sweden; [f]Max Planck Institute for Psycholinguistics, 6525 XD, Nijmegen, The Netherlands; [g]Comparative Linguistics, Radboud University Nijmegen, 6525 HP, Nijmegen, The Netherlands; and [h]School of Psychology, University of Auckland, Auckland, New Zealand

Edited by Tomoko Ohta, National Institute of Genetics, Mishima, Japan, and approved July 11, 2017 (received for review January 8, 2017)

Understanding how and why language subsystems differ in their evolutionary dynamics is a fundamental question for historical and comparative linguistics. One key dynamic is the rate of language change. While it is commonly thought that the rapid rate of change hampers the reconstruction of deep language relationships beyond 6,000–10,000 y, there are suggestions that grammatical structures might retain more signal over time than other subsystems, such as basic vocabulary. In this study, we use a Dirichlet process mixture model to infer the rates of change in lexical and grammatical data from 81 Austronesian languages. We show that, on average, most grammatical features actually change faster than items of basic vocabulary. The grammatical data show less schismogenesis, higher rates of homoplasy, and more bursts of contact-induced change than the basic vocabulary data. However, there is a core of grammatical and lexical features that are highly stable. These findings suggest that different subsystems of language have differing dynamics and that careful, nuanced models of language change will be needed to extract deeper signal from the noise of parallel evolution, areal readaptation, and contact.

language evolution | language dynamics | language phylogenies | typology | linguistics

Understanding how and why language systems differ in their evolutionary dynamics is a fundamental question for historical-comparative linguistics. One key dynamic is the rate of change: Are some subsystems of language more stable over time or less prone to borrowing than others? Attempts to trace the deep history of languages, and the peoples who spoke them, are hampered by the rate at which languages change. The orthodox view in historical linguistics, based on reconciling linguistic reconstruction with archaeological inferences, is that after 6,000–10,000 y, the genealogical signal becomes so weak, and so difficult to separate from chance similarities and borrowings, that attempts to infer deeper linguistic history will inevitably fail (1, 2). This limitation is unfortunate as it hampers our ability to make inferences about language relationships, and human prehistory, beyond this "time barrier."

Grammatical structures are sometimes claimed to be a solution to this time barrier problem. First, the abstract nature of the grammatical features of language means they are comparable between languages not known to be related, while comparison on the basis of the lexicon relies upon substantial linguistic work to identify sound correspondences and cognate items (3). Second, grammatical structures are more tightly integrated than lexical or phonological features (4–6). Tight systemic integration should make these structures much more resistant to change than the lexicon (5, 6). Third, while grammatical borrowing occurs, it is thought to be harder to borrow a grammatical pattern than a word, and grammatical borrowing should only happen when there is sustained and intimate contact between languages (6, 7).

Following these arguments, some scholars have used grammatical structures to trace deeper history. Nichols (8) describes some striking structural similarities shared between languages around the Pacific Rim from Australia and New Guinea to mainland Southeast Asia and into the Americas. If these similarities are due to deep connections, then this signal must date back at least 15,000 if not 50,000 y. A more recent set of studies analyzing the structure of 31 languages in Island Melanesia found results linking the non-Austronesian languages in the region. If true, then this signal must be a residue of language relationships dating back before the Austronesian expansion into the region ~3,500 y ago (3, 4, 9). Given the great disparity between the non-Austronesian languages, this signal could date to more than 10,000 y (2). What might this signal be? As Nichols (ref. 8, p. 208) says, "we can be quite confident that a group of stocks systematically sharing a number of such features has some historical identity as a group, although we cannot assume that the historical connection is specifically genealogical." Thus, for these grammatical structures to be highly stable over time, they must combine the effects of phylogenetic and areal inheritance. Genealogically stable features must spread into incoming languages, either directly or indirectly, through processes like reanalysis, reinterpretation, or grammaticalization (10), and subsequently remain relatively genealogically stable over time, leading to repeated readaptation of areal norms (11).

On the other hand, however, grammatical structures have a number of drawbacks that could limit their ability to trace language relationships. First, despite arguments that borrowing of grammatical structures requires sustained intimate contact (6), there are indications that at least some features readily diffuse between languages indirectly (7). Second, the abstract coding and limited design space of these structures means that the risk of chance similarity is much higher due to increased rates of convergence and parallel evolution (12). For example, an important grammatical structure is sentence word order, but there are only six possible ways of ordering the subject, object, and verb. Third, unlike the lexicon, many structural features are functionally linked such that a change in one causes a change in another (13). All of these factors could overwrite the historical information inherent in the evolved histories of these data, and cause problems for deep reconstruction.

Significance

Do different aspects of language evolve in different ways? Here, we infer the rates of change in lexical and grammatical data from 81 languages of the Pacific. We show that, in general, grammatical features tend to change faster and have higher amounts of conflicting signal than basic vocabulary. We suggest that subsystems of language show differing patterns of dynamics and propose that modeling this rate variation may allow us to extract more signal, and thus trace language history deeper than has been previously possible.

Author contributions: S.J.G., C.-H.W., M.D., S.C.L., and R.D.G. designed research; S.J.G., C.-H.W., X.H., M.D., S.C.L., and R.D.G. performed research; M.D. and S.C.L. contributed new reagents/analytic tools; S.J.G., C.-H.W., and X.H. analyzed data; and S.J.G., C.-H.W., M.D., S.C.L., and R.D.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: simon.greenhill@anu.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1700388114/-/DCSupplemental.

Because of the importance of understanding the rates at which different aspects of language change, and the potential payoffs of pushing the time barrier, there have been a number of efforts to investigate the relative stability of linguistic traits (14). For example, in a survey of the Polynesian languages, Pawley (15) found that tense-aspect, direction, and position markers are much more persistent than conjunctions or manner particles. Most prominent, however, are suggestions by Nichols (5) that certain features are highly stable and reflect genealogical relatedness. These features include head vs. dependent marking, alignment, word order, voice, inclusive vs. exclusive distinction in pronominals, plurality neutralization, inalienable possession, and noun classes. However, these claims were not based on a quantitative assessment of the rates of change (16), and conflate the role of stability and diffusion such that "stability" is either an outcome of inheritance of features tracing language relationships (genealogical stability) or occurs as a result of repeated diffusion within a language area (repeated readaptation to a regional norm).

To remedy the shortcomings of the previous work, we apply a Bayesian nonparametric approach to datasets of basic vocabulary and structural features from 81 Austronesian languages (map in Fig. 1). We collated a large database of structural information for these languages (3). This database was carefully constructed to encode the presence or absence of structural traits (e.g., whether a language shows a phonemic distinction between [l] and [r], whether a language has two or more contrastive tones, whether there are prenominal articles, or whether the language distinguishes gender in the third-person pronouns). These variables and their coding were selected to better recover deep signal (3). We compared these features with 210 items of cognate-coded basic vocabulary for the same languages extracted from the Austronesian Basic Vocabulary Database (ABVD) (17).

To investigate the evolutionary dynamics of these data, we first estimate rates of change using a Dirichlet process mixture model. This model is designed to counter the problems of estimating rates by assigning features to rate categories using a Chinese restaurant process (18). The number of rate categories and the assignment of each feature to a category are simultaneously estimated along with the tree topology in a Bayesian model-averaging framework. The Bayesian phylogenetic modeling enables us to investigate the genealogical stability and explore the dynamics of both lexicon and grammatical structures by directly quantifying the rates of change with an elegant model that coestimates phylogeny, rate categories, and the assignment of characters to different categories along with the uncertainty around all these parameters. Second, we assess how much regional influences shaped these languages and investigate the effect of the different design space sizes by quantifying the amount of conflicting signal (homoplasy) in these data. Finally, we investigate if the rates of change in these systems are equally affected by punctuated bursts of more rapid change when speakers act to differentiate themselves from their sister language when linguistic lineages separate, a process dubbed "schismogenesis" by Bateson (19).

Results

Rates Comparison. To identify the best-fitting model, we used the Akaike information criterion through Markov chain Monte Carlo (MCMC) [AICM (20)], where AICM scores with a difference greater than 7 are significant (21). The AICM indicated that the best-fitting model for the rates comparison was a relaxed clock along with a log-normally distributed base measure of the Dirichlet process prior (Table 1). Unfortunately, the current implementation of the Dirichlet process mixture model does not accommodate path-sampling calculations, which provide better estimates of the marginal likelihood and have higher power for model comparison (20). However, the rates results are highly consistent across models, suggesting that rate identification is robust to model choice (pairwise Spearman rank correlations range from 0.89 to 0.90, and all are significant at $P < 0.00001$).

Overall the combined lexical and structural data changed at a median rate of 1.48e-05 changes per feature per year (SD of median item rate estimates = 4.58e-05). The basic vocabulary changed at a median rate of 1.48e-05 (SD = 4.00e-05), while the grammatical features changed at a faster median rate of 7.93e-05 (SD = 5.74e-05). The mixture model analyses identified a posterior mean of three different rate categories across the lexical and structural data. These three rate categories can be thought of as "fast," "medium," and "slow" rates (we stress that these rates are relative to the features in this analysis and language family and not statements about the universality of these categories). Fig. 2 shows the average proportion of cognates and structural features falling into each of the three rate categories estimated in the posterior distribution. This is calculated by conditioning on the MCMC steps with the number of rate categories equal to the estimated posterior mode. For each of those steps, the proportion of cognates in each rate category is calculated and then averaged across all of the conditioned steps.

The posterior mean rates indicate that, overall, 119 of the 1,352 variables in the analysis fall into the fast rate category (8.80%), 204 of 1,352 (15.09%) were identified as a medium rate, and 1,029 of 1,352 (76.11%) fall within the slow rate category. The items identified as falling into the slow rate category were overwhelmingly lexical: 982 of 1,029 (95.43%) basic vocabulary vs. 47 of 1,030 (4.57%) grammatical. The slow, medium, and fast rate proportions in the grammatical data are, respectively, 47 of 157 (29.94%), 65 of 157 (41.40%), and 45 of 157 (28.66%), while for the lexicon, the proportions are 982 of 1,195 (82.18%), 139 of 1,195 (11.63%), and 74 of 1,195 (6.19%). The rate category distribution is fairly uniform for the grammatical data, whereas the distribution of lexical data has a greater weight on the slow category.

Homoplasy. We measured the amount of conflicting signal in these data using two metrics, the δ-score and Q-residual (22, 23), which score each language from 0 (lower conflict) to 1 (higher conflict). The median δ-score for the lexical data was 0.38 (SD = 0.024), and the median δ-score for the structural data was 0.44 (SD = 0.020) (Fig. 3). The median Q-residual was 0.0062 (SD = 0.0010) for the lexicon, and the median Q-residual for the structural data was 0.0354 (SD = 0.0042). According to both of these metrics, the lexical data show significantly lower levels of homoplasy than the structural data (Wilcoxon signed rank test, lexicon: $Z = -10.33$, $P < 0.001$, $r = -1.15$; structure: $Z = -10.99$, $P < 0.001$, $r = -1.22$). This difference is especially marked in the Q-residual, where the values differ by more than an order of magnitude. The correlation between the lexical and structural homoplasy scores was not significant for either the δ-score or the Q-residual.

Schismogenesis. We tested whether there were significant effects of schismogenesis by analyzing the posterior probability distribution of tree shapes and branch lengths estimated from either the lexical or structural data alone. The proportion of trees from the posterior with a significant effect of nodes on path length was high (lexicon = 100%, structure = 86.6%), and there was little support for a strong effect of the node density artifact identified by the δ-test (24, 25) in either the lexical data or the structural data (with $\delta < 1$ and significant β-scores in 93.5% and 86.6% of the trees, respectively). The amount of evolution in the lexical data attributable to punctuational effects was almost twice that of the structural data (29.10%, SD = 3.1 vs. 15.15% SD = 4.0).

Discussion

Grammatical vs. Lexical Rates. We find striking differences in the overall pattern of rates between the basic vocabulary and the grammatical features. On average, the grammatical features changed faster ($n = 157$, median rate = 7.93e-05, SD = 5.74e-05 changes per feature per year) than the basic vocabulary cognates ($n = 1195$, median rate = 1.48e-05, SD = 4.00e-05). The rule of

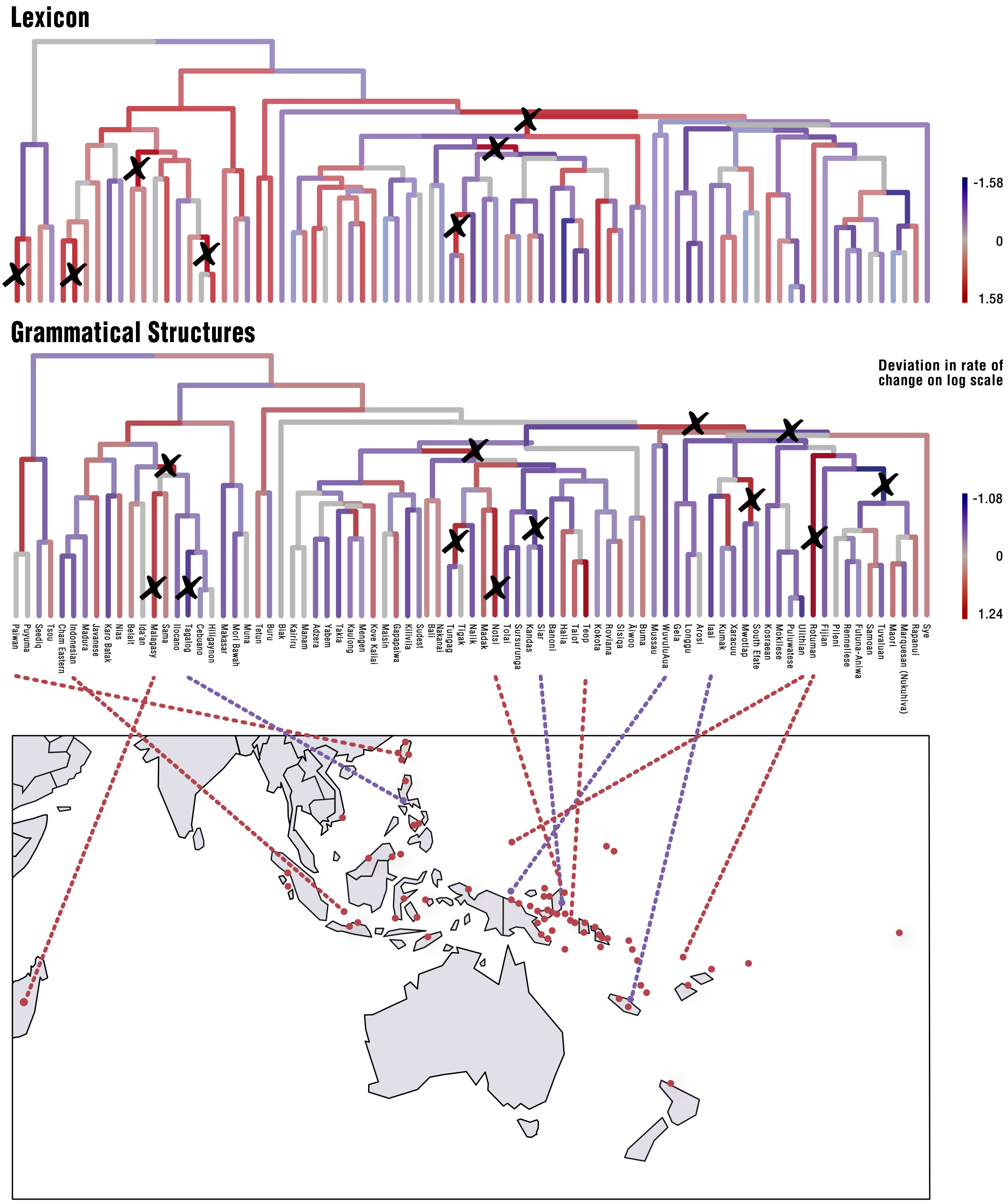

Fig. 1. Map showing locations of languages in this study. The phylogenies show the maximum clade credibility tree of the Austronesian languages in our sample. Each phylogeny is colored by the average rate of change, with branches showing more change colored redder, while bluer branches show reductions in rate. Branches with significant shifts are annotated with an asterisk, and the languages showing significantly different rates of change in their grammatical data are located on the map.

thumb in historical linguistics is that 20% of the basic vocabulary lexicon is replaced every 1,000 y, for a per-lineage rate of 0.02 (26). However, the trees we use here represent an average of 172,353 y of language evolution such that, on average, a given lexical cognate is expected to undergo a median of 2.5 changes, while the grammatical features undergo a median of 13.7 changes.

Table 1. Model fits of the rates analyses using the AICM

Dirichlet distribution	Clock model	AICM (SE)	Difference
Lognormal	Relaxed	48,935.85 (±11.82)	—
Exponential	Relaxed	48,956.74 (±11.86)	−20.90
Lognormal	Strict	49,389.53 (±10.24)	−453.68
Exponential	Strict	49,400.38 (±9.52)	−464.53

The grammatical features were relatively evenly spread across the three rate categories with 47 of 157 (29.94%) in the slow category, 65 of 157 (41.40%) in the medium category, and 45 of 157 (28.66%) in the fast category. However, the basic vocabulary characters were predominantly placed in the slow rate category, 982 of 1,195 (82.18%) characters, over the medium [139 of 1,195 (11.63%)] and fast [74 of 1,195 (6.19%)] rate categories. Overall, 95% of the items identified in the slow rate category were lexical cognates. In the medium and fast rate categories, approximately one-third of the features were structural and two-thirds were cognates. In the lexical data, the majority of the items fall into the slow rate category (Datasets S1 and S2).

This high level of genealogical stability in the basic vocabulary is perhaps expected, as all these items were preselected by their presumed virtue of being stable over time (26). The most stable words include reflexes of "leg," "live," and "hand." In contrast to previous work that has estimated the rate of change for word meaning categories (27, 28), we infer rates for cognate sets within meaning categories. However, some comparison with previous work is possible by averaging the rate category across each meaning category. For example, Dyen et al. (27) find that words for "five," "two," "eye," and "we" are highly stable in Austronesian languages. We find that "five" and "we" are highly stable, while we identify "two" and "eye" as medium rate items. Overall we find low similarity in word rates between Dyen et al.'s results (27) and ours (n = 156; $\rho = 0.26$, $P < 0.0001$). Interestingly, a comparison of our estimated rates with a study of rates in the Indo-European languages (28) shows no significant relationship despite claims for universality of rates (n = 146; $\rho = -0.11$, P = not significant). It is unclear whether this is due to differences in methodology or measurement schemes (cognate rates vs. word rates; *SI Materials and Methods* and Fig. S1) and suggests that more work needs to be done to explore the temporal dynamics of lexical evolution across a wider range of language families.

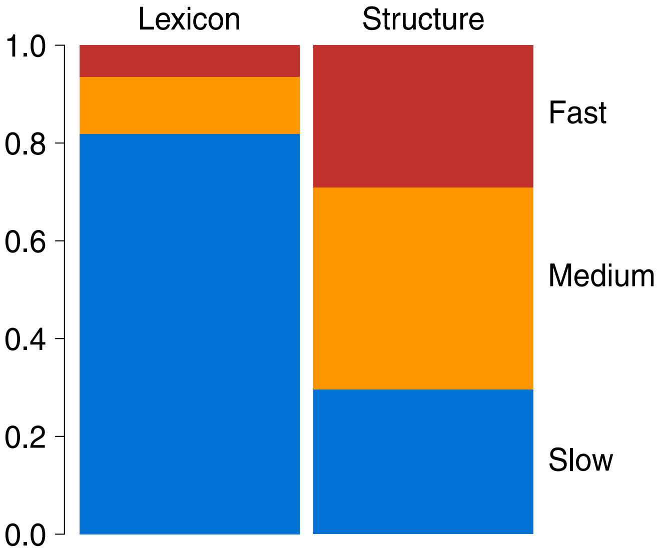

Fig. 2. Proportion of characters in each dataset falling into each of the three posterior rate categories.

At first glance, our finding that structural features evolve more rapidly than basic vocabulary is incompatible with hypotheses proposing extremely deep signal in grammatical features. However, we do identify a set of highly stable structural features (Dataset S2). The highly stable features include many that have been previously suggested (5, 29), such as inclusive vs. exclusive distinctions and gender distinctions. For example, gender distinction in third person, and gender distinction in third person only are highly stable, while the presence of gender distinctions is in the medium rate category. Other features suggested to be unstable also appear in the fast rate category; for example, the presence of numeral classifiers is rapidly changing consistent with predictions that these are highly areal features (30). Suggestions that definite articles are unstable are also borne out by our results (31).

In contrast, however, other structural features predicted to be stable are not identified as such by our results. For example, case systems are proposed to be stable over time (30), but the four case-marking features (presence of case marking on core nominal noun phrases, on oblique nominal noun phrases, on core pronouns, and on oblique pronouns) all fall into the medium rate category. Likewise, constituent order features (order of numeral and noun, order of subject and verb), which have been claimed to be stable cross-linguistically (31), are allocated to the fast or medium rate category in this analysis. A major difference between our results and these studies is that we precisely estimate rates of change within a single, well-known language family, whereas the other analyses produce aggregate measures estimated from a phylogenetically disparate sample (30, 31). Such differences raise the possibility that features may vary in genealogical stability across different linguistic lineages (13), and suggest that future work should take a dynamics approach to language stability and attempt to identify the situations in which features are stable and those in which they not.

The structural rate class that a feature belongs to cannot be predicted from the structural domain: Features involving structure of the verb, the noun, and argument structure are all distributed between the three different rate classes. There are some generalizations that can be made. Features to do with the relative order of elements tend to fall in the fast rate category. The medium and slow rate classes tend to include highly abstract features, such as those asking whether a particular grammatical category is relevant to the morphosyntax of the language (e.g., "Is there future tense regularly marked on the verb?" "Is there an inclusive/exclusive distinction?"). Features detailing the conflation of grammatical categories also enter into the medium and slow classes. Examples include "Do intransitive subject and transitive object operate in the same way, and differently from transitive subject, for the purpose of any syntactic construction?" and "Are second and third persons conflated in nonsingular numbers?"

What might be driving these differences in rates between grammatical features? One intriguing feature of our results is that the more slowly evolving structural features seem to be more abstract and less available to speaker reflection. These covert features are the ones less "attractive" to being copied between languages due to their deeper integration into the language system, reduced perceptibility, or lower transparency (32, 33). Here, the increased stability would be due to lower rates of transmission across language boundaries. Alternately, these more covert features are less available to "sociolinguistic reflection" (34, 35)*, and therefore less likely to be recruited to demarcate social groups. Here, the increased

*Labov W (1993) The unobservability of structure and its linguistic consequences. *Twenty-Second New Ways in Analyzing Variation Conference* (University of Ottawa, Canada).

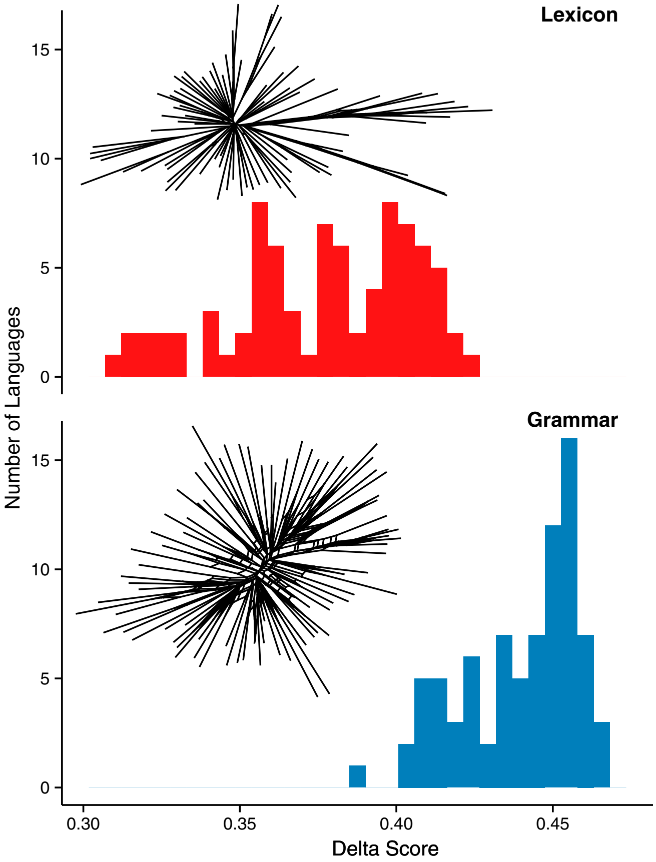

Fig. 3. Histograms of median δ-scores for each language calculated from the lexical and grammatical data. Larger scores indicate more reticulation. (*Inset*) NeighborNet network visualizations of the conflicting signal in these data, where edge lengths are proportional to support in the data and larger boxes indicate more conflicting signal.

stability would be due to lower rates of speaker-driven change. Research in contact linguistics and sociolinguistics suggests that neither of these possibilities is clear-cut (6, 33, 36), but perhaps the variation along these two dimensions of attractiveness and sociolinguistic awareness may play a role in shaping rates of change over thousands of years. One implication of this finding is that more covert features should be superior for tracing language history, and typological questionnaires could be designed with this in mind. Future work should formally explore the link between rates and covertness, integration, and usage in other language families to see if this prediction holds, and whether sociolinguistic and contact processes can shape language evolution over thousands of years.

Differing Language Dynamics in Different Language Systems. Why might the structural data be changing faster than the basic vocabulary data? One obvious difference between the two types of data is that, on average, the structural data show a much higher level of conflicting signal than the lexicon. The differing amount of conflict can be seen visually in NeighborNet figures in Fig. 3, and is quantified in the substantial differences in both the δ-scores and Q-residuals (22, 23). The higher levels of homoplastic change are one factor that might account for the higher rates of change. There are at least three possible reasons why structural features of language might evolve in a less treelike way: parallel evolution, areal diffusion, and multiple histories.

Identifying lexical cognates relies on identifying systematic sound correspondences between languages within similar semantic categories. The requirement that cognate forms are linked by both sound and meaning reduces the number of possible chance similarities (37). Grammatical features lack this constraint. Moreover, the limited design space for grammatical characters reduces the number of possible states dramatically. This poverty of choice subsequently increases the chances of parallel evolution of unrelated languages into the same state. In sentence word order, for example, there are only six possible configurations for the order of subject, object, and verb; thus, on the face of it, the probability of chance similarity is 1/6. However, some of these six configurations are much more common than others (e.g., subject-object-verb and subject-verb-object are much more common than verb-subject-object and verb-object-subject, while object-verb-subject and object-subject-verb are vanishingly rare) (38). Thus, even the 1/6 probability is an underestimation of the chance of parallel evolution. This "poverty of choice" regarding possible feature states in grammatical data (12) is the norm rather than the exception. Consider one of the other features coded in the structural data: whether there is a tonal system or not. As more than half of the world's languages use tonal contrasts and there are multiple ways of developing and expressing tone (39), again, the chances of parallel evolution are high.

The differing desiderata of these two types of data lead to drawbacks for both. Grammatical structures are generalizable and might be recognizable across putatively related languages at deeper time depths, but this increases the risk of chance similarity. The lexical cognates, on the other hand, do trace history well, but only at a rather shallow level. In addition, cognates are not easily identifiable without substantial descriptive work and detailed knowledge of the phonology and morphology of the languages of interest. Unfortunately, the required descriptive work is seriously lacking for most languages of the world (40).

The second possible explanation for the striking differences in rates between the lexicon and the structural data are that the grammatical structures may diffuse more readily between languages, and thus increase the rates of homoplastic changes in phylogenetic analyses. While aspects of the total lexicon are often borrowed across languages (41), basic vocabulary has been found to be relatively resistant to diffusion between lineages than the wider lexicon (42, 43). Similarly, while in most cases the diffusion of structures between languages is thought to be rare, there is evidence that diffusion readily happens in situations of long-term intimate contact and bilingualism (6, 7, 32, 33).

To quantify the relative regional patterns of lexical and structural change, we identified the branches in the posterior probability distribution that showed significant increases or decreases in rates of change. These results are visualized on the two trees in Fig. 1. This visualization shows that different regions of the Pacific have different patterns of relative lexical change or restructuring. For example, the Indonesian/Malay languages tend to show increased rates of lexical change, while the languages through the Bismarck archipelago have considerably more structural change, as expected, due to the increased contact with non-Austronesian languages in this region (e.g., ref. 7). These findings suggest contact effects tend to play out over small distances around contact "hotspots," and act on a limited number of grammatical and lexical features of regional importance unless there is substantial and deeply persistent language contact.

The third possible explanation for the rate differences is that the entire language system may not evolve in concert but rather contain different aspects that each have their own history. Does every word (or feature) have its own history as the famous dialectologists' slogan states? This is not quite the case, but our results suggest that the lexical and grammatical features have rather different histories driven by the different dynamics that are shaping them. This disconnect is clearly demonstrated by the rates, the homoplasy, and contact results.

Finally, we quantified the amount of punctuated evolution in these two datasets. Punctuated evolution occurs when lineages experience a burst of change associated with the formation of a new language. Earlier work has identified the presence of these punctuational effects in both lexical (44) and grammatical data (31). Here, we find strong evidence of this effect, but the birth of new languages is associated with almost twice as much change in the lexicon as in the grammatical data [29.10% (SD = 3.1) vs. 15.15% (SD = 4.0)]. There are two proposed reasons for punctuated evolution in linguistics (44): that language formation is associated with incomplete sampling of linguistic variants such that the new population loses some features ("founder effect") or that speakers act to differentiate themselves from their sister languages [schismogenesis (17)]. The impact of linguistic founder effects is strongly contested (45), and it is unlikely that they play a major role in shaping the structural rates as even a subsample of the speaker population will contain the full grammatical repertoire. However, lexical items, especially rarer ones, could well be forgotten by a small population of speakers. Alternatively, if speakers are actively modifying their languages to differentiate themselves, then there should be more change in the more sociolinguistically salient items. The less salient items of the lexicon and grammar should therefore be less likely to be recruited to differentiate languages (and if the lexical data have increased amounts of change at the locus of language formation, then this would help strengthen their phylogenetic patterning). Naturally these overall rates will be mediated by other processes shaping rates like frequency of use (28) and language contact (6), but this may provide another instance of demographic processes affecting rates of language evolution (46, 48).

Broader Implications for Tracing History. Our results suggest that the lexical and grammatical data have different evolutionary dynamics. It is clear that grammatical data are not necessarily a better source of information for resolving deep linguistic history. However, we do find some grammatical and lexical features are highly stable over time. This finding is consistent with the hints of deeper signal identified in earlier work on grammatical data (16, 31), and with suggestions that some basic vocabulary items have half-lives of around 10,000 y (28). We suggest that the path forward is to move beyond attempts to find highly stable "magic bullets" or to claim complete primacy for either type of data. There is substantial variation in rates and stability in both the structural and lexical data, and this stability may well vary substantially across language families and over time (13, 16). One of the major benefits of grammatical data is that such data enable comparison across unrelated languages (3). However, this benefit comes with the drawback of an inflated rate of false-positive deep links due to higher rates' parallel changes. One approach that could overcome this problem would be to combine both grammatical and basic vocabulary data into a single partitioned analysis with information on known shallow language relationships incorporated as a prior. This prior constraint, along with nuanced models of language change (47), would effectively down-weigh the contribution of features prone to high rates of parallel change, and thus increase the probability of genuine deep genealogical signals being recovered in the analysis. Such analyses offer a potentially powerful way forward for pushing back the time barrier while taking into account the different language subsystem histories that make up a given language (48).

Materials and Methods

Structural Data. The grammatical data were sourced from the Pioneers of Island Melanesia database (PIMdb) that was collected as part of the Pioneers of Island Melanesia project (2002–2005), focusing on the Papuan and Oceanic/Austronesian languages. The database was greatly extended to other Austronesian languages in a follow-up "Sahul" project (2006–2009) to contain data coded from the 81 Austronesian languages included in this study. The structural features coded in the PIMdb were selected to give a typological overview of the language at a level of detail such that it should be possible to code the features from a good-quality sketch grammar. The selection of characters was biased toward what is known about the typological diversity of Oceania and (in later iterations, of the questionnaire) the Sahul area. All characters represent abstract features of language, coded independent of the lexical and morphological form (49). Analysis of these data using a phylogenetic clustering method was able to recover known language families and their major subgroups, as well as proposing plausible affiliations for as-yet-uncategorized languages (3). This particular coding of grammatical structure should not be taken as representing the only way to do it. A typological questionnaire would have different properties if it were designed to solely survey Austronesian languages. The PIMdb data for these Austronesian languages are available in *SI Materials and Methods* and Dataset S3.

Lexical Data. We used the ABVD (17) to find lexical data for the languages in the PIMdb. We identified 81 languages in both the PIMdb and the ABVD. The ABVD contains word lists of 210 items of basic vocabulary (e.g., simple verbs and nouns, colors, numbers, body parts, kinship terms). These items of basic vocabulary are thought to be highly stable across languages and resistant to borrowing (26, 42, 43). We identified the cognate (homologous) words in these 81 languages using the linguistic comparative method to identify systematic sound correspondences (37), in consultation with experts in Austronesian languages (50). All lexical and cognate information is available online at https://abvd.shh.mpg.de/austronesian, and the cognate file is available in *SI Materials and Methods* and Dataset S4.

There were four languages in the PIMdb that were not available in the ABVD; to maximize the overlap in the two datasets, we used the lexical data from their closest sister language from the same language subgroup (Table S1). For example, the PIMdb contains Ulithian, which we matched to the ABVD's Woleian because they are sister taxa and share 85% intelligibility. Other languages that are affected here are the PIMdb's Sursurunga matched to the ABVD's Patpatar, Notsi matched to Lihir, and Adzera matched to Wampar (Table S1).

Data Selection and Coding. For both datasets, we removed characters that were identified as all missing (i.e., had no state information) or were singletons (i.e., were present in one language only as a unique state). These singletons were removed as they provide no subgrouping information and run the risk of unequal data collection (e.g., where languages have more comprehensive dictionaries available, it is easier to add more synonyms). Each dataset was recoded into a binary presence or absence coding to minimize inconsistency across coding schemes, essentially asking if a given grammatical feature or lexical cognate set was present or not. Following this coding scheme, we extracted 157 grammatical features and 1,195 lexical cognate sets (*SI Materials and Methods* and Fig. S2).

Inferring Genealogical Stability of Features. To estimate the relative rates of the lexical and structural data, we fitted a Dirichlet process mixture model (51) to these data to estimate the number of rate categories in a Bayesian phylogenetic framework (18). This method coestimates the assignments of cognates to rate categories and the rate of each category. The phylogenetic parameters (tree topology, node heights) are also jointly estimated. The mixture model was implemented in the substBMA package (18) using the BEAST2 framework (52).

The Dirichlet process (53) mixture model is a mixture model where the mixture corresponds to the multiple rate categories to be estimated. The Dirichlet process, which can be denoted as DP(α, G_0), is a distribution over distributions and has two components, the concentration parameter, α, and the base distribution, G_0. Here, we have considered two options for the base distribution: (*i*) the exponential distribution and (*ii*) the lognormal distribution. The direct prior on the rate of the *i*th cognate, r_i, is generated by the following procedure:

$$G \sim \mathrm{DP}(a,\ G_0) \quad \text{and}$$
$$r_i \sim G,$$

where G is a discrete distribution and the supported values of that distribution are random draws from the base distribution. Increasing α leads to increasing the number of categories. As in general practice, G is a latent variable that is not sampled directly and is integrated out in our analyses.

We explored two different distributions for the base measure of the Dirichlet process: the exponential distribution and the lognormal distribution. Diffuse priors have been applied to the parameters of the base measure. If the base measure is an exponential distribution, the mean parameter of the exponential distribution has a prior distribution defined by a gamma-distribution with its shape and rate set to 0.001. If the base measure is a lognormal distribution, the prior on the log-space mean is a normal distribution with a mean of 0 and a SD of 1,000, while the prior on the log-space precision is a gamma-distribution with its shape and rate equal to 0.001. The (log-space) precision parameter is the inverse of the (log-space) variance.

To ensure that the rates comparison was as fair as possible to both the lexical and structural data, we constrained the tree topology to match the expected language subgroups (54, 55). Constraining the tree in this way minimizes any bias toward the lexical data, as there are many more lexical cognates than there are structural items, and forces the tree to match the results of the linguistic comparative method based on total evidence.

To scale the trees according to time, we used two different clock models and a continuous time Markov model of cognate change (CTMC) (50). The first clock model was a strict clock, which assumes a constant rate of feature replacement over time. The second clock model was a relaxed clock, which allows for variation in the rate of feature change across lineages (56). The relaxed clock model assumes that rates can vary freely across the tree. The divergence times of the phylogeny are assumed to follow a Yule process as all languages were contemporaneous.

To calibrate these clocks, we incorporated historical evidence of language divergence times as described by Gray et al. (50). We implemented five calibrations on the tree using normally distributed priors on the node heights. These calibrations were the following: (*i*) Proto-Oceanic (mean of 3,300 y, SD = 100 y), (*ii*) Proto-Central Pacific (mean of 3,000 y, SD = 100), (*iii*) Proto-Malayo-Polynesian (mean of 4,000 y, SD = 250), (*iv*) Proto-Micronesian (mean of 2,000 y, SD = 100), and (*v*) Proto-Austronesian (mean of 5,200 y, SD = 300).

The dataset has been analyzed with each of the four combinations of base-measure distributions and clock models. For each of the four combinations, we ran five MCMC chains of 50 million steps (Datasets S5–S8). After removing the first 10% of each analysis as burn-in, we sampled 10,000 generations from the posterior probability distribution across all replicates. For each analysis, the values of the effective sample size are all >100 for the phylogenetic estimations and Dirichlet process mixture model parameters, indicating that each analysis had converged. The maximum clade credibility tree of the best-fitting analysis is presented in Fig. S3.

Quantifying Homoplasy. To estimate the degree of conflicting signal, or homoplasy, in each language, we used two metrics: the δ-score (23) and the Q-residual (22). These two metrics were calculated in SplitsTree v4.13.1 (57) using "uncorrected *P*" (= Hamming) distances.

Inferring Punctuational Effects. To test whether punctuational effects have strongly shaped these languages, we constructed trees for the lexical and structural data separately using BEAST2 (52). To allow the branch lengths to be estimated most accurately, we fitted a relaxed clock model and CTMC model to these data, with constraints and calibrations on the trees as above, and obtained trees with branch lengths proportional to the amount of change in each lineage (i.e., substitution-scaled branches).

We fitted a linear model to the number of nodes and amount of phylogenetic change (i.e., pathlength) in these trees using a generalized least squares framework (24, 25, 44) implemented online at www.evolution.rdg.ac.uk/pe/. This approach estimates the slope β of the relationship between nodes and pathlength, while controlling for phylogenetic similarity, and tests for significant effects using a likelihood-ratio test. From the posterior tree distributions, we sampled 800 trees to test to incorporate uncertainty in the tree topology. This method is sensitive to the "node density" artifact, where regions of the phylogeny with increased lineage sampling can show longer branches, which may resemble a punctuational effect. This artifact can be identified by a curvilinear relationship between the number of nodes and pathlength and can be detected using a δ-test, where significant node density artifacts occur when $\delta > 1$ (25).

In the lexical data, we find strong significant evidence of punctuational evolution ($\beta > 0$) in 100% of the trees and in 93.5% of those showing no evidence for the artifact ($\delta < 1$). We also find evidence of punctuational evolution in the structural data, with 86.6% of the trees showing significant $\beta > 0$ and all of those 86.6% showing no evidence of the artifact. The percentage of evolution on the tree attributable to punctuational effects in the lexical data was 29.10% (SD = 3.1), while in the structural data, it was 15.15% (SD = 4.0).

Quantifying Relative Patterns of Lexical and Grammatical Structure Change. To test whether a language lineage had a significantly higher or lower rate of lexical or structural change, we took the posterior probability distribution from the previous analysis. Using the posterior, we estimated the parameters of the lognormal distribution where the estimated branch rate is independently and identically drawn from the discretized lognormal distribution a priori (56). For each posterior sample, we calculated the lower tail probability of the rate of each branch from this empirical lognormal distribution. The average value of these lower tail probabilities on each branch was used to identify branches at the extremes of the distribution with lower (≤0.05) or higher (≥0.95) rates. These results are plotted in Fig. 1, with each phylogeny colored according to the deviation in rate of change on a logarithm scale, where the rate is the mean rate across all posterior samples after burn-in. Branches with significant changes are annotated with an asterisk, and languages showing significant structural changes are identified on the map.

ACKNOWLEDGMENTS. We thank Lindell Bromham, Nick Evans, Martin Haspelmath, Miriam Meyerhoff, Ger Reesink, Malcolm Ross, Hedvig Skirgård, and two anonymous reviewers for discussion. S.J.G. was supported under the Australian Research Council's Discovery Projects funding scheme (Grants DE120101954 and CE140100041), and C.-H.W. is funded by the Medical Research Council and Cancer-Research UK (Grant ANR00310).

1. Ringe D (1995) "Nostratic" and the factor of chance. *Diachronica* 12:55–74.
2. Gray R (2005) Evolution. Pushing the time barrier in the quest for language roots. *Science* 309:2007–2008.
3. Reesink G, Singer R, Dunn M (2009) Explaining the linguistic diversity of Sahul using population models. *PLoS Biol* 7:e1000241.
4. Dunn M, Levinson SC, Lindström E, Reesink G, Terrill A (2008) Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84:710–759.
5. Nichols J (1992) *Linguistic Diversity in Space and Time* (Univ of Chicago Press, Chicago).
6. Thomason SG, Kaufman T (1988) *Language Contact, Creolization, and Genetic Linguistics* (Univ of California Press, Berkeley, CA).
7. Ross MD (1996) Contact-induced change and the comparative method: Cases from Papua New Guinea. *The Comparative Method Reviewed*, eds Durie M, Ross MD (Oxford Univ Press, Oxford), pp 180–218.
8. Nichols J (1994) The spread of language around the pacific rim. *Evol Anthropol* 3: 206–215.
9. Dunn M (2009) Contact and phylogeny in Island Melanesia. *Lingua* 119:1664–1678.
10. Aikhenvald AY (2003) Mechanisms of change in areal diffusion: New morphology and language contact. *J Linguist* 39:1–29.
11. Bickel B, Nichols J (2005) Inclusive/exclusive as person vs. number categories worldwide. *Clusivity*, ed Filimonova E (Benjamins, Amsterdam), pp 47–70.
12. Harrison SP (2003) On the limits of the comparative method. *The Handbook of Historical Linguistics*, eds Joseph BD, Janda RD (Blackwell, Malden, MA), pp 213–243.
13. Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011) Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473:79–82.
14. Wichmann S (2014) Diachronic stability and typology. *Routledge Handbook of Historical Linguistics*, eds Bowern C, Evans B (Routledge, London), pp 212–225.
15. Pawley A (1970) Grammatical reconstruction and change in Polynesia and Fiji. *Pacific Linguistic Studies in Honour of Arthur Capell*, eds Wurm SA, Laycock DC (Pacific Linguistics, Canberra, Australia), pp 301–367.
16. Greenhill SJ, Atkinson QD, Meade A, Gray RD (2010) The shape and tempo of language evolution. *Proc R Soc B Biol Sci* 277:2443–2450.
17. Greenhill SJ, Blust R, Gray RD (2008) The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evol Bioinform Online* 4:271–283.
18. Wu C-H, Suchard MA, Drummond AJ (2013) Bayesian selection of nucleotide substitution models and their site assignments. *Mol Biol Evol* 30:669–688.
19. Bateson G (1935) Culture contact and schismogenesis. *Man (Lond)* 35:178–183.
20. Baele G, et al. (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* 29: 2157–2167.
21. Burnham KP, Anderson DR (1998) *Model Selection and Inference—A Practical Information-Theoretic Approach* (Springer, New York).
22. Gray RD, Bryant D, Greenhill SJ (2010) On the shape and fabric of human history. *Philos Trans R Soc Lond B Biol Sci* 365:3923–3933.
23. Holland BR, Huber KT, Dress A, Moulton V (2002) Delta plots: A tool for analyzing phylogenetic distance data. *Mol Biol Evol* 19:2051–2059.
24. Webster AJ, Payne RJH, Pagel M (2003) Molecular phylogenies link rates of evolution and speciation. *Science* 301:478–478.
25. Venditti C, Meade A, Pagel M (2006) Detecting the node-density artifact in phylogeny reconstruction. *Syst Biol* 55:637–643.
26. Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. *Int J Am Linguist* 21:121–137.
27. Dyen I, James AT, Cole JWL (1967) Language divergence and estimated word retention rate. *Language* 43:150–171.
28. Pagel M, Atkinson QD, Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449:717–720.
29. Wichmann S, Holman EW (2009) *Temporal Stability of Linguistic Typological Features* (Lincom Europa, Munich).
30. Nichols J (2003) Diversity and stability in language. *The Handbook of Historical Linguistics*, eds Joseph BD, Janda RD (Blackwell, Oxford), pp 283–310.

31. Dediu D, Levinson SC (2012) Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS One* 7:e45198.
32. Weinreich U (1953) *Languages in Contact: Findings and Problems* (Publications of the Linguistic Circle of New York, New York).
33. Johanson L (2002) *Structural Factors in Turkic Language Contacts* (Curzon, Surrey, UK).
34. Silverstein M (1977) *The Limits of Awareness* (Southwest Education Development Laboratory, Austin, TX).
35. Labov W, et al. (2011) Properties of the sociolinguistic monitor. *J Sociolinguist* 15: 431–463.
36. Meyerhoff M, Walker JA (2013) An existential problem: The sociolinguistic monitor and variation in existential constructions on Bequia (St. Vincent and the Grenadines). *Lang Soc* 42:407–428.
37. Durie M, Ross MD (1996) *The Comparative Method Reviewed* (Oxford Univ Press, Oxford).
38. Dryer MS (1992) The Greenbergian word order correlations. *Language* 68:81–138.
39. Yip M (2002) *Tone* (Cambridge Univ Press, Cambridge, UK).
40. Hammarström H, Nordhoff S (2012) The languages of Melanesia: Quantifying the level of coverage. *Lang Doc Conserv* 5:13–33.
41. Curnow TJ (2001) What language features can be "borrowed"? *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*, eds Aikhenvald AY, Dixon RMW (Oxford Univ Press, Oxford), pp 412–436.
42. Tadmor U, Haspelmath M, Taylor B (2010) Borrowability and the notion of basic vocabulary. *Diachronica* 2:226–246.
43. Greenhill SJ, Gray RD (2012) Basic vocabulary and Bayesian phylolinguistics: Issues of understanding and representation. *Diachronica* 29:523–537.
44. Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M (2008) Languages evolve in punctuational bursts. *Science* 319:588.
45. Hunley K, Bowern C, Healy M (2012) Rejection of a serial founder effects model of genetic and linguistic coevolution. *Proc Biol Sci* 279:2281–2288.
46. Bromham L, Hua X, Fitzpatrick TG, Greenhill SJ (2015) Rate of language evolution is affected by population size. *Proc Natl Acad Sci USA* 112:2097–2102.
47. Greenhill SJ (2014) Demographic correlates of language diversity. *Routledge Handbook of Historical Linguistics*, eds Bowern C, Evans B (Routledge, London), pp 555–578.
48. Reesink G, Dunn M (2012) Systematic typological comparison as a tool for investigating language history. *Lang Doc Conserv* 5:34–71.
49. Dunn M, Terrill A, Reesink G, Foley RA, Levinson SC (2005) Structural phylogenetics and the reconstruction of ancient language history. *Science* 309:2072–2075.
50. Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323:479–483.
51. Huelsenbeck JP, Suchard MA (2007) A nonparametric method for accommodating and testing across-site rate variation. *Syst Biol* 56:975–987.
52. Bouckaert R, et al. (2014) BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Comput Biol* 10:e1003537.
53. Frigyik BA, Kapila A, Gupta MR (2010) Introduction to the Dirichlet Distribution and Related Processes. University of Washington Electrical Engineering Technical Report UWEETR-2010-006. Available at https://www2.ee.washington.edu/techsite/papers/refer/UWEETR-2010-0006.html. Accessed September 24, 2017.
54. Blust RA (2009) *The Austronesian Languages* (Pacific Linguistics, Canberra, Australia).
55. Lewis PM, ed (2009) *Ethnologue: Languages of the World* (SIL International, Dallas, TX), 16th Ed.
56. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
57. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267.
58. Pagel M, Atkinson QDS, S Calude A, Meade A (2013) Ultraconserved words point to deep language ancestry across Eurasia. *Proc Natl Acad Sci USA* 110:8471–8476.