LETTER

doi:10.1038/nature09923

Evolved structure of language shows lineage-specific trends in word-order universals

Michael Dunn[1,2], Simon J. Greenhill[3,4], Stephen C. Levinson[1,2] & Russell D. Gray[3]

Languages vary widely but not without limit. The central goal of linguistics is to describe the diversity of human languages and explain the constraints on that diversity. Generative linguists following Chomsky have claimed that linguistic diversity must be constrained by innate parameters that are set as a child learns a language[1,2]. In contrast, other linguists following Greenberg have claimed that there are statistical tendencies for co-occurrence of traits reflecting universal systems biases[3–5], rather than absolute constraints or parametric variation. Here we use computational phylogenetic methods to address the nature of constraints on linguistic diversity in an evolutionary framework[6]. First, contrary to the generative account of parameter setting, we show that the evolution of only a few word-order features of languages are strongly correlated. Second, contrary to the Greenbergian generalizations, we show that most observed functional dependencies between traits are lineage-specific rather than universal tendencies. These findings support the view that—at least with respect to word order—cultural evolution is the primary factor that determines linguistic structure, with the current state of a linguistic system shaping and constraining future states.

Human language is unique amongst animal communication systems not only for its structural complexity but also for its diversity at every level of structure and meaning. There are about 7,000 extant languages, some with just a dozen contrastive sounds, others with more than 100, some with complex patterns of word formation, others with simple words only, some with the verb at the beginning of the sentence, some in the middle, and some at the end. Understanding this diversity and the systematic constraints on it is the central goal of linguistics. The generative approach to linguistic variation has held that linguistic diversity can be explained by changes in parameter settings. Each of these parameters controls a number of specific linguistic traits. For example, the setting 'heads first' will cause a language both to place verbs before objects ('kick the ball'), and prepositions before nouns ('into the goal')[1,7]. According to this account, language change occurs when child learners simplify or regularize by choosing parameter settings other than those of the parental generation. Across a few generations such changes might work through a population, effecting language change across all the associated traits. Language change should therefore be relatively fast, and the traits set by one parameter must co-vary[8].

In contrast, the statistical approach adopted by Greenbergian linguists samples languages to find empirically co-occurring traits. These co-occurring traits are expected to be statistical tendencies attributable to universal cognitive or systems biases. Among the most robust of these tendencies are the so-called "word-order universals"[3] linking the order of elements in a clause. Dryer has tested these generalizations on a worldwide sample of 625 languages and finds evidence for some of these expected linkages between word orders[9]. According to Dryer's reformulation of the word-order universals, dominant verb–object ordering correlates with prepositions, as well as relative clauses and genitives after the noun, whereas dominant object–verb ordering predicts postpositions, relative clauses and genitives before the noun[4]. One general explanation for these observations is that languages tend to be consistent ('harmonic') in their order of the most important element or 'head' of a phrase relative to its 'complement' or 'modifier'[3], and so if the verb is first before its object, the adposition (here preposition) precedes the noun, while if the verb is last after its object, the adposition follows the noun (a 'postposition'). Other functionally motivated explanations emphasize consistent direction of branching within the syntactic structure of a sentence[9] or information structure and processing efficiency[5].

To demonstrate that these correlations reflect underlying cognitive or systems biases, the languages must be sampled in a way that controls for features linked only by direct inheritance from a common ancestor[10]. However, efforts to obtain a statistically independent sample of languages confront several practical problems. First, our knowledge of language relationships is incomplete: specialists disagree about high-level groupings of languages and many languages are only tentatively assigned to language families. Second, a few large language families contain the bulk of global linguistic variation, making sampling purely from unrelated languages impractical. Some balance of related, unrelated and areally distributed languages has usually been aimed for in practice[11,12].

The approach we adopt here controls for shared inheritance by examining correlation in the evolution of traits within well-established family trees[13]. Drawing on the powerful methods developed in evolutionary biology, we can then track correlated changes during the historical processes of language evolution as languages split and diversify. Large language families, a problem for the sampling method described above, now become an essential resource, because they permit the identification of coupling between character state changes over long time periods. We selected four large language families for which quantitative phylogenies are available: Austronesian (with about 1,268 languages[14] and a time depth of about 5,200 years[15]), Indo-European (about 449 languages[14], time depth of about 8,700 years[16]), Bantu (about 668 or 522 for Narrow Bantu[17], time depth about 4,000 years[18]) and Uto-Aztecan (about 61 languages[19], time-depth about 5,000 years[20]). Between them these language families encompass well over a third of the world's approximately 7,000 languages. We focused our analyses on the 'word-order universals' because these are the most frequently cited exemplary candidates for strongly correlated linguistic features, with plausible motivations for interdependencies rooted in prominent formal and functional theories of grammar.

To test the extent of functional dependencies between word-order variables, we used a Bayesian phylogenetic method implemented in the software BayesTraits[21]. For eight word-order features we compared correlated and uncorrelated evolutionary models. Thus, for each pair of features, we calculated the likelihood that the observed states of the characters were the result of the two features evolving independently, and compared this to the likelihood that the observed states were the result of coupled evolutionary change. This likelihood calculation was

[1]Max Planck Institute for Psycholinguistics, Post Office Box 310, 6500 AH Nijmegen, The Netherlands. [2]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Kapittelweg 29, 6525 EN Nijmegen, The Netherlands. [3]Department of Psychology, University of Auckland, Auckland 1142, New Zealand. [4]Computational Evolution Group, University of Auckland, Auckland 1142, New Zealand.

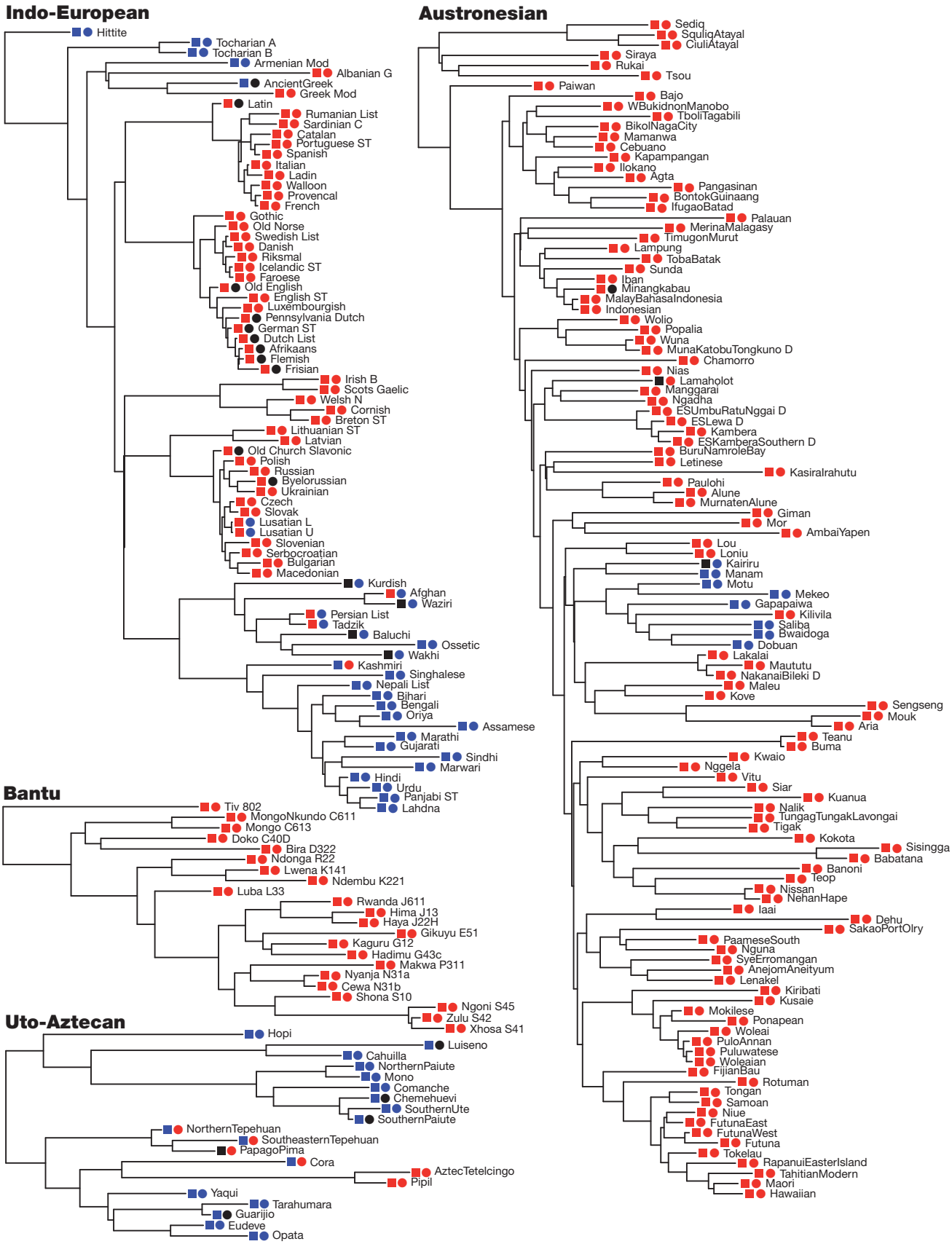

Figure 1 | Two word-order features plotted onto maximum clade credibility trees of the four language families. Squares represent order of adposition and noun; circles represent order of verb and object. The tree sample underlying this tree is generated from lexical data[16,22]. Blue-blue indicates postposition, object–verb. Red-red indicates preposition, verb–object. Red-blue indicates preposition, object–verb. Blue-red indicates postposition, verb–object. Black indicates polymorphic states.

conducted over a posterior probability distribution of phylogenetic trees constructed using basic vocabulary data from each of the language families: 79 Indo-European languages[16,22], 130 Austronesian languages[15,23], 66 Bantu languages[24] and 26 Uto-Aztecan languages (R. Ross & R.D.G., manuscript in preparation). Information on word-order typology was derived partly from the World Atlas of Language Structure database[25] and expanded with additional coding from grammatical descriptions (Supplementary Information section 1.3 and 2). As an illustration, the states of two of these features mapped against a summary of the posterior tree samples for all four language families are shown in Fig. 1. In this case, visual inspection shows that these characters appear to be linked in some families. However, the Bayesian phylogenetic approach allows us to assess this formally by quantifying the relative fits of dependent and independent models of character evolution across all trees in the posterior probability distribution. This method incorporates the uncertainty in the estimates of the tree topology, the rates of change and the branch lengths. The extent to which a dependent model of evolution provides a superior explanation of the variation of word-order features to an independent model is measured using Bayes factors (BF) calculated from the marginal likelihoods over the posterior tree distribution. BF $>$ 5 are conventionally taken as strong evidence that the dependent model is preferred over the independent model[13,26].

The results of the BayesTraits analysis of correlated trait evolution are summarized in Fig. 2. These differ considerably from the expectations derived from both universal approaches. The Greenbergian approach suggests robust tendencies towards linkages due to intrinsic system biases, while the generative approach assumes these will be 'hard' systems constraints set by discrete choices over a small innate parameter set[1,27]. Instead, our major finding is that, although there are linkages or dependencies between word-order characters within language families, these are largely lineage-specific, that is, they do not hold across language families in the way the two universals approaches predict.

Dryer's study of the Greenberg word-order universals[4] across a world-wide sample of related and unrelated languages found a set of dependent word-order relations that show correlations with the order of verb and object, and another set of word-order relations that were independent of this. We extracted from his analyses two predictions of strong tendencies across all languages. First, all the word-order relations in the dependent set should be correlated: these are verb–object order, adposition–noun order, genitive–noun order, relative-clause–noun order. Second, no dependencies are expected between the dependent set and the independent set (including demonstrative–noun, numeral–noun, adjective–noun and subject–verb orders).

Contrary to the first expectation, we found no pairs of word-order features that were strongly dependent in all language families. Only two of these predicted dependencies were found in more than one language family: a dependency between adposition–noun and verb–object order was found in Austronesian and Indo-European, and a dependency between genitive–noun order and object–verb order was found in Indo-European and Uto-Aztecan.

Contrary to the second prediction, we found eight strong dependencies between members of the dependent set and members of the independent set, including two that occurred in two language families. The evolution of adjective–noun order and relative-clause–noun order is correlated within both Austronesian (BF = 5.33) and Uto-Aztecan (BF = 5.02), and the demonstrative–noun, object–verb features are correlated in Bantu (BF = 5.24) and Indo-European (BF = 7.55). Many dependencies are unique to just one language family; for example, only Uto-Aztecan shows strongly coupled (BF = 13.57) changes between subject and object ordering with respect to the verb, only Indo-European shows strongly coupled (BF = 21.23) changes between adjective and genitive ordering, and only Austronesian shows strongly coupled (BF = 18.26) changes between numeral–noun and genitive–noun orders. These family-specific linkages suggest that evolutionary processes of language diversification explore alternative ways to construct coherent language systems unfettered by tight universal constraints. They also demonstrate the power of phylogenetic methods to reveal structural linkages that could not be detected by cross-linguistic sampling.

The lineage-specificity of these dependencies is striking. There is a poor correspondence between dependencies across the families, and even where we find dependencies shared across language families, the phylogenetic analyses show family-specific evolutionary processes at work. Take, for example, the dependency between object–verb and adposition–noun orders shared by two of the language families. Examination of the transition probabilities between linked states reveals that different patterns of change are responsible for the observed linkage in each language family, as shown in Fig. 3. Here changes in the Austronesian family funnel evolving systems towards a single solution, while Indo-European shunts changes towards two solutions. Thus similarities in word-order dependencies may hide underlying differences in how these linkages come about, which once again reflect lineage-specific processes.

If the central goal of linguistic theory is to understand constraints on linguistic variation and language change, then the methods outlined here promise systematic insights of a kind only possible with the recent development of phylogenetic methods and large linguistic databases. As more large linguistic databases become available[28], the approach

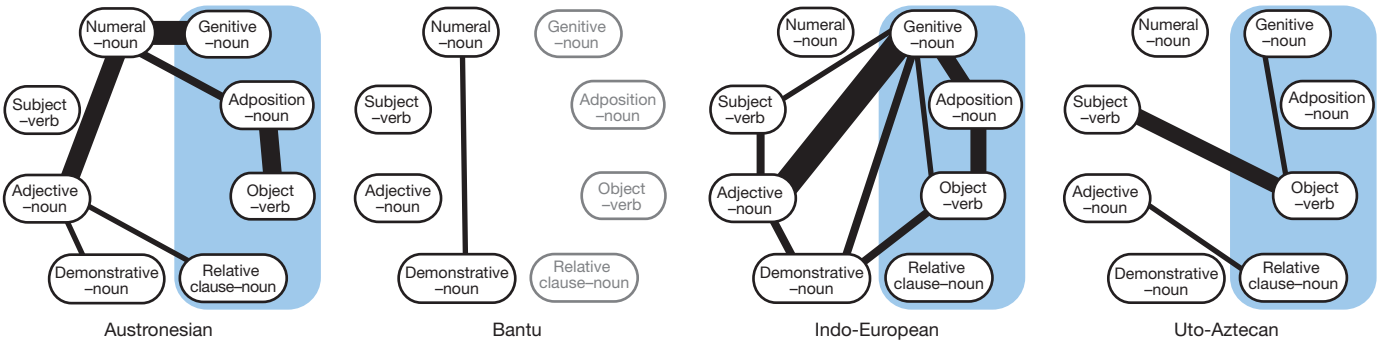

Figure 2 | Summary of evolutionary dependencies in word order for four language families. All pairs of characters where the phylogenetic analyses detect a strong dependency (defined as BF $\geq$ 5) are shown with line width proportional to BF values (indicating a range from 5.01 to 21.23, see Supplementary Information section 3). In the case of the Bantu language family, four invariant features (indicated in grey) were excluded from the analyses. Following Dryer's reformulation of Greenberg's word-order universals, we expected dependencies between all the features in the blue shaded area. However, only two dependencies (object–verb order and adposition–noun order; and object–verb order and genitive–noun order) are found in more than one language family, and no dependencies were found involving relative-clause order and any of the other three features. Of the other thirteen strongly supported dependencies, nine were unexpected (no prediction was made about feature pairs outside the blue area). Most of these 19 dependencies occur in only one language family (three occur in two families, and one in three families).

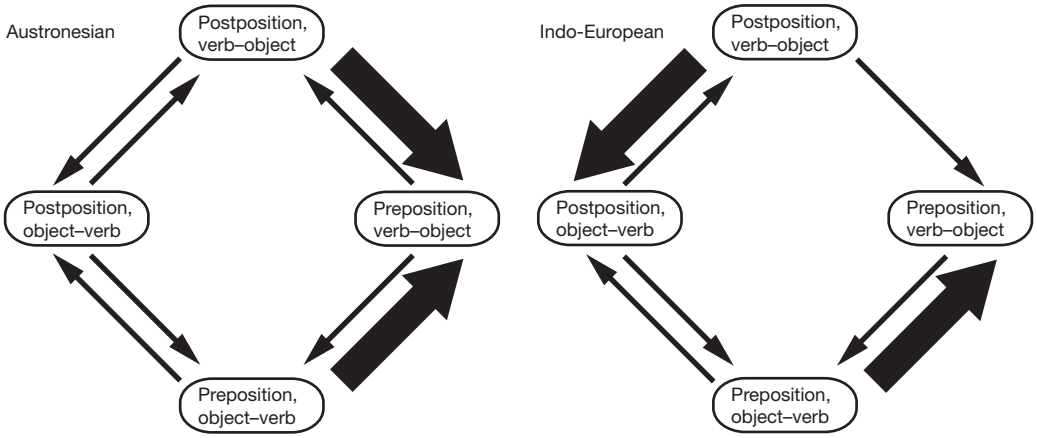

Figure 3 | The transition probabilities between states leading to object–verb and adposition–noun alignments in Austronesian and Indo-European. Data were taken from the model most frequently selected in the analyses; probability is indicated by line weight. The state pairs across the midline of each figure (postposition, object–verb; and preposition, verb–object) are Greenberg's 'harmonic' or stable word orders. Nevertheless, each language family shows tendencies for specific directions and probabilities of state transitions.

developed here could be used to explore the dependency relationships between a wide range of linguistic features. Nearly all branches of linguistic theory have predicted such dependencies. Here we have examined the paradigm example (word-order universals) of the Greenbergian approach, taken also by the Chomskyan approach as "descriptive generalizations that should be derived from principles of UG [Universal Grammar]".[1,27] What the current analyses unexpectedly reveal is that systematic linkages of traits are likely to be the rare exception rather than the rule. Linguistic diversity does not seem to be tightly constrained by universal cognitive factors specialized for language[29]. Instead, it is the product of cultural evolution, canalized by the systems that have evolved during diversification, so that future states lie in an evolutionary landscape with channels and basins of attraction that are specific to linguistic lineages.

Received 10 June 2009; accepted 8 February 2011.
Published online 13 April 2011.

1. Baker, M. *The Atoms of Language* (Basic Books, 2001).
2. Chomsky, N. *Lectures on Government and Binding* (Foris, 1981).
3. Greenberg, J. H. in *Universals of Grammar* (ed. Joseph H. Greenberg) 73–113 (MIT Press, 1963).
4. Dryer, M. in *Language Typology and Syntactic Description* Vol. I *Clause Structure* 2nd edn (ed. Shopen, T.) 61–131 (Cambridge University Press, 2007).
5. Hawkins, J. A. in *Language Universals* (eds Christiansen, M. H., Collins, C. & Edelman, S.) 54–78 (Oxford University Press, 2009).
6. Croft, W. Evolutionary linguistics. *Annu. Rev. Anthropol.* **37,** 219–234 (2008).
7. Chomsky, N. *Knowledge of Language: its Nature, Origin and Use* (Praeger, 1986).
8. Lightfoot, D. W. The child's trigger experience—degree-0 learnability. *Behav. Brain Sci.* **12,** 321–334 (1989).
9. Dryer, M. S. The Greenbergian word order correlations. *Language* **68,** 81–138 (1992).
10. Mace, R. & Pagel, M. The comparative method in anthropology. *Curr. Anthropol.* **35,** 549–564 (1994).
11. Bakker, P. in *Oxford Handbook of Linguistic Typology* (ed. Song, J. J.) (Oxford University Press, 2010).
12. Cysouw, M. in *Quantative Linguistics: An International Handbook* (eds Altmann, G., Köhler, R. & Piotrowski, R.) 554–578 (Mouton de Gruyter, 2005).
13. Pagel, M., Meade, A. & Barker, D. Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* **53,** 673–684 (2004).
14. Gordon, R. G. J. *Ethnologue: Languages of the World* 15th edn (SIL International, 2005).
15. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323,** 479–483 (2009).
16. Gray, R. D. & Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426,** 435–439 (2003).
17. Guthrie, M. *Comparative Bantu* Vol. 2 (Gregg International, 1971).
18. Diamond, J. & Bellwood, P. Farmers and their languages: the first expansions. *Science* **300,** 597–603 (2003).
19. Campbell, L. *American Indian Languages: The Historical Linguistics of Native America* 133–138 (Oxford University Press, 1997).
20. Kemp, B. M. *et al.* Evaluating the farming/language dispersal hypothesis with genetic variation exhibited by populations in the Southwest and Mesoamerica. *Proc. Natl Acad. Sci. USA* **107,** 6759–6764 (2010).
21. Pagel, M. & Meade, A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* **167,** 808–825 (2006).
22. Dyen, I., Kruskal, J. B. & Black, P. An Indo-European classification, a lexicostatistical experiment. *Trans. Am. Phil. Soc.* **82,** 1–132 (1992).
23. Greenhill, S. J., Blust, R. & Gray, R. D. The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evol. Bioinform.* **4,** 271–283 (2008).
24. Holden, C. J. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. R. Soc. Lond. B* **269,** 793–799 (2002).
25. Haspelmath, M., Dryer, M. S., Gil, D. & Comrie, B. *The World Atlas of Language Structures Online* (Max Planck Digital Library, 2008).
26. Raftery, A. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83,** 251–266 (1996).
27. Cela-Conde, C. & Marty, G. Noam Chomsky's minimalist program and the philosophy of mind. An interview. *Syntax* **1,** 19–36 (1998).
28. Reesink, G., Singer, R. & Dunn, M. Explaining the linguistic diversity of Sahul using population models. *PLoS Biol.* **7,** e1000241 (2009).
29. Evans, N. & Levinson, S. C. The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32,** 429–492 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Liberman for comments on our initial results and F. Jordan and G. Reesink for comments on drafts of this paper. L. Campbell, J. Hill, W. Miller and R. Ross provided and coded the Uto-Aztecan lexical data.

Author Contributions R.D.G. and M.D. conceived and designed the study. S.J.G., R.D.G. and M.D. provided lexical data and phylogenetic trees. M.D. coded word-order data, and conducted the phylogenetic comparative analyses with S.J.G. All authors were involved in discussion and interpretation of the results. All authors contributed to the writing with S.C.L. and M.D. having leading roles; M.D., R.D.G. and S.J.G. produced the Supplementary Information.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.D. (michael.dunn@mpi.nl).