ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Perrault N, Farrell MJ, Davies TJ. 2017 Tongues on the EDGE: language preservation priorities based on threat and lexical distinctiveness. *R. Soc. open sci.* **4**: 171218. http://dx.doi.org/10.1098/rsos.171218

Received: 23 August 2017
Accepted: 10 November 2017

Subject Category:
Biology (whole organism)

Subject Areas:
evolution/ecology

Keywords:
conservation, language preservation, biodiversity, phylogeny, evolutionarily distinct and globally endangered, linguistic diversity

Author for correspondence:
T. Jonathan Davies
e-mail: j.davies@mcgill.ca

†Authors contributed equally to this work.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3938971.

THE ROYAL SOCIETY PUBLISHING

Tongues on the EDGE: language preservation priorities based on threat and lexical distinctiveness

Nicolas Perrault[1,†], Maxwell J. Farrell[2,†] and T. Jonathan Davies[2,3]

[1]University of Oxford, School of Archaeology, Oxford, UK
[2]McGill University, Department of Biology, Montréal, Québec, Canada
[3]African Centre for DNA Barcoding, Department of Botany and Plant Biotechnology, University of Johannesburg, Johannesburg, South Africa

NP, 0000-0002-7357-8415; MJF, 0000-0003-0452-6993

Languages are being lost at rates exceeding the global loss of biodiversity. With the extinction of a language we lose irreplaceable dimensions of culture and the insight it provides on human history and the evolution of linguistic diversity. When setting conservation goals, biologists give higher priority to species likely to go extinct. Recent methods now integrate information on species evolutionary relationships to prioritize the conservation of those with a few close relatives. Advances in the construction of language trees allow us to use these methods to develop language preservation priorities that minimize loss of linguistic diversity. The evolutionarily distinct and globally endangered (EDGE) metric, used in conservation biology, accounts for a species' originality (evolutionary distinctiveness—ED) and its likelihood of extinction (global endangerment—GE). Here, we use a similar framework to inform priorities for language preservation by generating rankings for 350 Austronesian languages. Kavalan, Tanibili, Waropen and Sengseng obtained the highest EDGE scores, while Xârâcùù (Canala), Nengone and Palauan are among the most linguistically distinct, but are not currently threatened. We further provide a way of dealing with incomplete trees, a common issue for both species and language trees.

1. Introduction

There is growing evidence that we are in the midst of a sixth mass extinction event and mankind is probably the cause [1,2]. Since the 1950s, scientific and public awareness of the loss of biodiversity has increased considerably [3], but we lack both resources and time to save all endangered species. Some species

will go extinct and we must make choices and set priorities in species conservation [4]. Many human languages are equally, if not more threatened [5]. It is estimated that one of the world's 7000 languages vanishes every other week and half might not survive the twenty-first century [6]. Languages are the spark of a people, the bearing of cultures, and are tied to a special understanding of native environments. Their disappearance is a loss to humanity, scholarship and science [7]. Prehistorians study languages to trace back population movements [8,9] and anthropologists use language trees to test hypothesis of cultural evolution [10,11]. Linguists use the variety of parlances to understand language as a human phenomenon; every single tongue gives them additional insight [7]. Traditional ecological knowledge, often used in biodiversity conservation efforts [12–15], is imperilled if languages are lost [16,17]. The rapid rate of language loss coupled with limited resources for preservation indicates that formal prioritization schemes may be useful tools to maximizing the retention of linguistic diversity.

In conservation biology, there have been efforts to prioritize species based on their evolutionary distinctiveness (ED) with the idea that highly distinct species might have unique traits that contribute to biodiversity [18–20] and that communities that capture greater phylogenetic diversity may enhance ecosystem functioning (e.g. [21,22]). For example, species with many close relatives might provide few unique ecosystem services. Conversely, species with few relatives are usually the most functionally original [20] and may thus provide irreplaceable services (see arguments in [23]). Likewise in linguistics, the more isolated a language is in its family tree, the more unique information it contains and ultimately contributes to linguistic diversity. Prioritizing the documentation of threatened and isolated languages is a key goal in linguistics [6]. Recently developed methods for quantifying similarity among languages [24] offer new opportunities to inform these prioritizations.

In biology, phylogenetic trees (trees of life) depict species ancestor-to-descendant relationships. Two populations of a single species will evolve into two species when gene flow is interrupted, often by geographical isolation [25]. One can consider speciation complete when two populations can no longer interbreed [26]. Speciation is depicted in the tree by the splitting of branches. Likewise, though a simplification, dialects become languages when the speakers of one dialect can no longer understand speakers of the other. Like new species, diverged dialects are splits in a language tree [27].

We can quantify a species' ED by measuring how isolated it is on a phylogenetic tree. Species isolated in the tree are said to be evolutionarily distinct. Similarly, we can quantify linguistic distinctiveness from language trees. Once a set of features is selected and a tree built from them, distinctiveness scores can be calculated and used as empirical and objective estimates of uniqueness among languages. There are many distinctiveness metrics [28,29], but all aim to favour species with a few close relatives.

Early distinctiveness metrics counted only the number of splits in a species' ancestry, giving higher scores to fewer splits [4,30]. Such metrics are highly sensitive to missing data (absent splits in the tree). More recent measures treat the lengths of tree branches as units of distinctiveness, usually counted in millions of years. In these cases, a species' distinctiveness is equal to the length of its branch plus a fraction of that of its ancestors. Like money that people inherit from their mother, fewer siblings mean a larger inheritance. If the mother herself had few siblings, she inherited more from her parents and in turn would have more to leave to her children. Further, with a constant salary, the longer she lived, the more money she would have to leave them. Devised by Redding [31] and employed by Isaac *et al.* [18], we used a metric of ED in which ancestral distinctiveness is divided evenly among all living descendants, although distinctiveness may be calculated in other ways [29,32].

Isaac *et al.* [18] determined the ED from a near-complete species-level phylogenetic tree for mammals with branch lengths proportional to time. Implicit within their calculation is an assumption that species differentiate at a constant rate through time, i.e. that branch lengths measured in evolutionary time capture the expected differences between species. ED, being a weighted sum of branch lengths, also represents time in millions of species-years. The platypus, for example, has an ED of approximately 97.6 million years, the greatest ED in the mammal phylogeny.

The assumption of constant divergence through time, however, does not hold for languages. As Icelandic and Norwegian diverged from Old Norse one thousand years ago, the basic vocabulary of Norwegian has changed five times faster than that of Icelandic [33]. This is not an isolated example—time is a poor estimator of linguistic distinctiveness. A language's ED is better computed from a language tree whose branch lengths convey distinctiveness directly. Here, we use a tree based on the proportion of ancestral words substituted for newer words in a language's basic vocabulary.

To prioritize conservation efforts so as to minimize the expected loss of diversity, distinctiveness can be weighted by the probability of extinction—P(extinction) [34]. To estimate this probability, Isaac *et al.* [18] used the endangerment levels of the IUCN Red List [35], an objective qualitative scale of species extinction risk, assuming each increase in Red List threat category represents a doubling in P(extinction).

Taking a species' ED and global endangerment (GE) as proxies for its contribution to diversity and probability of extinction, a species' EDGE score is calculated as follows:

$$\mathrm{EDGE} = \ln(1 + \mathrm{ED}) + \mathrm{GE} \cdot \ln(2). \tag{1.1}$$

At least four endangerment assessments analogous to the IUCN Red List exist for languages: a list by Sutherland [5], a conservation biologist; UNESCO's atlas of the world's languages in danger [36]; a database by the Endangered Languages Project (www.endangeredlanguages.com); and the EGIDS scale [37] used by the Ethnologue, an online database of 7000 languages [38]. Given a detailed language phylogeny, it is hence possible to apply techniques from conservation biology to language preservation.

Here, we illustrate how the EDGE framework can be applied to linguistic diversity using a tree of several hundred Austronesian languages built on differences in basic vocabulary (210 words), typically stable through time and resistant to borrowing from other languages. This tree represents one of the largest language families in the world, which probably originated from Taiwan 4000–6000 years ago and then rapidly expanded through islands of the Pacific [39]. The exercise of ranking languages with the EDGE metric can identify languages that are both distinct and threatened, which might be considered important targets for documentation and preservation, if not done already. Although we analyse only a subset of Austronesian language diversity, we present a method that corrects for limited sampling, and show that our results are surprisingly robust to missing languages in the phylogeny.

2. Material and methods

2.1. Measuring evolutionary distinctiveness

The tree used in this analysis has 1215 tips representing the 1215 living ISO 639-3 Austronesian languages. The tree is a composite of two datasets: a 350-tip tree with branch lengths from Gray *et al.* [40] and, provided in the electronic supplementary material, language classification data by the Ethnologue. Gray *et al.*'s tree is the core dataset. It is based on lexical data from Greenhill *et al.* [41], consisting for each language of 210 basic words thought to be stable over time and resistant to borrowing. Branch lengths in Gray *et al.*'s tree represent the median number of cognate changes undergone on that branch across trees sampled from a Bayesian posterior distribution. Importantly, Gray *et al.* made no assumption that words change at a constant rate over time.

The tree in Gray *et al.* [40] consists of 400 languages chosen based on data availability and to provide 'a representative sample of each recognized Austronesian subgroup' [40]. From this set we removed 16 languages that were extinct, not Austronesian or without an ISO 639-3 code from the International Organization for Standardization [42]. We further removed 34 Austronesian dialects that shared an ISO 639-3 code with another language in the tree, always keeping the dialect with the greatest ED. This resulted in a tree with 350 languages (hereafter 'Gray *et al.*'s tree').

Of the 1215 living Austronesian languages, 71.2% are not represented in Gray *et al.*'s tree, and missing languages may be expected to affect ED scores. To account for this effect, we complemented the phylogeny with language classification data from the Ethnologue, which groups all ISO 639-3 languages into families and subfamilies. The Ethnologue classification for Austronesian languages was converted into a tree with no meaningful branch lengths, and then missing languages were inserted into Gray *et al.*'s tree (figure 1; details in the electronic supplementary material). This resulted in a 1215-tip tree (hereafter the 'reconstructed full Austronesian tree') used to calculate ED following the fair proportion method devised by Redding [31]. ED was estimated for only those 350 living ISO 639-3 Austronesian languages present in Gray *et al.*'s tree.

2.2. Measuring global endangerment

To measure the probability of extinction, we converted the 10-point EGIDS scale of language endangerment into a Global Endangerment index (GE, table 1). We chose this approach to quantifying GE as it parallels the IUCN Red List [35] and most closely matches to the original EDGE framework published by Isaac *et al.* [18] in which increases of one threat level double the probability of extinction (details and conversion scheme in electronic supplementary material, table S1). Because GE is simply a multiplier in the calculation of EDGE, it would be straightforward to substitute our GE index for alternative estimates of P(extinction), such as those of Sutherland [5], UNESCO's atlas of the world's languages in danger [36], or if data are available, estimations of P(extinction) based on the total

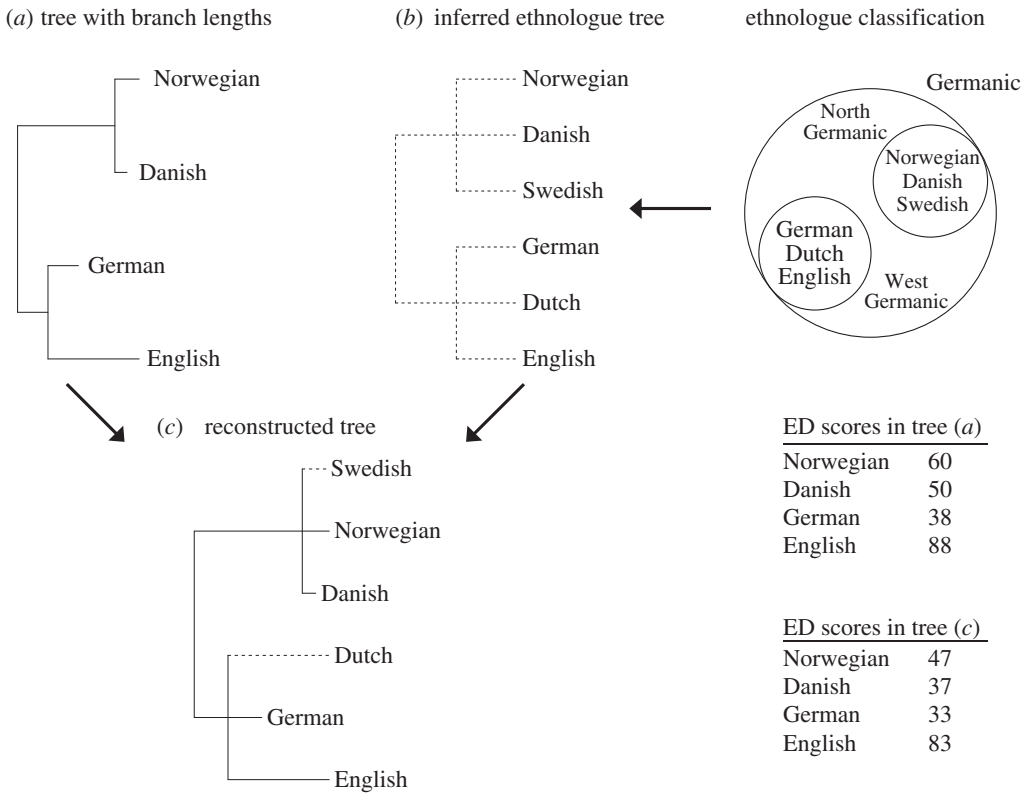

Figure 1. How the Austronesian tree was reconstructed to compute evolutionary distinctiveness more accurately, using Germanic languages as an example. A tree (*a*) of Germanic languages with (here invented) branch lengths can be used to compute evolutionary distinctiveness (ED), but missing languages (Dutch and Swedish) will bias this score. Language classifications into families and subfamilies by the Ethnologue (simplified for illustration) can partially compensate for this bias. It can be used to infer a tree (*b*) with no meaningful branch lengths. Those languages or groups of languages missing from tree (*a*) are imported from tree (*b*) to form a reconstructed tree (*c*). ED, as calculated from tree (*c*), is usually more accurate than when calculated from tree (*a*); see main text. This method does not allow computing ED of languages missing from tree (*a*). In this analysis, we used the Austronesian equivalent of tree (*c*). Details in the electronic supplementary material.

Table 1. Definition of global endangerment (GE) scores for language endangerment. GE is a conversion of the EGIDS endangerment scale that parallels Isaac's conversion of the IUCN Red List, in which increases of one unit in GE represent a doubling in the probability of extinction. The age of youngest users is the most important criterion for the EGIDS scale (details in electronic supplementary material, table S1).

GE	EGIDS endangerment	youngest users	other criteria
4	nearly ext.	grandparents	rarely used
3.5	moribund	grandparents	—
3	shifting	parents	—
2	threatened	children	losing users
1	vigorous	children	stable user base
$\frac{1}{2}$	developing	children	standardized lit.
$\frac{1}{4}$	educational	children	used in schools
$\frac{1}{8}$	wider Comm.	children	used in mass media
$\frac{1}{16}$	provincial	children	local govt. lang.
$\frac{1}{32}$	national	children	national govt. lang.

number of speakers [43]. If uncorrelated with EGIDS, we would expect different endangerment scales to yield different EDGE scores. The EGIDS, however, is the only complete scale for the languages in our sample.

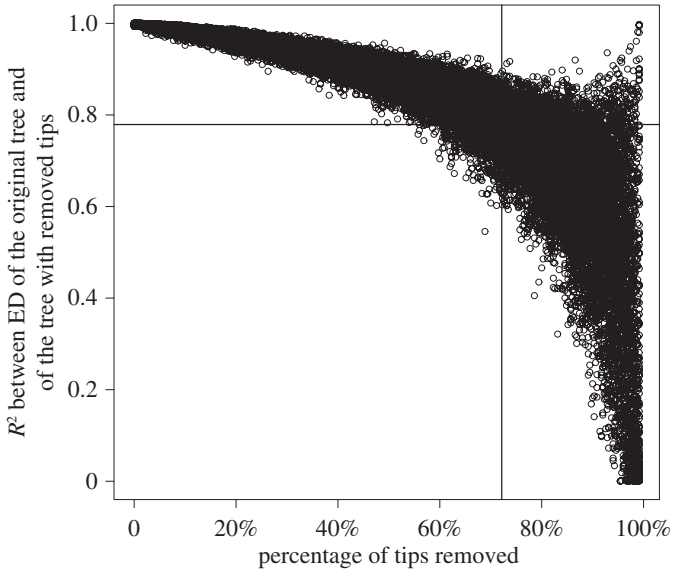

Figure 2. How robust ED is to missing languages. Every one of these 35 000 points represents the R^2 between the ED scores of Gray *et al.*'s tree (350 languages) and the ED scores of one of its subtrees. Each subtree was obtained by randomly removing from Gray *et al.*'s tree a fixed proportion of languages represented on the *x*-axis. Even when 71.2% of tips are removed (vertical line), ED scores correlate well to that of Gray *et al.*'s tree, with $R^2 = 0.78$ on average (horizontal line). If subtrees with 101 languages (i.e. with 71.2% of the 350 languages removed) are reconstructed to 350 languages with Ethnologue data, as detailed in Materials and methods but not depicted here, the average R^2 rises from 0.78 to 0.82. In Gray *et al.*'s tree of 350 languages, 71.2% of the 1215 ISO 639-3 Austronesian languages are missing. We therefore expect that the ED scores of the reconstructed tree used in this analysis are good approximations.

2.3. Effect of missing languages on evolutionary distinctiveness

As mentioned above, missing languages may be expected to affect ED scores, an effect that data from the Ethnologue cannot be expected to correct entirely because of unresolved polytomies. To assess the effect of missing languages, we performed the following sensitivity analyses. Given that 71.2% of the 1215 Austronesian languages are missing in Gray *et al.*'s 350-language tree, we randomly pruned from it 249 languages. This yielded a 101-language tree (hereafter a 'reduced Gray tree'), lacking 71.2% of the languages in Gray *et al.*'s original tree. We then calculated the R^2 between the ED scores from Gray *et al.*'s tree and the reduced Gray tree. We repeated this process 10 000 times, each time obtaining a new reduced Gray tree by randomly pruning 249 languages. On average, the ED of the reduced Gray tree and that of Gray *et al.*'s tree correlated to an R^2 of 0.78, and to 0.75 ± 0.14, 99% of the time. One may then apply, on the reduced Gray trees, Ethnologue data with the procedure mentioned in §2.1 to partially reconstruct Gray *et al.*'s tree, yielding 'reconstructed Gray trees', for which ED scores may be computed for the 101 languages present in the reduced tree. On average, the 101 computable ED scores of each reconstructed Gray tree and those of the corresponding 101 languages in Gray *et al.*'s tree correlated to an R^2 of 0.82, and to 0.78 ± 0.14, 99% of the time.

We then generalized the pruning procedure from 249 pruned languages to any number of pruned languages (figure 2, for a similar generalization of the pruning-and-reconstruction procedure; see electronic supplementary material figure S2). It appears that the R^2 between the ED scores of Gray *et al.*'s tree and that of the tree with pruned languages does not decrease linearly with the number of tips removed. The ED scores appear initially resilient. By contrast, if this sensitivity analysis of the pruning procedure is performed not on Gray *et al.*'s tree but on a random tree generated with the ape [44] package of the R statistical language, the R^2 decreases linearly and is on average equal to the percentage of tips left in the reduced tree.

These sensitivity analyses assume that the 1215 languages Gray *et al.* [40] included in their tree were a random subset of all Austronesian languages. Inclusion of languages in the phylogeny is influenced by the availability of data, and there may be bias in favour of well-documented languages, whereas those languages least well-documented might also be among the most endangered. On average, languages present in Gray *et al.*'s tree have an endangerment score of 1.21 and languages absent from the tree, a score of 1.49. Of languages present in Gray *et al.*'s tree, 33% of languages are threatened ($GE \geq 2$), whereas this is the case for 41% of excluded languages. Of languages present in Gray *et al.*'s tree, 13% are

endangered (GE ≥ 3) meaning that they are only spoken by the parent generation and older, while this is the case for 15% of excluded languages. Of languages present in Gray *et al.*'s tree, 6.9% are moribund (GE ≥ 3.5), meaning that they are only known to the grandparent generation and older. This figure is 7.4% in languages absent from Gray *et al.*'s tree. These figures suggest some bias: less endangered languages are slightly over-represented in the tree.

2.4. Calculating the evolutionarily distinct and globally endangered scores

As branch lengths in the language tree were not proportional to time (as is often the case with species trees), an absolute ED score is difficult to interpret. We therefore chose to use the relative ED (ED_R), computed by dividing all ED scores by the average ED score. By construction, ED_R scores have a mean of 1, and a language with an ED_R of 2 is twice as distinct as the average language.

The weightings of ED and GE in the original EDGE metric are arbitrary [45]. To give importance to both ED and endangerment (GE), we adapted the EDGE metric given in equation (1.1) by dividing the weight of the GE by 4, its maximum value:

$$\mathrm{EDGE} = \ln(1 + \mathrm{ED_R}) + \tfrac{1}{4} \cdot \mathrm{GE} \cdot \ln(2). \tag{2.1}$$

Had we stuck to the original definition of the EDGE, rankings would have been dominated by endangerment scores with little regard to distinctiveness (see Pearse *et al.* [46] for a similar approach). An extension of EDGE, HEDGE ('heightened' EDGE), also includes information on the P(extinction) of close relatives, such that an endangered language would be up-weighted if closely related languages were also endangered [47]. While this is a useful approach when setting global conservation priorities, the HEDGE metric is not appropriate in our case due to the large number of missing languages (i.e. we cannot be certain that we were not missing a closely related language that had a very different GE score to a language within our sample).

We ran all analyses with the R statistical language [48] with the following libraries: we used the ade4 [49], ape [44] and phytools [50] R packages to manipulate phylogenetic trees, picante [51] to compute ED, phangorn [52] and geiger [53] to idenfity ancestral nodes, and ggplot2 [54] to generate plots.

3. Results

ED_R scores are approximately log-normally distributed (full list in electronic supplementary material), ranging from that of Indonesian (0.15) to those of Kavalan (3.36) and Xârâcùù (Canala) (3.66). ED_R scores have a geometric mean of 0.836 and a median of 0.847. Of the 350 languages for which we can measure ED, 113 (32%) are threatened (GE ≥ 2), representing 34% of the total measurable Austronesian ED. EDGE scores are approximately normally distributed but slightly right-skewed (figure 3), and range from 0.15 (Indonesian) to 2.17 (Kavalan, table 2), with an average of 0.786 and a s.d. of 0.35.

ED and GE do not appear correlated ($R^2 = 0.008,\ p = 0.09$), although this could change with different ED and GE metrics. Nonetheless, our choice of GE is just one possible index of language endangerment, and alternative scales or transformations of language threat might reveal the relationship between ED and GE. As discussed above, we assessed the effect of missing languages on ED. We expect a coefficient of determination R^2 of ≈ 0.82 between their ED scores in the reconstructed full Austronesian tree (the one used in this analysis) and their ED scores in the hypothetical full Austronesian tree (within 0.78 ± 0.14, with 99% probability, details in the electronic supplementary material).

Neither ED_R nor EDGE are randomly distributed geographically—the Philippines are a striking example. Of the 350 languages studied here, 53 are of the Philippines (although Filipino itself, also an Austronesian language, is not included). In the Philippines, 48 languages (91%) have ED_R below average, and 51 languages (96%) are in vigorous use (GE ≤ 1). The Philippine language with the highest ED_R, Inabaknon, only ranks 83 out of 350, which is significantly lower than expected by chance ($p < 10^{-6}$; see the electronic supplementary material). Similarly, the Philippine language with the highest EDGE, Central Tagbanwa, ranks 90 out of 350, again significantly lower than expected by chance ($p < 10^{-7}$).

As for EDGE scores across other countries, all five French Polynesian languages except Tahitian are endangered, which makes French Polynesia the country with the highest average GE among those countries with more than two languages (avg. GE = 2.41, $n = 5$). Under the HEDGE framework, these languages would have been given even higher conservation priority. Formosan languages, spoken by indigenous peoples of Taiwan, have the second-highest average GE (1.91, $n = 14$). French Polynesian languages, however, have a much lower average ED_R (0.44) than Formosan languages (1.91); losing one average Formosan language would reduce the measured Austronesian ED more than losing all

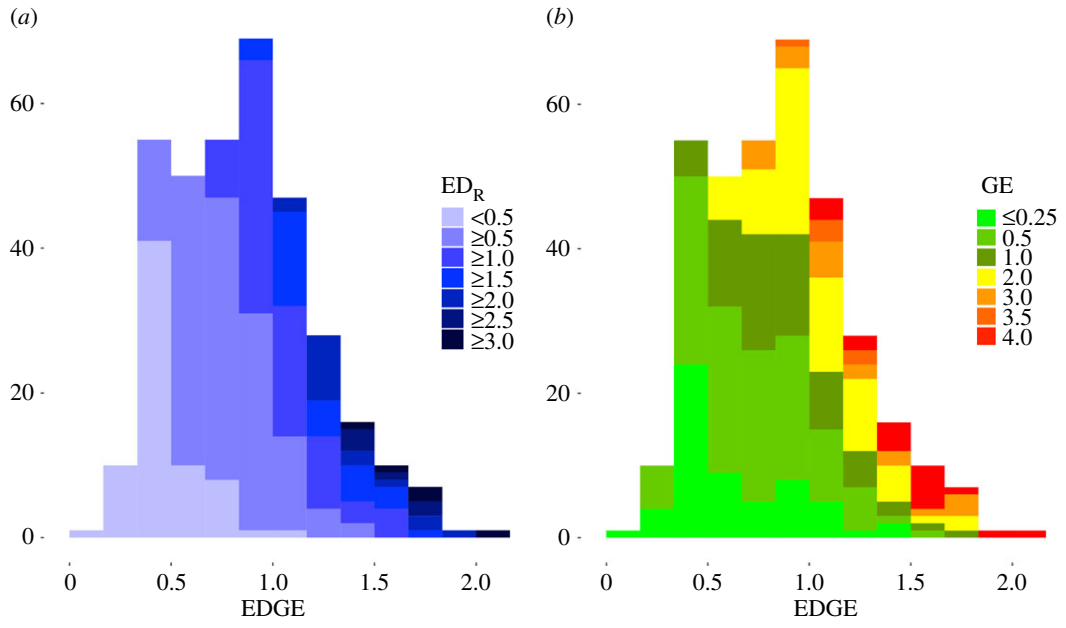

Figure 3. EDGE distribution of the 350 Austronesian languages shaded by their relative evolutionary distinctiveness (*a*) and endangerment level (*b*).

Table 2. Languages by EDGE score (full list in the electronic supplementary material .csv file).

	language	EDGE	ED_R	endangerment	GE
1	Kavalan	2.17	3.36	nearly extinct	4
2	Tanibili	1.86	2.21	nearly extinct	4
3	Waropen	1.774	2.50	shifting	3
4	Sengseng	1.765	3.13	threatened	2
5	Magori	1.75	1.88	nearly extinct	4
6	Xârâcùù	1.7133	3.66	vigorous	1
7	Irarutu	1.7130	2.92	threatened	2
⋮					
349	Tuvaluan	0.24	0.25	wider comm.	$\frac{1}{8}$
350	Indonesian	0.15	0.16	national	$\frac{1}{32}$

four endangered French Polynesian languages. Only New Caledonian languages have higher average ED_R (1.93) than Formosan languages, but New Caledonian languages are not as threatened (avg. GE = 1.25, $n = 7$). High ED and GE make Formosan languages the highest EDGE scoring on average (avg. EDGE = 1.38), followed by New Caledonia (1.20).

4. Discussion

In linguistics, as in conservation biology, limited resources in conjunction with rapid rates of extinction mean that efforts need to be optimized to maximize the preservation of diversity. Here, we suggest how efforts to preserve linguistic diversity could benefit from approaches used in conservation biology that include both distinctiveness and GE. Applying these types of metrics to languages requires only an endangerment score for each language, and a language tree whose branches reflect linguistic distinctiveness, data that are already available for many languages.

We illustrate the linguistic EDGE on a 350-language Austronesian family tree. Our results reveal striking disparities in the ED among languages, here reflecting a measure of lexical contribution to linguistic diversity. For example, the language with the highest ED, Xârâcùù, contributes 23 times

more than the language that contributes least. The six highest ranking EDGE languages (table 2) were Kavalan ($ED_R = 3.36,\ GE = 4$), Tanibili ($ED_R = 2.21,\ GE = 4$), Waropen ($ED_R = 2.50,\ GE = 3$), Sengseng ($ED_R = 3.13,\ GE = 2$), Magori ($ED_R = 1.88,\ GE = 4$) and Xârâcùù ($ED_R = 3.66,\ GE = 1$).

Kavalan is an exceptionally distinct yet nearly extinct language indigenous to Northeastern Taiwan. In 2000, it had 24 speakers [38] and an ethnic population of 1000 living mostly in Eastern Taiwan [55]. It is spoken in only one village, Sinshe, chiefly by elderly speakers. There have been recent efforts to revive it in schools, but without proper funding the village could not train language teachers [56]. Tanibili is one of three highly endangered languages of Utupua in Temotu Province, Solomon Islands, none of which have more than a few hundred speakers and are almost completely undocumented [57]. Waropen and Sengseng are languages of New Guinea spoken by a few thousand people. There are some word lists and other resources for Waropen [58], while there are word lists and a sketch grammar for Sengseng [59,60]. Waropen is no longer spoken by children, and only half of the children of Sengseng users speak it [38]. Magori is a nearly extinct language of Papua New Guinea that had 100 users in 2000 [38]. It is known, however, to have undergone large-scale lexical and structural borrowings from Magi, a Papuan language [61], and because unaccounted borrowings are ignored when computing ED, our estimate of ED_R might overestimate the distinctiveness of the language. Xârâcùù is a language of southern New Caledonia spoken by some 6000 people [62], and although not currently endangered, it is considered near threatened.

There are multiple complementary approaches for language preservation. Yet, for largely undocumented languages close to extinction, recording is an essential first step, for if there is no record of a language beyond its current speakers, there will be no reviving it once those speakers are lost. The exercise of ranking languages by both level of endangerment and distinctiveness is useful for identifying global priorities that maximize linguistic diversity. Such prioritization lists, however, can at best only help to inform preservation programs, and do not take into account other factors such as the quantity and quality of existing documentation, the practicality of working in particular regions, or the cultural, social and political contexts unique to each language [63,64]. This is an important observation, as in addition to identifying languages that might be prioritized, we show that neither ED_R nor EDGE are randomly distributed geographically. Both linguistic diversity and the drivers of language extinction risk are known to be geographically patterned [65,66], which may offer opportunities to prioritize groups of languages by proximity, leveraging the resources necessary for documentation to multiple languages at once. Similar challenges and opportunities arise in species conservation.

We should be cognisant that our measures of ED reflect only the information that is used to create the tree, and other metrics of ED are available. Any single language tree or metric is unlikely, therefore, to fully capture linguistic diversity. Aside from lexical change (new or modified words for the same things), linguistic change involves semantic change (existing words that shift meanings), phonetic change (change in pronunciation), phonological change (change in the frequency or number of phonemes) and syntactic change (change in syntax). Similarly, different ED metrics can give more or less weight to branches deeper in the tree, and thus capture different language features. These different types of language changes can occur together, either because a change in one aspect of a language provokes changes in the other, or because external factors induce changes on several of these aspects simultaneously. They may not, however, necessarily evolve in synchrony, as changes in one dimension can be independent of changes in another dimension. Our case study is based on lexical diversity, but could well be extended to encompass other dimensions of linguistic diversity [24], and account for uncertainties in the resulting trees. We present here a first attempt at merging threat and distinctiveness for language preservation.

As is the case for the species EDGE program, we anticipate and hope that our approach will be revised and improved through time as alternative phylogenies are constructed, methods are improved and as we refine our knowledge of the status of languages around the globe.

5. Conclusion

The EDGE scores presented here provide an illustration of the potential benefits in borrowing methods and theory from one field, here conservation biology, and applying them to another, here language preservation. In other examples, the similarity of language and species trees might find the flow of information reversed [67]. We considered over 350 languages, yet these represent only a subset of Austronesian languages. We show that such missingness has only a limited effect on the ED scores of included languages. Importantly, tree incompleteness never lowers EDGE scores, though it is possible

that relative rankings could change. In addition we present a novel method to evaluate robustness of ED measures estimated from incomplete trees, which has utility in biology and linguistics. Languages, however, cannot be assessed if we lack data for them. It is notable that while only 210 words are needed to include additional languages in the phylogeny we used, even these data are missing for the majority of Austronesian languages. Perhaps one of the most pressing priorities, therefore, is to gather the data required to build more inclusive language trees. Large, well-sampled species trees have transformed our understanding of macroevolution [68–71] and helped shape conservation priorities (see Mace *et al.* [72]). The construction of more comprehensive language trees is likely to benefit linguists, anthropologists and historians, as well as biocultural diversity for its own sake.

Data accessibility. The data supporting this article have been uploaded as part of the electronic supplementary material, and includes the ED_R, GE, EDGE, ISO 639-3 code and country for each language included in the analysis, as well as each of the phylogenetic trees. ED was calculated based on the Austronesian phylogeny of Gray *et al.* [40] built on data from the Austronesian Basic Vocabulary Database [41], freely available online (https://github.com/D-PLACE/dplace-data/tree/master/phylogenies/gray_et_al2009).

Authors' contributions. All authors designed the study. N.P. performed the analyses. N.P. & M.J.F. wrote the article, with input from T.J.D.

Competing interests. The authors declare no competing interests, financial or otherwise.

Funding. M.J.F. was supported by a Natural Sciences and Engineering Research Council of Canada Vanier CGS. T.J.D. was supported by an NSERC Discovery Grant.

Acknowledgements. We thank Simon Greenhill for kindly providing us with the Austronesian phylogeny, and Michael Dunn for providing us with an Indo-European language tree published in Dunn *et al.* (2011) [73], which was used in earlier drafts of this paper. We thank Kara Crabb for helpful discussion in framing the manuscript. We also thank Arne Mooers and Claire Bowern for their careful reading of our manuscript and helpful comments and suggestions as reviewers.

References

1. Wake DB, Vredenburg VT. 2008 Are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proc. Natl Acad. Sci. USA* **105**, 11 466–11 473. (doi:10.1073/pnas.0801921105)
2. Barnosky AD *et al*. 2011 Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51–57. (doi:10.1038/nature09678)
3. Meine C. 2010 Conservation biology: past and present. In *Conservation biology for all* (eds NS Sodhi, PR Ehrlich), ch. 1, pp. 7–26. Oxford, UK: Oxford University Press.
4. Vane-Wright RI, Humphries CJ, Williams PH. 1991 What to protect?—Systematics and the agony of choice. *Biol. Conserv.* **55**, 235–254. (doi:10.1016/0006-3207(91)90030-D)
5. Sutherland WJ. 2003 Parallel extinction risk and global distribution of languages and species. *Nature* **423**, 276–279. (doi:10.1038/nature01607)
6. Krauss M. 1992 The world's languages in crisis. *Language* **68**, 4–10. (doi:10.1353/lan.1992.0075)
7. Hale K. 1998 On endangered languages and the importance of linguistic diversity. In *Endangered languages* (eds LA Grenoble, LJ Whaley), pp. 192–216. Cambridge, UK: Cambridge University Press.
8. Bellwood P. 1995 Austronesian prehistory in Southeast Asia: homeland, expansion and transformation. In *The Austronesians: historical and comparative perspectives* (eds P Bellwood, JJ Fox, D Tryon), ch. 5, 2006 edn, pp. 103–118. Canberra, Australia: ANU E Press.
9. Gruhn R. 2006 Reconstructing prehistoric population movements: seeking congruence in genetics, linguistics, and archaeology. *Rev. Anthropol.* **35**, 345–372. (doi:10.1080/00938150600988182)
10. Holden CJ, Mace R. 2003 Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. *Proc. R. Soc. Lond.* **270**, 2425–33. (doi:10.1098/rspb.2003.2535)
11. Watts J, Sheehan O, Atkinson QD, Bulbulia J, Gray RD. 2016 Ritual human sacrifice promoted and sustained the evolution of stratified societies. *Nature* **532**, 228–231. (doi:10.1038/nature17159)
12. Hellier A, Newton AC, Ochoa-Gaona S. 1999 Use of indigenous knowledge for rapidly assessing trends in biodiversity: a case study from Chiapas, Mexico. *Biodivers. Conserv.* **8**, 869–889. (doi:10.1023/A:1008862005556)
13. Nabhan GP. 2000 Interspecific relationships affecting endangered species recognized by O'odham and Comcáac cultures. *Ecol. Appl.* **10**, 1288–1295. (doi:10.1890/1051-0761(2000)010[1288:IRAESR]2.0.CO;2)
14. Mapinduzi AL, Oba G, Weladji RB, Colman JE. 2003 Use of indigenous ecological knowledge of the Maasai pastoralists for assessing rangeland biodiversity in Tanzania. *Afr. J. Ecol.* **41**, 329–336. (doi:10.1111/j.1365-2028.2003.00479.x)
15. Sutherland WJ, Gardner TA, Haider JL, Dicks LV. 2014 How can local and traditional knowledge be effectively incorporated into international assessments?. *Oryx* **48**, 1–2. (doi:10.1017/S0030605313001543)
16. Kimmerer RW. 2002 Weaving traditional ecological knowledge into biological education: a call to action. *Bioscience* **52**, 432–438. (doi:10.1641/0006-3568(2002)052[0432:WTEKIB]2.0.CO;2)
17. Settee P. 2008 Native languages supporting indigenous knowledge. In *International Expert Group Meeting on Indigenous Languages, New York, NY, 8–10 January*, PFII/2008/EGM1/13. New York, NY: United Nations Department of Economic and Social Affairs.
18. Isaac NJB, Turvey ST, Collen B, Waterman C, Baillie JEM. 2007 Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PLoS ONE* **2**, e296. (doi:10.1371/journal.pone.0000296)
19. Redding DW, Mooers AO. 2015 Ranking mammal species for conservation and the loss of both phylogenetic and trait diversity. *PLoS ONE* **10**, 1–11. (doi:10.1371/journal.pone.0141435)
20. Redding DW, DeWolff CV, Mooers AO. 2010 Evolutionary distinctiveness, threat status, and ecological oddity in primates. *Conserv. Biol.* **24**, 1052–1058. (doi:10.1111/j.1523-1739.2010.01532.x)
21. Cadotte MW, Cardinale BJ, Oakley TH. 2008 Evolutionary history and the effect of biodiversity on plant productivity. *Proc. Natl Acad. Sci. USA* **105**, 17 012–17 017. (doi:10.1073/pnas.0805962105)
22. Cadotte MW, Cardinale BJ, Oakley TH. 2013 Experimental evidence that evolutionarily diverse assemblages result in higher productivity. *Proc. Natl Acad. Sci. USA* **110**, 8996–9000. (doi:10.1073/pnas.1301685110)
23. Srivastava DS, Cadotte MW, Macdonald AAM, Marushia RG, Mirotchnick N. 2012 Phylogenetic diversity and the functioning of ecosystems. *Ecol. Lett.* **15**, 637–648. (doi:10.1111/j.1461-0248.2012.01795.x)
24. Dunn M. 2015 Language phylogenies. In *Routledge handbook of historical linguistics* (eds C Bowern, B Evans), pp. 190–211. London, UK: Routledge. ISBN 9780415527897.
25. Dobzhansky T. 1970 *Genetics of the evolutionary process*, revised edn, 505p. New York, NY: Columbia University Press .
26. Safran RJ, Nosil P. 2012 Speciation: the origin of new species. *Nat. Educ. Knowledge* **3**, 17.

27. Schleicher A. 1863 *Die Darwinsche Theorie und die Sprachwissenschaft—offenes Sendschreiben an Herrn Dr. Ernst Haeckel*. Weimar, Germany: H. Boehlau.
28. Pavoine S, Ollier S, Dufour A-B. 2005 Is the originality of a species measurable? *Ecol. Lett.* **8**, 579–586. (doi:10.1111/j.1461-0248.2005.00752.x)
29. Redding DW, Mazel F, Mooers AØ. 2014 Measuring evolutionary isolation for conservation. *PLoS ONE* **9**, 1–15. (doi:10.1371/journal.pone.0113490)
30. May RM. 1990 Taxonomy as destiny. *Nature* **347**, 129–130. (doi:10.1038/347129a0)
31. Redding DW. 2003 Incorporating genetic distinctness and reserve occupancy into a conservation prioritisation approach. Masters thesis, University of East Anglia, Norwich, UK.
32. Redding DW, Mooers AØ. 2006 Incorporating evolutionary measures into conservation prioritization. *Conserv. Biol. ONE* **20**, 1670–1678. (doi:10.1111/j.1523-1739.2006.00555.x)
33. Bergsland K, Vogt H. 1962 On the validity of glottochronology. *Curr. Anthropol.* **3**, 115–153. (doi:10.2307/2739527)
34. Mooers A, Faith DP, Maddison WP. 2008 Converting endangered species categories to probabilities of extinction for phylogenetic conservation prioritization. *PLoS ONE* **3**, e3700. (doi:10.1371/journal.pone.0003700)
35. IUCN. 2013 The IUCN red list of threatened species. Version 2013.2.
36. Moseley C. 2010 Atlas of the world's languages in danger, 3rd edn. UNESCO. Online version: http://www.unesco.org/culture/en/endangeredlanguages/atlas.
37. Lewis MP, Simons GF. 2010 Assessing endangerment: expanding Fishman's GIDS. *Revue Roumaine de Linguistique* **55**, 103–120. (doi:10.1017/CBO9780511783364.003)
38. Ethnologue: languages of the world, seventeenth edition, 2014. http://www.ethnologue.com.
39. Bellwood P, Sanchez-Mazas A. 2005 Human migrations in continental East Asia and Taiwan: genetic, linguistic, and archaeological evidence. *Curr. Anthropol.* **46**, 480–484. (doi:10.1086/430018)
40. Gray RD, Drummond AJ, Greenhill SJ. 2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483. (doi:10.1126/science.1166858)
41. Greenhill SJ, Blust R, Gray RD. 2008 The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evol. Bioinform. Online* **4**, 271–283. (doi:10.4137/EBO.S893)
42. ISO 639-3 Registrar. 2014 *ISO 639-3*. Dallas, Texas: SIL International.
43. Volkmann L, Martyn I, Moulton V, Spillner A, Mooers AO. 2014 Prioritizing populations for conservation using phylogenetic networks. *PLoS ONE* **9**, e88945. (doi:10.1371/journal.pone.0088945)
44. Paradis E, Claude J, Strimmer K. 2004 Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)
45. Isaac NJB, Redding DW, Meredith HM, Safi K. 2012 Phylogenetically-informed priorities for amphibian conservation. *PLoS ONE* **7**, e43912. (doi:10.1371/journal.pone.0043912)
46. Pearse WD *et al.* 2015 Beyond the EDGE with EDAM: prioritising british plant species according to evolutionary distinctiveness, and accuracy and magnitude of decline. *PLoS ONE* **10**, e0126524. (doi:10.1371/journal.pone.0126524)
47. Steel M, Mimoto A, Mooers AO. 2007 Hedging one's bets: quantifying a taxon's expected contribution to future phylogenetic diversity. *Evol. Bioinf.* **3**, 237–244. (doi:10.1111/j.1461-0248.2005.00752.x)
48. R Development Core Team *et al*. 2011 R: A language and environment for statistical computing. R foundation for Statistical Computing.
49. Dray S, Dufour A-B. 2007 The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20. (doi:10.18637/jss.v022.i04)
50. Revell LJ. 2012 Phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223. (doi:10.1111/j.2041-210X.2011.00169.x)
51. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010 Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464. (doi:10.1093/bioinformatics/btq166)
52. Schliep KP. 2011 Phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593. (doi:10.1093/bioinformatics/btq706)
53. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008 GEIGER: investigating evolutionary radiations. *Bioinformatics* **24**, 129–131. (doi:10.1093/bioinformatics/btm538)
54. Wickham H. 2009 ggplot2: elegant graphics for data analysis. New York, NY: Springer. See http://had.co.nz/ggplot2/book.
55. Tsukida N, Tsuchida S. 2007 Indigenous languages of Formosa. In *The vanishing languages of the Pacific rim* (eds O Miyaoka, O Sakiyama, ME Krauss), pp. 285–300. New York, NY: Oxford University Press.
56. Hsieh F, Huang S. 2007 Documenting and revitalizing Kavalan. In *Documenting and revitalizing austronesian languages* (eds DV Rau, M Florey), pp. 93–110. Honolulu, HI: University of Hawai'i Press.
57. VaagOlsesen A. 2014 Documenting the Utupua languages. See https://elar.soas.ac.uk/Collection/MPI1032001.
58. Hammarström H, Forkel R, Haspelmath M, Bank S, Waropen 2017. See http://glottolog.org/resource/languoid/id/waro1242.
59. Chinnery EWP. 1928 Certain natives in South New Britain and Dampier straits. Territory of New Guinea Anthropological Report, 3. Melbourne, Australia: H. J. Green, Government Printer.
60. Chowning A. 1978 Comparative grammars of five New Britain languages. In *Proc. of the 2nd Int. Conf. on Austronesian Linguistics, Canberra, Australia* (eds S Wurm, L Carrington), pp. 1129–1157. Pacific Linguistics.
61. Foley WA. 1986 *The Papuan languages of New Guinea*. Cambridge Language Surveys. Cambridge, UK: Cambridge University Press.
62. Moyse-Faurie C. 2015 Valency classes in Xârâcùù (New Caledonia). In *Valency classes in the world's languages* (eds A Malchukov, B Comrie), pp. 1015–1068. Berlin, Germany: De Gruyter Mouton.
63. Crystal D. 2000 *Language death*. Cambridge, UK: Cambridge University Press.
64. Krauss M. 2007 Mass language extinction, and documentation: the race against time. In *The vanishing languages of the Pacific rim* (eds O Miyaoka, O Sakiyama, ME Krauss), pp. 19–39. New York, NY: Oxford University Press.
65. Amano T, Sandel B, Eager H, Bulteau E, Svenning J-C, Dalsgaard B, Rahbek C, Davies RG, Sutherland WJ. 2014 Global distribution and drivers of language extinction risk. *Proc. R. Soc. B* **281**, 20141574. (doi:10.1098/rspb.2014.1574)
66. Gavin MC *et al.* 2013 Toward a mechanistic understanding of linguistic diversity. *BioScience* **63**, 524–535. (doi:10.1525/bio.2013.63.7.6)
67. Atkinson QD, Gray RD. 2005 Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Syst. Biol.* **54**, 513–526. (doi:10.1080/10635150590950317)
68. Bininda-Emonds ORP *et al.* 2007 The delayed rise of present-day mammals. *Nature* **446**, 507–512. (doi:10.1038/nature05634)
69. Fritz SA, Rahbek C. 2012 Global patterns of amphibian phylogenetic diversity. *J. Biogeogr.* **39**, 1373–1382. (doi:10.1111/j.1365-2699.2012.02757.x)
70. Jetz W, Thomas G, Joy J, Hartmann K, Mooers A. 2012 The global diversity of birds in space and time. *Nature* **491**, 444–448. (doi:10.1038/nature11631)
71. Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V. 2004 Darwin's abominable mystery: insights from a supertree of the angiosperms. *Proc. Natl Acad. Sci. USA* **101**, 1904–1909. (doi:10.1073/pnas.0308127100)
72. Mace GM, Gittleman JL, Purvis A. 2003 Preserving the tree of life. *Science* **300**, 1707–1709. (doi:10.1126/science.1085510)
73. 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. (doi:10.1038/nature09923)