Journal of Language Evolution, 2026, **11**, lzag007
https://doi.org/10.1093/jole/lzag007
Published: 8 August 2026
Research article

OXFORD

Is phoneme inventory size influenced by speaker population size and isolation?

Samuel Ingloff-Richards[1], Xia Hua [2], Lindell Bromham [1]*

[1]Macroevolution and Macroecology, Research School of Biology, Australian National University, Canberra ACT 2601, Australia
[2]Mathematical Sciences Institute, Australian National University, Canberra, ACT 2601 Australia
*Corresponding author. Research School of Biology, Australian National University, Canberra ACT 2601, Australia. E-mail: lindell.bromham@anu.edu.au

Phoneme inventory size is related to speaker population size, but the cause of this relationship has remained a mystery. The correlation was first reported by Hay and Bauer (2007), who considered several possible causes of this relationship, including ascertainment bias, similarity between relatives, loss of phonemes in small, isolated populations, and maintenance of larger phoneme inventories in larger speaker populations. Here, we test these proposed causes of the relationship and show that the correlation between phoneme inventory size and speaker population size is largely driven by similarity between relatives and neighbours in both speaker population size and phoneme inventory size. However, we also show that phoneme inventory size is related to aspects of linguistic isolation—languages confined to islands or with few bordering languages tend to have lower phoneme inventory sizes than their relatives. We also explore the geographic bias of phoneme inventory databases and show that in some cases there is a positive bias towards regions with larger phoneme inventories. However, we find no evidence that less-studied languages tend to have lower reported phoneme inventory sizes. Our analyses demonstrate the importance of taking sources of non-independence and bias in typological data into account when conducting global analyses of linguistic diversity.

Keywords phoneme, language evolution, null model, Galton's problem, phylogenetic non-independence, language contact, phylogeny, spatial autocorrelation, island languages, geographic sampling bias

1. Introduction

> 'To our knowledge, no-one has ever reported a statistical correlation between the number of speakers of a language, and how many phonemes that language has. This, of course, is not surprising: why would anyone look for such an association in the first place? It is certainly not an association one would necessarily expect.'
>
> HAY, J. & BAUER, L. 2007. Phoneme inventory size and population size. *Language,* 83, 388–400.

A short paper published in *Language* in 2007 reported a curious result (Hay and Bauer 2007). The analysis had been prompted not by a theoretical question but by an observation. When collating typological features of a sample of languages, the authors noticed an apparent association between number of vowels and speaker population size that suggested further investigation might be fruitful. For a sample of 216 languages, they found significant correlations between population size and phoneme inventory size, not only for total phoneme count but also subsets of types of phonemes, including vowels, consonants, obstruents, and sonorants. Hay and Bauer (2007) noted that language families differ in both population size and phoneme inventory size, so they added family as an independent variable in an analysis and also repeated the analysis on family average values. However they analysed their data, the basic pattern remained: a positive association between population size and phoneme inventory size. The authors seemed frankly surprised at the generality and robustness of the relationship: 'In this short report we show that there is, indeed, an association between phoneme inventory and population size. We do not have well-developed theoretical arguments to offer about why this should be. However the correlation seemed intriguing enough that it was worth simply publishing the result, and leaving it up to readers to draw their own conclusions' (Hay and Bauer 2007).

They did discuss some possible explanations of the observed association between phoneme inventory size and population size based on previous proposals:

1. small, isolated populations lose phonemes because lower accuracy of phonological definition is tolerable;
2. small, isolated populations lose phonemes because close social connections permit reduction in phonological redundancy (Trudgill 2004);
3. large speaker populations can support larger phoneme inventories because child learners are exposed to a greater number of speakers, and increased exposure allows robust learning, permitting more categories to be distinguished.

Hay and Bauer (2007) also considered whether data bias or analytical artefacts could have produced their result. For example, languages with large numbers of speakers may be better studied so they may have more thoroughly recorded phoneme inventories, inflating phoneme inventory counts relative to smaller, less well-studied languages. They rejected this explanation on the grounds that the positive relationship between population size and phoneme inventory size occurred for family averages, which they argued would reduce the influence of well-studied families such as Indo-European. They also considered whether their result could be driven by family bias in both phoneme inventory size and population size—for example, Australian languages have small vowel inventories and small population sizes, while Indo-European languages have larger vowel inventories and larger populations. But they rejected these family-level trends as a cause of the relationship because the same correlation was observed for family average values of both phoneme inventory size and population size. They conclude that the 'positive correlation between how many phonemes a language has, and how many speakers it has….. is not an artefact of language family. We do not know what the underlying causes of this correlation are. But it is certainly intriguing, and we hope that this short discussion note will generate some discussion of the possible causes of such a relationship' (Hay and Bauer 2007). In this paper, we test several of the proposed causes of the relationship between phoneme inventory size and population size: artefacts of analysis (relatedness and proximity of languages), uneven research effort (geographic bias and level of documentation) and linguistic isolation (degree of contact between languages).

The relationship between phoneme inventory size and population size is worth revisiting for both theoretical and practical reasons. It represents an important example of a potentially ubiquitous effect shaping the nature of languages, yet is neither a strictly language-internal phenomenon (because it does not invoke interaction between language elements) nor entirely language-external (since it concerns a property of the speaker population rather than the social or physical environment it exists in). Phoneme inventory size also represents an interesting case study for the role of neutral (non-selective) factors on language change. While selective hypotheses have been offered for the use or disuse of particular sets of phonemes, most of the explanations invoked to explain patterns of phoneme inventory size as a whole involve neutral mechanisms of founder effect (loss of variants through subsampling from an ancestral language) or ongoing small population size (loss of variants through a process akin to genetic drift). Phoneme inventory size is also an informative case study for comparing the relative influence of 'vertical transmission' (inheritance from ancestor) with 'horizontal transmission' (acquisition from contact): both phenomena shape phoneme inventories, but the relative effect depends on the mechanism (e.g. whether small inventories are due to lack of diffusion from neighbours or loss of phonemes in small populations). If horizontal transfer is a major driver of phoneme inventory sizes, then we expect the contact history to be the best predictor of differences in phoneme inventories, but if vertical inheritance is more important then we expect relationships to have the strongest explanatory power. Finally, the relationship between phoneme inventory size and population size is an excellent exemplar of several practical considerations in comparative global studies of language variation. Most linguistic and cultural features are more similar between neighbours and relatives, and this fact alone is sufficient to generate significant correlations between language features, populations, and environments (Bromham et al. 2018; Bromham and Yaxley 2023; Claessens et al. 2023). Furthermore, when analysing global languages, many additional aspects of the data may impact on the analysis, such as the nature of the variable (binary, count or continuous data) and biased sampling (greater proportion of observations from some regions or families) (Hua and Bromham 2025).

There are three ways for a phoneme to enter a language: vertically (inherited from ancestral language), horizontally (borrowed from another language), or *de* *novo* origin (innovation within a language, including by splitting of existing phonemes). In addition, phonemes can be lost from a language through attrition, redundancy, or merger. Size of phoneme inventory might therefore be influenced by features that increase the chance of gain of phonemes (contact, innovation, splits) or that affect rates of loss of phonemes (attrition, merger, sampling error). While studies at the individual, population, or generational level can highlight processes of gain and loss of phonemes within speaker populations (e.g. Kerswill and Shockey 2007; Baranowski 2013), comparing phoneme inventories of different languages allows evaluation of general drivers of phoneme gain, retention, and loss over longer time periods (e.g. Atkinson 2011; Creanza et al. 2015; Dediu et al. 2017; Maddieson 2023). By comparing overall phoneme inventory size between languages with different population sizes and language ecologies, we can identify factors that are associated with phoneme gain, retention, and/or loss. Here, we consider inheritance from shared ancestors (vertical transmission of phonemes), contact with neighbouring languages (horizontal transmission of phonemes), and population size (potentially influencing origin, retention, and loss of phonemes).

Phoneme inventories undergo descent with modification. Languages inherit their phoneme inventories from their ancestors, but these inventories can then become modified by gain and loss of phonemes. Because they are derived from the same ancestral phoneme inventory, we would expect two related languages to be more similar in phoneme inventory size and composition than two randomly selected languages. Phoneme inventory sizes cluster both on the phylogeny and in space; for example, many of the smallest phoneme inventories are from Oceanic languages from the Austronesian family, while many of the largest are from languages from southern Africa (Fig. 1). Population size is also non-randomly distributed in both space and phylogeny (Bromham 2024a). This non-random distribution across regions and families could lead to correlation between population size and phoneme inventory size even in the absence of a causal connection between population size and the gain, retention, or loss of phonemes (Roberts and Winters 2013; Urban and Moran 2021; Bromham and Yaxley 2023). Furthermore, phoneme inventories can be influenced by borrowing, so contact and proximity with other languages may also influence phoneme inventory size and content (Van Coetsem 1988; Creanza et al. 2015; Moran et al. 2021; Blevins and Egurtzegi 2023).

Our focus in this paper is on testing possible causes of Hay and Bauer's (2007) observed relationship between population size and phoneme inventory size. Because the theories cited by Hay and Bauer focus on the effect of small, isolated populations on reduction in phoneme inventories, we use a number of measures of language isolation in addition to population size and area. Isolation and contact are notoriously tricky to quantify, so our approach here is to use a range of variables that might reflect the ease of contact between speakers of different languages, such as how many other language distributions are in contact, and whether a language occurs on an island or in a mountainous area or area of rough terrain. But in order to examine the possible influence of these population factors on phoneme inventories, we must take language distribution and relationships into account.

Because languages are similar to their relatives and neighbours in many aspects, we expect to find correlations between many aspects of language, population, and environment. Associations due to phylogenetic non-independence (similarity due to descent) and spatial autocorrelation (similarity due to proximity and shared environments) can be sufficient to drive a strong correlation between aspects of language, population, and environment (Bromham and Yaxley 2023; Claessens et al. 2023; Bromham 2024a, 2024b). For example, a correlation between tonality and humidity has been reported (Everett et al. 2016), but this could be due to a greater concentration of languages in tropical areas, or non-random patterns of tones among areas and language groups (Collins 2017). When language data is analysed with information on location and relationships, there is no significant association between tone and humidity beyond that accounted for by similarity between neighbours and relatives (Bromham and Yaxley 2023; Hartmann et al. 2024). Similarly, an association between ejectives and altitude has been proposed (Everett 2013), but has been questioned on the grounds that mountain communities might be similar due to language relationships and distribution, and a shared history that resulted in contact-induced change in phoneme inventories between languages (Urban and Moran 2021).

Correlation due to descent from a common ancestor is referred to as phylogenetic non-independence, also known as 'Galton's problem' (Bromham 2024b). Moran et al (2012) addressed the problem of non-independence of phoneme inventories due to genealogical relationships by adding language family as a factor in an analysis. Using a hierarchical mixed model with family as a factor, they found that language families tend to have a greater influence on phoneme inventory size than population size factors. They concluded that 'although overall the trend across languages is modestly positive, for many families we can expect the relationship between the phoneme inventory size and population to be negative or zero' (Moran et al. 2012). Similarly, Donohue and

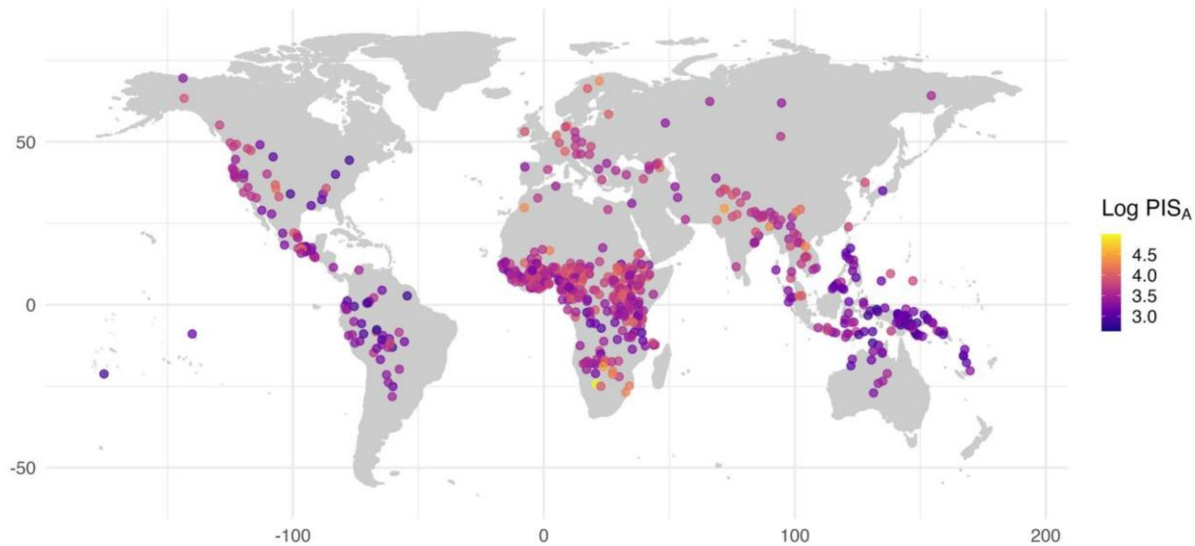

Figure 1 Map of phoneme inventory size (PIS$_A$) for 766 languages (see Methods for details). Phoneme inventory sizes have been log transformed and colour scaled (legend on the right). Phoneme inventory sizes and spatial coordinates are from Anderson et al. (2023).

Nichols (2011) examined the relationship between phoneme inventory size and speaker population size and found that the relationship varied between families, with some showing a positive association, but others no association or a negative trend (Donohue and Nichols 2011). These studies indicate that the relationship between phoneme inventory size and population size is moderated by language family and region, so to describe general relationships and test proposed mechanisms, we need to take language relationships and distribution into account.

However, adding language or region as a factor in an analysis does not fully account for phylogenetic non-independence (Hua and Bromham 2025). For example, the Austronesian language family contains over 1,200 languages (Hammarström et al. 2019), 89 of which are represented in PHOIBLE, with phoneme inventory counts ranging from 17 to 55 (Moran and McCloy 2019). Within this family, there are clear areal and genetic patterns, for example, a west-to-east decline in inventory size for the Oceanic languages (Blust 2013). A number of processes have been proposed to shape phoneme inventory sizes within this family, such as a gradual 'erosion' as populations expand eastward across Oceania, due to founder effects, isolation and small population size, or increase in phoneme inventories through contact and borrowing at the western end of the range (Blust 1990; Trudgill 2004; Naranjo et al. 2025). Whatever the cause of these west-to-east patterns in phoneme inventory size, we expect two closely related and co-located Austronesian languages to be more similar in their phoneme inventory count than two languages chosen at random from the family—for example, within the Austronesian language family, eastern Polynesian languages all tend to have relatively small phoneme inventories compared to other languages in the family. Therefore, we need to take relationships between all languages into account when analysing data on phoneme inventory size, not simply their family or macro-area (Naranjo and Mertner 2023).

In this study, we assess the generality of the association between phoneme inventory size and speaker population size. We test several possible causes of the previously reported correlation between phoneme inventory size and population size (Hay and Bauer 2007): population size and isolation (phonemes inventories undergo reduction in small isolated populations), ascertainment bias (well-studied languages have larger phoneme inventories), phylogenetic non-independence (relatives are more similar in both phoneme inventory and population size), and spatial autocorrelation (neighbours are more similar in both phoneme inventory and population size).

2. Methods

2.1 Is phoneme inventory size related to population size?

Several phoneme inventory databases are available, differing in their coverage and their reported phoneme inventory sizes, and some databases contain multiple values for the same language. Our selection of phoneme inventory data was based on the practical guideline of maximizing the number of languages for which we could obtain comparable data for phoneme inventory size, speaker population size, and other relevant variables (Table 1). For the first analysis, we use the standardized language phoneme count data provided by Anderson et al. (2023), who matched phonemes according to the Cross Linguistic Transcription System so that all datasets shared the same phoneme representation convention. Only segmental phonemes were included (vowels and consonants); other suprasegmental features such as tone were not (Anderson et al. 2023). We match phoneme inventory sizes from Anderson et al. (2023) to the language database of Bromham et al. (2022) by mapping Glottocodes to ISO-693 codes (using PHOIBLE values as they have the greatest coverage). In total, after the removal of duplicates and alignment with phylospatial and spatial distance matrices, there are 766 languages with the relevant information (see Supplementary Information for details). We refer to these Phoneme Inventory Size data derived from Anderson et al. (2023) as PIS_A (see Table 1).

To represent the speaker population size, we use estimated number of first language (L1) speakers for each spoken language (Bromham et al. 2022). Clearly, these are imperfect measures of speaker population size because they do not include second language speakers and are dependent on census size from a range of sources. Furthermore, speaker population size and distribution are dynamic so any values can represent only a snapshot at a point in time (Bromham 2023). Although historic population sizes would be a more relevant reflection of the language ecology in which each language developed its phoneme inventory, available population size data is limited to reflecting contemporary speaker population sizes (Bromham 2023), for example, some languages in our dataset have a reported L1 speaker population of zero (Fig. 2). However, there is currently no alternative source of historic L1 speaker population size available for a sufficient sample of languages (see discussion in Bromham 2023). In any case, previous studies have explored the association between contemporary L1 population size and phoneme inventory size (Hay and Bauer 2007; Moran et al. 2012), so these are the appropriate data to explore that observed relationship. We do not use estimates of number of second language (L2) speakers because they are available for relatively few languages (e.g. 120 in Shcherbakova et al. 2023).

To ask whether phoneme inventory size is related to L1 population size, we first do a standard regression. We then use a phylospatial analysis that takes into account the location of each language and its relationship to all other languages (Hua et al. 2019). This method assumes spatial error model (SEM) and uses generalized least squares (GLS) method to fit regression models, using two matrices of covariation based on the relatedness of languages (based on the taxonomy in Glottolog (Hammarström et al. 2019)) and spatial proximity (based on the centroids of the language distributions: Bromham et al. 2022). We treat phoneme inventory size as a continuous variable because its distribution is wide and smooth, ranging between 14 and 147 with a median value of 35 (Fig. 3). In addition, we repeat all analysis described in this section for an alternative database of phoneme inventory sizes (PIS_C) from Creanza et al. (2015). Phoneme inventory size values differ between these two databases but are strongly correlated (Supplementary Figure S1).

2.2 Is phoneme inventory size shaped by area and isolation?

Relative degree of linguistic isolation and contact is challenging to represent in a format that is comparable across all languages. We use several variables to represent different aspects of contact based on information that is available for the majority of the world's spoken languages (Table 1). Language area has previously been used to represent linguistic isolation, on the assumption that speakers from a more widespread language are more likely to come into contact with L1 speakers of other languages, or adult L2 learners within the language area (Lupyan and Dale 2010). To represent language area, we include the area of each language from Bromham et al (2022), which represents autochthonous language distribution and does not include emigrant populations. We also include the number of autochthonous languages whose distributions overlap or touch the language area (given the variable name 'bordering') and the L1 population size (see Table 1).

We include two variables representing the potential for physical isolation: altitudinal range ('altitude') and island endemism ('island': see Table 1). These variables are chosen to represent ease of movement over the landscape; however, we acknowledge that physical barriers of water or steep terrain do not necessarily represent barriers to contact between language groups. Because the aim is to represent the degree to which speakers of one language can be in regular contact with other L1 language speakers in the surrounding area, altitudinal range is calculated for a 10,000 km^2 'neighbourhood' defined as a circle centred on the centroid of the language's polygon. Altitudinal range is defined as the difference

Table 1 Variables included in all analysis, listed by the numbered sections of the Methods section.

Analysis	Information	Variable	Description	Transform	Source
1	Phoneme Inventory Size	PIS_A	Number of phonemes (Anderson database)	log (PIS_A)	Anderson et al. (2023)
2, 3		PIS_C	Number of phonemes (Creanza database)	log (PIS_C)	Creanza et al. (2015)
1, 2	Population	L1_Pop	Estimated number of L1 speakers	log (L1_pop+0.5)	Bromham et al. (2022)
2		Area	Language distribution area (km^2)	log (area)	Bromham et al. (2022)
2	Contact	Bordering	Number of languages in contact	log (bordering+0.5)	Bromham et al. (2022)
2		Altitude	Average altitudinal range of 10,000 km^2 neighbourhood	log (altitude+0.5)	Bromham et al. (2022)
2		Island	Language with >90 per cent of distribution on landmasses <11,000 km^2		Bromham et al. (2024)
2		Mainland	Coast-to-coast distance to nearest landmass ≥25,000 km^2	log (mainland+0.5)	Bromham et al. (2024)
2		Continent	Coast-to-coast distance to nearest landmass ≥2.5 million km^2	log (continent+0.5)	Bromham et al. (2024)
3	Ascertainment bias	Documentation	Availability of textual linguistic resources		Bromham et al. (2022)
3		Region	Which of 12 global regions language is in		Bromham et al. (2022)

Variables with zeros have 0.5 added before the log transformation.

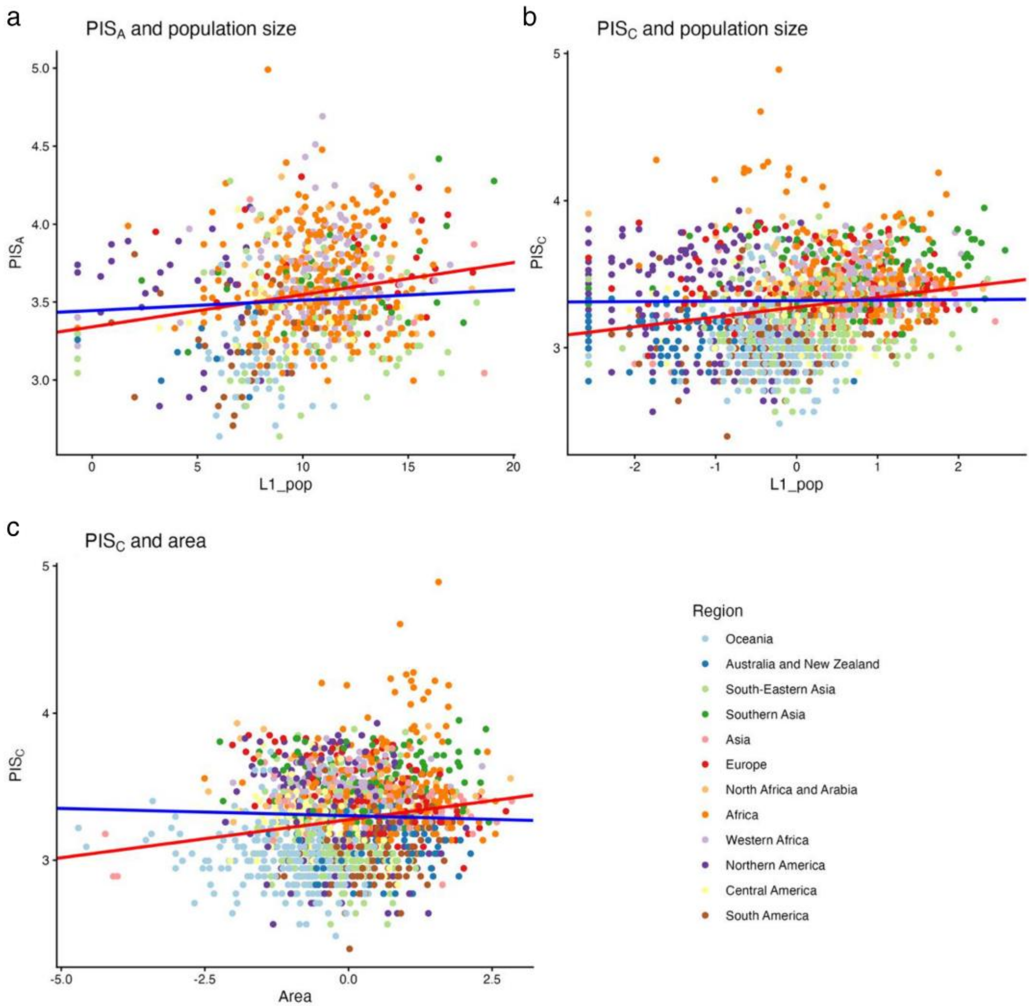

Figure 2 Phoneme inventory size vs L1 population size and area. All variables have been log-scaled. Blue lines represent the outcome of phylospatial regression; red lines represent the outcome of uncorrected regression. See Table 1 for variable descriptions and sources of data.

between the maximum and minimum altitude within each 5 km^2 grid cell, averaged across all grid cells within the language neighbourhood (Bromham et al. 2022). Island endemism, defined as a language where >90 per cent of the language area is on landmasses <11,000 km^2, is taken from Bromham et al. (2024). Because we are concerned with the historical development of phoneme inventories, we do not include measures that might reflect recent contact, such as roads or change in the built environment (Bromham et al. 2022).

We use a phylospatial analysis method described in Hua et al. (2019) to evaluate whether any of these factors—L1 speaker population size, area, bordering, altitude, and island—are significantly associated with phoneme inventory size (for Supplementary Information for more detail: for code availability see below). The Anderson et al. (2023) phoneme inventory size database (PIS$_A$) has a relatively poor representation of island endemic languages, with only 42 island endemic languages identified by matching languages by Glottocode and ISO-code to the database of Bromham et al (2022) (see Supplementary Information for details of matching). Instead, for analysing the relationship between isolation and phoneme inventory size, we use the phoneme inventory sizes from the database of Creanza et al. (2015), which we refer to as PIS$_C$ (see Table 1), from which 1,710 languages have sufficient information to be included in this analysis. We generate all combinations of these variables to create 32 models (including intercept-only model), then find GLS maximum likelihood solution for each model. Models are ranked by Bayesian

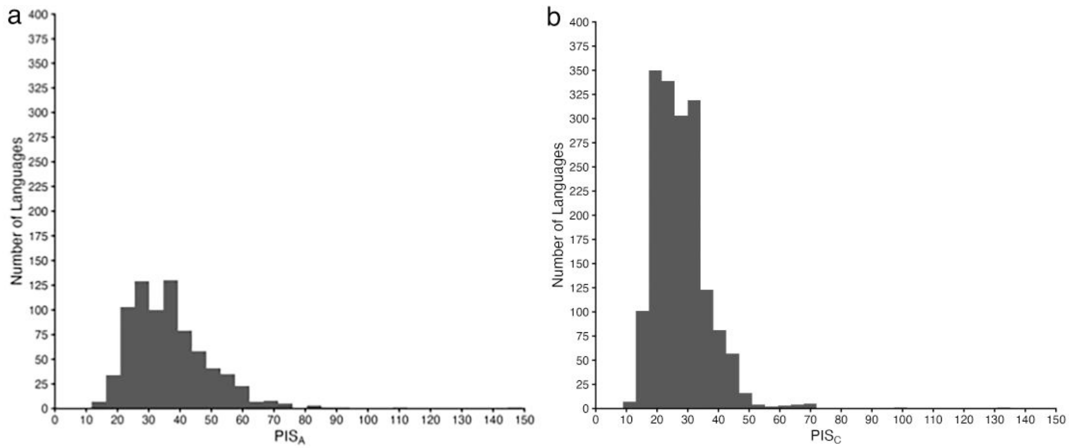

Figure 3 Histogram of phoneme inventory size for two different databases: a) Anderson et al 2023 (PIS_A) for 766 languages, and Creanza et al. (2015) (PIS_C) for 1,716 languages: see Methods for details of data curation.

Information Criterion (BIC), and only those models with a delta BIC (ΔBIC) less than or equal to 10 are considered 'good' models for phoneme inventory size. This approach follows the widely used strategy of using Jeffrey's scale to select models using a 'rule of thumb' that differences in BIC greater than 10 suggest strong evidence for support of one model over another (Jeffreys 1961). Coefficients used in selected models are tested for significance at the level 0.05.

In addition, we ask whether isolation in terms of distance from mainland or continental languages is associated with reduced phoneme inventory size by reanalysing data from Bromham et al. (2024) (see Supplementary Information for details). We include phoneme inventory size (PIS_C) for 1,716 global languages including 153 island endemics, and we ask whether phoneme inventory size is predicted by L1 population size, area, island endemism, distance to mainland, and distance to continent. We test all combinations of variables, giving 32 alternative models, then rank models by BIC.

2.3 Does ascertainment bias influence global analysis of phoneme inventory size?

Sampling bias has been put forward as a possible explanation for the relationship between phoneme inventory size and population size, on the assumption that larger languages tend to be better studied, and well-studied families may not be a globally representative sample of phoneme inventory sizes (Hay and Bauer 2007). We investigate both geographic coverage and study effort as potential sources of sample bias in phoneme inventory sizes (Supplementary Figure S2).

We test whether our dataset of phoneme inventory sizes is biased towards particular regions compared to a random sample of the world's languages. To generate a null model of the expected number of languages per area, we randomly sampled N languages from the global database where N is the number of languages for which we have phoneme inventory size (for the two different datasets, PIS_A and PIS_C), and record how many of the randomly sampled languages are in each of 12 global regions, using the region designations from Bromham et al. (2022) which is based on the global regions from the Natural Earth dataset (https://www.naturalearthdata.com/). We repeat the random sample 1,000 times. This generates the number of languages that we would expect from each region from a geographically unbiased sampling process of an equivalent number of languages as our available data. We can then compare the distribution across regions in our observed sample—the languages for which we have comparable phoneme inventory size data—to the number expected if sampling was unbiased with respect to region (Bromham et al. 2021). Note that this is a test of geographic distribution of languages included in our analysis (i.e. for which we can derive appropriate data for our analysis) and not a general test of the geographic distribution of languages in the source databases (Creanza et al. 2015; Anderson et al. 2023).

To generate this null model, we use the database of Bromham et al (2022) containing 6,511 languages (~90 per cent of the world's L1 spoken languages), which has been curated to only include languages for which comparable data is available on population size and distribution. While Glottolog contains a larger set of languages, it includes languages for which population size data is not currently available. In addition, Glottolog includes a wide range of language types including cants (e.g. Polari, pld), constructed languages (like Esperanto, eco), market languages and pidgins (like Fanagalo, fng), and speech registers (like Shelta, sth), as well as signed languages and unattested languages. Because none of these categories of language are included in the database that we use to test the relationship between population size and phoneme inventory size, we could not have included them in this analysis, so they are not appropriate for the null model of sampling bias. For example, if relatively more non-L1 languages such as cants and speech registers are recorded within the Indo-European language family, then it will give the impression that our empirical sample of phoneme inventories (which didn't include cants and speech registers) is under-sampling Indo-European languages.

For documentation, we use an ordered categorical variable based on information on language documentation available in Glottolog (v4.2.1): 0 = little or none; 1 = basic; 2 = detailed (Bromham et al. 2022). While this variable does not fully reflect linguistic research attention, it is the only globally comparative metric that would allow us to ask if language that have been more extensively studied have relatively higher phoneme counts. We ask if languages with greater levels of documentation tend to have higher phoneme inventory sizes using PIS_C for 1,716 languages, taking relationships and location of languages into account using method of Hua et al (2019)—see available code.

3. Results

3.1 Is phoneme inventory size related to population size?

Both L1 population size (L1_pop) and phoneme inventory size (PIS_A) were log transformed as both have a right-skewed distribution (Fig. 3; Bromham et al 2024). PIS_A is significantly positively correlated with L1 population size (Fig. 2a: $N = 766$, BIC = 477.3, coefficient = 0.021, SE = 0.003, $P < 0.001$). However, this relationship is weak and not significant when relationships and locations of languages are taken into account using the GLS model (BIC = 232.8, coefficient = 0.007, SE = 0.004, $P = 0.074$), suggesting that the association between phoneme inventory sizes and population size is at least partly driven by similarity between relatives and neighbours. Comparing Nagelkerke/Cragg–Uhler pseudo-R^2 for models incorporating spatial or phylogenetic covariance or both suggests that the geographic proximity of languages provides the strongest autocorrelation in these data (for $PIS_A \sim$ L1_pop: spatial and phylogenetic $R^2 = 0.6433$, spatial only $R^2 = 0.6253$, phylogenetic only $R^2 = 0.3663$, none $R^2 = 0.0893$).

Similar results were obtained for an alternative database of phoneme inventory size (PIS_C, Creanza et al 2015: Fig. 2b). There was a weak but significant association with ordinary least squares (N = 1716, BIC = 563.9, coefficient = 0.017, $P < 0.001$), but no association was found when taking phylogenetic and spatial relationships into account (BIC = −986.9, coefficient = 0.001, $P = 0.582$). The results are also consistent when repeated for the same set of languages for both PIS_A and PIS_C. There are 306 languages that have phoneme inventory sizes in both PIS_A and PIS_C, and phylospatial analysis reveals no relationship between population size and phoneme inventory size for either (PIS_A: BIC = 147.2, coefficient = 0.007, SE = 0.006, $P = 0.272$; PIS_C: BIC = −91.6, coefficient = −0.011, SE = 0.016, $P = 0.510$).

3.2 Is phoneme inventory size shaped by area and isolation?

We tested 32 models with all combinations of five variables (L1_pop, area, bordering, altitude, island) as predictors of phoneme inventory size using PIS_C (as there are insufficient island languages in the PIS_A dataset). While the null model has the lowest BIC, the next-best models are all <10 difference in BIC, so we do not have strong evidence to favour any of these models in terms of explanatory power. Models with ΔBIC less than or equal to 10 suggest that, after accounting for autocorrelation due to relationships and proximity, island endemism (isolation) and language area are significant negative predictors of phoneme inventory size, and bordering language richness (isolation) is a significant positive predictor of phoneme inventory size (Table 2). In fact, models with predictors reflecting both area and isolation are considered the best models according to AIC (i.e. models 7 and 8 in Table 2). Altitudinal range and population size were also included in the good models but were not significant predictors. These results suggest that languages that are more isolated from other languages (on islands or with fewer immediate neighbours) tend to have smaller phoneme inventory sizes, even after accounting for similarities between neighbours and relatives, with more limited support for reduced phoneme inventory sizes in smaller populations or those occurring in steeper terrain. Our reanalysis of the data from Bromham et al (2024) confirms that island endemic languages tend to have smaller phoneme inventory sizes, and that phoneme inventory sizes tend to decrease with distance from mainland and continent (Supplementary Table S1).

Language area is positively associated with phoneme inventory size for an uncorrected correlation (PISc, Fig. 2c). However, after accounting for covariance due to relationships and proximity between languages, area is a consistent negative predictor of phoneme inventory size. The change in the sign of the relationship between the uncorrected regression and the phylospatial regression suggests that there are additional trends in the data that are not captured at the global level.

3.3 Does ascertainment bias influence global analysis of phoneme inventory size?

By comparing the available phoneme inventory size data to a null model generated by random sampling from the global language database, we show that the available data is not an unbiased sample of the world's languages, but that the degree and direction of bias depend on the dataset. For the data derived from PHOIBLE (PIS_A, Anderson et al 2023), languages from Africa are relatively over-represented in the phoneme inventory size database compared to a random sample of languages from the global database, and languages from Oceania and Asia (including South Eastern Asia, Southern Asia, and Asia) are relatively underrepresented (Fig. 4a). For the PIS_C data derived from Creanza et al (2015), languages from Europe and North America are relatively oversampled, and languages from South-Eastern Asia and Africa are relatively undersampled (Supplementary Figure S3a). These results are specific to the data included in this study (i.e. only including languages that can be matched to relevant variables for our analyses) and may not reflect the source databases as a whole.

Regions differ in the distribution of phoneme inventory sizes (Figs 4a and 4b). The largest reported phoneme inventory sizes are from Africa, and languages from Oceania, Australasia, and South America tend to have smaller phoneme inventory sizes than the global median in both databases. For PIS_A, there is a significant correlation between the sample bias and phoneme inventory size, such that the regions that have more languages in the database than would be expected from a random sample of the world's languages also tend to have larger phoneme inventory sizes (Supplementary Figure S4). This suggests that available data is biased toward languages with richer phoneme inventories. However, there is no such trend for the PIS_C measure of phoneme inventory size (Supplementary Figure S4). Furthermore, there is no evidence in these data for a link between study effort (represented by documentation level) and phoneme inventory size ($N = 1{,}716$, BIC = −986.6, coefficient = −0.0018, $P = 0.821$: see Supplementary Figure S3). This makes an indirect causal connection between population size, study effort, and phoneme inventory size less plausible, though it would be possible through a rather indirect path (population size —> study effort —> database bias <— regional bias —> phoneme inventory size: Supplementary Figure S2).

4. Discussion

We confirm Hay and Bauer's (2007) observation of a correlation between population size and size of phoneme inventories. We analyse values drawn from two different phoneme inventory size databases (Creanza et al. 2015; Anderson et al. 2023) creating datasets which are, respectively, four times and eight times the size of Hay and Bauer's original dataset. Hay and Bauer offered several potential mechanistic explanations for their observation, including loss of phonemes in small, isolated populations (due to tolerance of low accuracy or loss of phonological redundancy) and maintenance of large phoneme inventories with larger speaker populations (due to greater exposure to more speakers and more robust learning). They also considered several kinds of ascertainment bias, for example, if languages with lots of speakers are better studied, leading to larger phoneme inventory counts, or due to similarities among languages within families in both phoneme inventories and speaker population size. Hay and Bauer considered these potential explanations qualitatively. In this paper, we have offered a quantitative examination of these potential causes.

We find that population size itself does not seem to be a convincing causal factor for phoneme inventory size because there is no significant association between population size and phoneme inventory size above and beyond that expected from phylogenetic and spatial autocorrelation. This result suggests that the association between L1 speaker population size and phoneme inventory size is largely an indirect result of neighbours and relatives being more similar to each other (Donohue and Nichols 2011; Bromham and Yaxley 2023; Bromham 2024a, 2024b). Therefore, our results do not support mechanisms directly related to the number of L1 speakers. If the number of speakers is not a predictor of phoneme inventory size, this suggests that having more speakers does not drive greater phoneme size by virtue of greater rates of innovation per generation, or better retention of phonemes due to more reliable transmission. These

Table 2 Effect of population size, area, and isolation on phoneme inventory size: results of multivariate phylospatial analysis ($N = 1{,}710$: see Methods, Section 2).

Model	AIC	BIC	ΔBIC	LogL	Island	Area	Altitude	Bordering	L1_pop
1	−1,000	−989.4	0	502.1					
2	−1,002	−985.3	4.015	503.8	−0.045 (0.064)				
3	−1,001	−984.8	4.532	503.6		−0.009 (0.088)			
4	−999	−982.7	6.705	502.5				0.004 (0.390)	
5	−999	−982.3	7.050	502.3					0.004 (0.530)
6	−999	−982.3	7.086	502.3			−0.003 (0.550)		
7	−1,004	−982.0	7.331	505.9		**−0.017 (0.009)**		**0.013 (0.031)**	
8	−1,004	−981.8	7.527	505.8	**−0.052 (0.035)**	**−0.0106 (0.048)**			
9	−1,002	−979.8	9.568	504.8		**−0.013 (0.027)**			0.010 (0.121)

Models with a delta BIC (ΔBIC) less than 10 are reported here, with the log likelihood value (logL). Coefficients are given with *P*-values below in parentheses ($P < 0.05$ are given in bold). Empty cells mean that the predictor was not included in the model, such that Model 1 represents only the intercept being included. This analysis uses PIC_C values for phoneme inventory size: for definitions and sources of variables see Table 1. The predictors Area, Altitude, Bordering, and L1_pop were standardized after being log-transformed (Island was not standardized).

results also do not support mechanisms associated with small numbers of speakers, such as loss due to stochastic sampling effects (equivalent to genetic drift), or erosion of phonemes due to close social relationships among speakers (see Hay and Bauer 2007).

The results of our analyses are consistent with a role for isolation in shaping phoneme inventories. Isolation is challenging to quantify, and the variables included in this study are indirect and imperfect measures of the likely degree of contact with speakers of other languages. Nonetheless, languages with fewer neighbouring languages in direct contact, or languages confined to islands, tend to have lower phoneme inventory sizes, above and beyond similarities due to relationships and proximity of languages. One potential explanation is that contact between neighbouring languages facilitates horizontal transfer of phonemes that allows building or maintaining diverse phoneme inventories. For example, while most Bantu languages do not use clicks, the few that do are spoken in the vicinity of languages from other families that have rich inventories of click phonemes (Güldemann and Stoneking 2008). While this transfer of clicks may be initiated through lexical borrowing, it is then possible for clicks to be incorporated into more words, becoming part of the phoneme inventory, which may then also spread among neighbouring languages (Bostoen and Sands 2009). Similarly, it has been suggested that phoneme inventory sizes in Polynesian languages are influenced by contact with non-Polynesian languages at the western end of the distribution, resulting in relatively higher phoneme inventory sizes than languages from the more remote eastern islands (Naranjo et al. 2025). Areal relationships such as these might contribute to the within-region trends in phoneme inventory size with population size or area (Fig. 2). Alternatively, as Hay and Bauer suggested, the negative relationship between isolation and phoneme inventory size could be explained by erosion of phonemes in languages whose speakers rarely talk to strangers, through a process of loss of redundancy or tolerance of lower accuracy. While the results of this global analysis do not allow interrogation of the mechanism of increased phoneme inventories with a greater degree of contact between languages, they do point to general patterns that may be fruitfully investigated at the language level, where origin and loss of phonemes can be examined in more detail.

Notably, if languages can directly influence each other's phoneme inventories by contact, this interaction between languages may indicate that a SEM is not the best model for these data because it models distance between observations as a nuisance parameter, filtering out the effects of proximity on the covariation between languages. But if neighbours influence each other, for

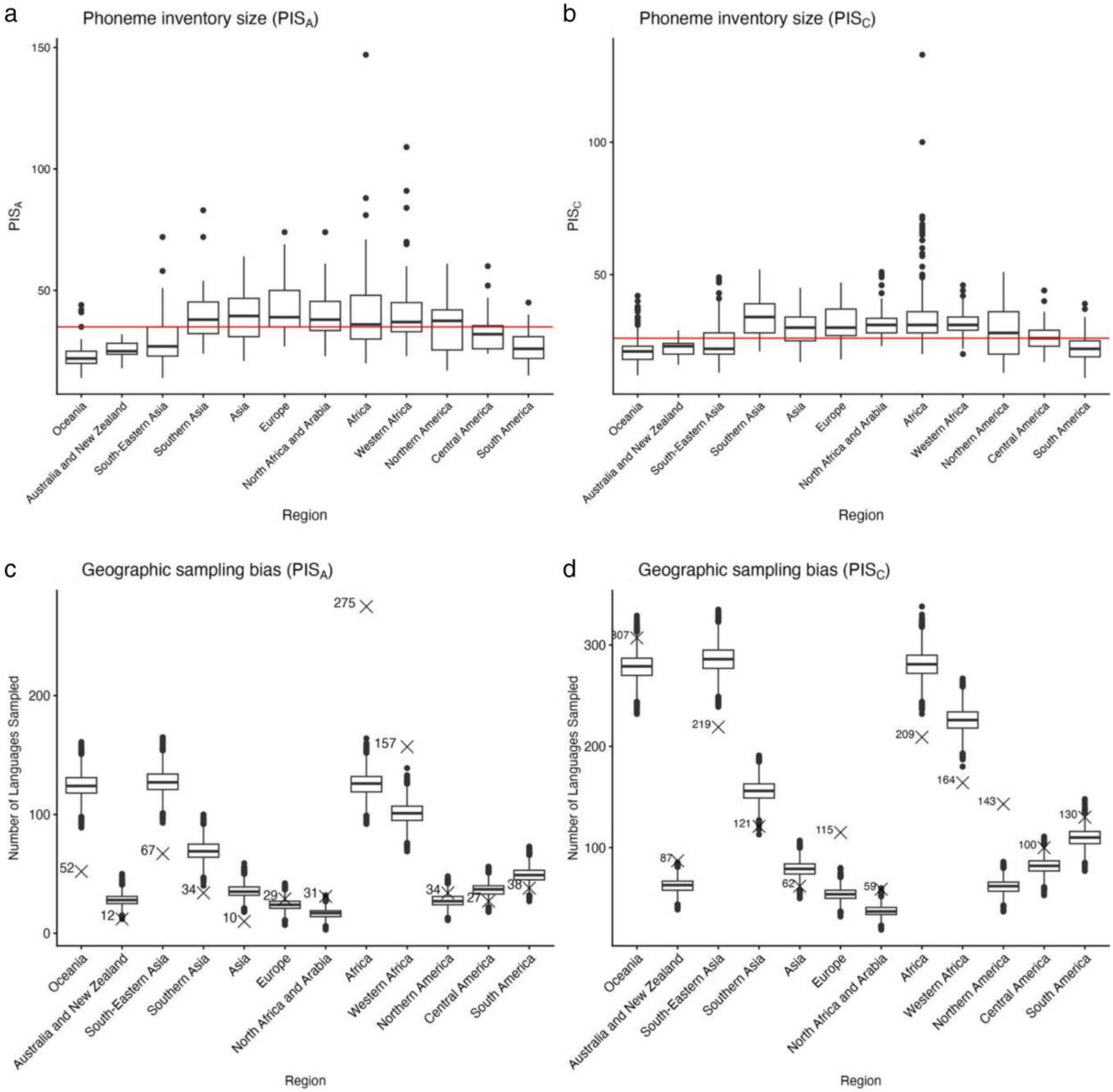

Figure 4 Regional bias in phoneme inventory size for a) PIS$_A$ (766 languages) and b) PIS$_C$ (1,716 languages): red line shows global median phoneme inventory size (35), box = interquartile range (IQR); centre line = median; top whisker = Q3 + 1.5 * IQR, bottom whisker = Q1–1.5 * IQR; dots = outliers which are greater than or less than top or bottom whisker values, respectively. Sampling effort per region for available data on phoneme inventory size for c) PIS$_A$ and d) PIS$_C$ is marked with an X and compared to the expected number of languages per region based on a random sample from the global database of languages: box = interquartile range (IQR); centre line = median; top whisker = Q3 + 1.5 * IQR, bottom whisker = Q1–1.5 * IQR); dots = outliers which are greater than or less than top or bottom whisker values, respectively.

example through horizontal transfer of phonemes, then a spatial autoregressive model (SAR) might be more appropriate, as it can incorporate a spatial lag to represent diffusion between neighbours (so that the phoneme inventory of a language influences its neighbours and *vice versa*).

We also find no evidence that the relationship between population size and phoneme inventory size is driven by research effort because the level of language documentation is not associated with phoneme inventory size. Clearly, this three-level coding of documentation is a very broad-brush measure of research effort; nonetheless, there is no suggestion in the data that under-documented languages have fewer recorded phonemes. However, we do detect sample biases in some of the analysed data, as one of the databases is biased toward relative overrepresentation of regions that have richer phoneme inventories, with relative undersampling of languages from regions with typically smaller phoneme inventories (Fig. 4). This raises the potential that indirect relationships may be driving the observed association between population size and phoneme inventories, for example, if larger languages are more likely to be better studied, resulting in more thorough phoneme inventory sizes and over-representation in databases, which have a bias towards regions with higher phoneme inventory sizes (Supplementary Figure S2). One potential way to model the various factors shaping the phoneme inventory sizes would be to use directed acyclic graphs (DAGs), which can be used to model the influence of interacting variables (e.g. Byrnes and Dee 2025). While DAG methods have been applied to understand sociolinguistic factors shaping individual variation within speaker populations (e.g. Tomić et al. 2025), there are currently no DAG methods that are applicable to cross-linguistic data, which violate several assumptions of these analyses (Roberts et al. 2020), most notably by non-independence of observations due to relatedness and proximity (Bromham and Yaxley 2023). This currently limits the applicability of DAGs to observational data that is shaped by descent with modification or similarity between neighbours, such as languages, cultures, or species.

5. Conclusions

The relationship between phoneme inventory size and speaker population size noted by Hay & Bauer is confirmed for two larger datasets. Further analysis allows discrimination between alternative explanations for this observed relationship. We find little evidence to support a direct influence of number of L1 speakers on phoneme inventory size because L1 population size is not a significant predictor of phoneme inventory size once relationships and proximity of languages are taken into account. Instead, we find that languages with distributions that suggest a greater degree of isolation—having few neighbours in contact or being confined to islands—have lower phoneme inventory sizes than would be expected on the basis of their relationships and location alone. In conclusion, this global study hints at a greater influence of features that influence borrowing of phonemes (i.e. languages in contact) than those that affect the generation or loss of phonemes (i.e. effect of population size). However, we stress that these results are correlative and cannot reveal underlying mechanisms. In any case, these measures are imperfect measures of isolation, and the results do not offer more than a partial explanation of phoneme inventory size. Nonetheless, as for the initial paper by Hay and Bauer (2007), we hope these results may stimulate interest and prompt further investigation.

Supplementary material

Supplementary material is available at *Journal of Language Evolution* online.

Conflicts of interest

None declared.

Funding

None declared.

Data and code availability

All data is from previously published sources (see Table 1). All code used in the analysis is available at https://github.com/Aejlaexctra/PHONEME_2025.git.

References

Anderson, C. et al. (2023) 'Variation in Phoneme Inventories: Quantifying the Problem and Improving Comparability', *Journal of Language Evolution*, 8: 149–68. https://doi.org/10.1093/jole/lzad011

Atkinson, Q. D. (2011) 'Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa', *Science*, 332: 346–9. https://doi.org/10.1126/science.1199295

Baranowski, M. (2013) 'On the Role of Social Factors in the Loss of Phonemic Distinctions', *Language and Linguistics*, 17: 271–95. https://doi.org/10.1017/S1360674313000038

Blevins, J. and Egurtzegi, A. (2023) 'Refining Explanation in Evolutionary Phonology: Macro-Typologies and Targeted Typologies in Action', *Linguistic Typology*, 27: 289–311. https://doi.org/10.1515/lingty-2021-0036

Blust, R. (1990) 'Patterns of Sound Change in the Austronesian Languages', *Linguistic Change and Reconstruction Methodology*, 45: 231–67. https://doi.org/10.1515/9783110886092.231

Blust, R. (2013) *The Austronesian: Revised Edition. Asia-Pacific Linguistics, Research School of Pacific and Asian Studies*. Canberra: The Australian National University.

Bostoen, K. and Sands, B. (2009) 'Clicks in South-Western Bantu Languages: Contact-Induced vs. Language-Internal Lexical Change, *Proceedings of the 6th World Congress of African Linguistics Cologne*, pp. 129–40. Cologne: Köppe Verlag.

Bromham, L. (2023) 'Language endangerment: Using analytical methods from conservation biology to illuminate loss of linguistic diversity'. *Cambridge Prisms: Extinction*, 1: e3.

Bromham, L. (2024a) 'Population Size and Language Change: An Evolutionary Perspective', *Annual Review of Linguistics*, 11: 183–208. https://doi.org/10.1146/annurev-linguistics-031422-123959

Bromham, L. (2024b) 'Solving Galton's Problem: Practical Solutions for Analysing Language Diversity and Evolution', in B. Evans, , M. K. Gallego, and L. Miceli, (eds.) *Historical Linguistics 2019: Selected Papers from the 24th International Conference on Historical Linguistics*, pp. 75–109. Canberra: John Benjamins.

Bromham, L. et al. (2018) 'Parasites and Politics: Why Cross-Cultural Studies Must Control for Relatedness, Proximity and Covariation', *Royal Society Open Science*, 5: 181100. https://doi.org/10.1098/rsos.181100

Bromham, L. et al. (2021) 'There is Little Evidence That Spicy Food in hot Countries is an Adaptation to Reducing Infection Risk', *Nature Human Behaviour*, 5: 878–91. https://doi.org/10.1038/s41562-020-01039-8

Bromham, L. et al. (2022) 'Global Predictors of Language Endangerment and the Future of Linguistic Diversity', *Nature Ecology & Svolution*, 6: 163–73. https://doi.org/10.1038/s41559-021-01604-y

Bromham, L. and Yaxley, K. J. (2023) 'Neighbours and Relatives: Accounting for Spatial Distribution When Testing Causal Hypotheses in Cultural Evolution', *Evolutionary Human Sciences*, 5: e27. https://doi.org/10.1017/ehs.2023.23

Bromham, L., Yaxley, K. J. and Cardillo, M. (2024) 'Islands are Engines of Language Diversity', *Nature Ecology & Evolution*, 8: 1991–2002. https://doi.org/10.1038/s41559-024-02488-4

Byrnes, J. E. K. and Dee, L. E. (2025) 'Causal Inference with Observational Data and Unobserved Confounding Variables', *Ecology Letters*, 28: e70023. https://doi.org/10.1111/ele.70023

Claessens, S., Kyritsis, T. and Atkinson, Q. D. (2023) 'Cross-national Analyses Require Additional Controls to Account for the Non-Independence of Nations', *Nature Communications*, 14: 5776. https://doi.org/10.1038/s41467-023-41486-1

Collins, J. (2017) 'Real and Spurious Correlations Involving Tonal Languages, *Dependencies in Language: On the Causal Ontology of Linguistics Systems*, pp. 129–39. Berlin: Language Science Press.

Creanza, N. et al. (2015) 'A Comparison of Worldwide Phonemic and Genetic Variation in Human Populations', *Proceedings of the National Academy of Sciences of the United States of America*, 112: 1265–72. https://doi.org/10.1073/pnas.1424033112

Dediu, D., Janssen, R. and Moisik, S. R. (2017) 'Language is not Isolated from its Wider Environment: Vocal Tract Influences on the Evolution of Speech and Language', *Language & Communication*, 54: 9–20. https://doi.org/10.1016/j.langcom.2016.10.002

Donohue, M. and Nichols, J. (2011) 'Does Phoneme Inventory Size Correlate with Population Size?', *Linguistic Typology*, 15: 161–70. https://doi.org/10.1515/lity.2011.011

Everett, C. (2013) 'Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives', *PLoS One*, 8: e65275. https://doi.org/10.1371/journal.pone.0065275

Everett, C., Blasi, D. E. and Roberts, S. G. (2016) 'Language Evolution and Climate: The Case of Desiccation and Tone', *Journal of Language Evolution*, 1: 33–46. https://doi.org/10.1093/jole/lzv004

Güldemann, T. and Stoneking, M. (2008) 'A Historical Appraisal of Clicks: A Linguistic and Genetic Population Perspective', *Annual Review of Anthropology*, 37: 93–109. https://doi.org/10.1146/annurev.anthro.37.081407.085109

Hammarström, H., Forkel, R. and Haspelmath, M. (2019) *Glottolog 4.0*. Jena: Max Planck Institute for the Science of Human History.

Hartmann, F. et al. (2024) 'Investigating Environmental Effects on Phonology Using Diachronic Models', *Evolutionary Human Sciences*, 6: e8. https://doi.org/10.1017/ehs.2023.33

Hay, J. and Bauer, L. (2007) 'Phoneme Inventory Size and Population Size', *Language*, 83: 388–400. https://doi.org/10.1353/lan.2007.0071

Hua, X. and Bromham, L. (2025) 'Reply to Koplenig & Wolfer: Global Language Analyses Must Account for Relationships, Location, and Unbalanced Binary Data', *Proceedings of the National Academy of Sciences of the United States of America*, 122: e2518837122. https://doi.org/10.1073/pnas.2518837122

Hua, X. et al. (2019) 'The Ecological Drivers of Variation in Global Language Diversity', *Nature Communications*, 10: 2047. https://doi.org/10.1038/s41467-019-09842-2

Jeffreys, H. (1961) *The Theory of Probability*, 3rd edn. Oxford: Oxford University Press.

Kerswill, P. and Shockey, L. (2007) 'The Description and Acquisition of variable Phonological Patterns: Phonology and Sociolinguistics, *Phonology in Context*, pp. 51–75. Basingstoke: Springer.

Lupyan, G. and Dale, R. (2010) 'Language Structure is Partly Determined by Social Structure', *PLoS One*, 5: e8559. https://doi.org/10.1371/journal.pone.0008559

Maddieson, I. (2023) 'Investigating the 'What', 'Where' and 'Why' of Global Phonological Typology', *Linguistic Typology*, 27: 245–66. https://doi.org/10.1515/lingty-2022-0076

Moran, S., Lester, N. A. and Grossman, E. (2021) 'Inferring Recent Evolutionary Changes in Speech Sounds', *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences*, 376: 20200198. https://doi.org/10.1098/rstb.2020.0198

Moran, S. and McCloy, D. (2019) *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

Moran, S., McCloy, D. and Wright, R. (2012) 'Revisiting Population Size vs. Phoneme Inventory Size', *Language*, 88: 877–93. https://doi.org/10.1353/lan.2012.0087

Naranjo, M. G. et al. (2025) 'Why Modelling Space is Hard: No Evidence for a Serial Founder Effect in Polynesian Phoneme Inventories', *Linguistics*, 63: 1493–546. https://doi.org/10.1515/ling-2024-0016

Naranjo, M. G. and Mertner, M. (2023) 'Estimating Areal Effects in Typology: A Case Study of African Phoneme Inventories', *Linguistic Typology*, 27: 455–80. https://doi.org/10.1515/lingty-2022-0037

Roberts, S. and Winters, J. (2013) 'Linguistic Diversity and Traffic Accidents: Lessons from Statistical Studies of Cultural Traits', *PLoS One*, 8: e70902. https://doi.org/10.1371/journal.pone.0070902

Roberts, S. G. et al. (2020) 'CHIELD: The Causal Hypotheses in Evolutionary Linguistics Database', *Journal of Language Evolution*, 5: 101–20. https://doi.org/10.1093/jole/lzaa001

Shcherbakova, O. et al. (2023) 'Societies of Strangers do not Speak Less complex Languages', *Science Advances*, 9: eadf7704. https://doi.org/10.1126/sciadv.adf7704

Tomić, A. et al. (2025) 'Individual Language Experience Determinants of Morphosyntactic Variation in Heritage and Attriting Speakers of Bosnian and Serbian: A Causal Inference Approach', *Linguistic Approaches to Bilingualism*, 15: 776–811. https://doi.org/10.1075/lab.24016.tom

Trudgill, P. (2004) 'Linguistic and Social Typology: The Austronesian Migrations and Phoneme Inventories', *Linguistic Typology*, 8: 305–20. https://doi.org/10.1515/lity.2004.8.3.305

Urban, M. and Moran, S. (2021) 'Altitude and the Distributional Typology of Language Structure: Ejectives and Beyond', *PLoS One*, 16: e0245522. https://doi.org/10.1371/journal.pone.0245522

Van Coetsem, F. (1988) *Loan Phonology and the two Transfer Types in Language Contact*. Dordrecht: Foris Publications.