



Investigating diachronic trends in phonological inventories using BDPROTO

Steven Moran¹  · Eitan Grossman² ·
Annemarie Verkerk³

Published online: 7 January 2020
© The Author(s) 2020

Abstract Here we present an expanded version of BDPROTO, a database comprising phonological inventory data from 257 ancient and reconstructed languages. These data were extracted from historical linguistic reconstructions and brought together into a single unified, normalized, accessible, and Unicode-compliant language resource. This dataset is publicly available and we aim to engage language scientists doing research on language change and language evolution. Furthermore, we identify a hitherto undiscussed temporal bias that complicates the simple comparison of ancient and reconstructed languages with present-day languages. Due to the sparsity of the data and the absence of statistical and computational methods that can adequately handle this bias, we instead directly target rates of change within and across families, thereby providing a case study to highlight BDPROTO's research viability; using phylogenetic comparative methods and high-resolution language family trees, we investigate whether consonantal and vocalic systems differ in their rates of change over the last 10,000 years. In light of the compilation of BDPROTO and the findings of our case study, we discuss the challenges involved in comparing the sound systems of reconstructed languages with modern day languages.

✉ Steven Moran
steven.moran@uzh.ch

✉ Annemarie Verkerk
annemarie.verkerk@uni-saarland.de

Eitan Grossman
eitan.grossman@mail.huji.ac.il

¹ Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland

² Department of Linguistics, Hebrew University of Jerusalem, Mount Scopus, 91905 Jerusalem, Israel

³ Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, 07745 Jena, Germany

Keywords Historical linguistics · Language evolution · Phylogenetics · Phonological inventories

1 Overview

In this article, we first provide some background on ancient language reconstruction and why it is interesting for the study of the evolution of human language. We then describe the BDPROTO data extraction, curation, and aggregation pipelines, before we discuss the challenges and statistical biases that arise from comparing ancient with modern-day languages. In particular, we identify a temporal bias that is problematic for traditional sampling methods in typology. Afterwards, we present a case study using phonological inventory data both for ancient and reconstructed languages, based on BDPROTO, and for currently spoken languages, using PHOIBLE (Moran et al. 2014). Finally, we discuss how consonant and vowel inventories have changed over the last 10,000 years and we detail avenues for further research in historical and evolutionary linguistics.

2 Background

The development of the comparative method in the 19th century is one of the major achievements of modern linguistics. It allows language scientists to reliably reconstruct ancient languages from prehistory in terms of form and meaning, including their vocabulary, phonology, grammar, and even, to an extent, their speakers' cultural practices. The comparative method is used to demonstrate genealogical relationships between languages and to reconstruct a PROTO-LANGUAGE, i.e., the common ancestor of a set of languages before their divergence. In short, when applying the comparative method to modern languages, linguists perform a feature-by-feature comparison of two or more languages that share a common ancestor, so that they can extrapolate the properties of their linguistic ancestor (the so-called parent language).

Comparative reconstruction is accomplished through systematic comparison of word forms in descendant languages. By identifying groups of potentially related words, so-called COGNATES are identified (words that share form and meaning and are genealogically related). Consider, for example, the modern word for 'tooth' in four Romance languages: Spanish 'diente', Italian 'dente', French 'dent', and Portuguese 'dente'. The similarities in these words could be due to chance correspondence, word borrowing, or linguistic universals (cf. Blasi et al. 2016). However, these three factors are highly unlikely in this example because each descendent form shares the same meaning and a similar phonetic form. That is, each word begins with a 'd' sound and it contains evidence of a consonant cluster 'nt'. Thus it is highly probable that the ancestral word contained a form resembling 'd ... nt ...' (Fortson 2004).

Using the principle of maximum parsimony (the preferred phylogenetic tree is the one that minimizes the total number of character changes), the proto-language word form contained simply an 'e', instead of the diphthong 'ie', as in Spanish

today. Additionally, it is not uncommon in the world's languages to drop sounds at the end of words, particularly vowels, as was probably done in French. Therefore the proto-form for the word 'tooth' shared by these Romance languages had the shape 'dente', written *dente to denote a reconstructed form. Sometimes such forms are preserved in historical records (in which they have been used to verify the accuracy of the comparative method), but more often they are hypothetical reconstructions. With a large enough amount of reconstructed vocabulary, language scientists can posit that the parent language of modern languages, in this case so-called Proto-Romance, contained a 'd' sound in its phonological inventory, i.e., its repertoire of contrastive speech sounds (Moran 2019).

Another illustration of the comparative method with a small dataset from four Polynesian languages is given in Table 1. These four related languages belong to the large Austronesian language family. The data are taken from Crowley and Bowerman (2010). Again, the words in each line are cognates and together they form cognate sets. It is immediately visible that where Tongan, Samoan and Rarotongan have /t/, Hawaiian has /k/. Similarly, where Tongan and Rarotongan have /k/, Samoan and Hawaiian have a glottal stop /ʔ/ (Table 2). Such facts, where they recur across the vocabulary, are called REGULAR SOUND CORRESPONDENCES, and they are the bread and butter of comparative reconstruction.

Since it is highly implausible that such facts are unmotivated, historical linguists assume that these cognates developed from a common ancestral word. As a result, it is assumed that the sounds attested in each cognate reflect either (i) the original sound or (ii) a sound change. The considerations for identifying a particular sound change involve a number of heuristics, including (i) majority rules, i.e., the sound that is most frequent in the cognate set is the original one, (ii) knowledge of relative frequency of sound changes, e.g., a change from /h/ to zero is more frequent than the reverse, (iii) plausible phonetic motivations, and so on (for a standard list of such heuristics, see Campbell 2004, chap. 3). In the present case, historical linguists argue that in the /t:k/ correspondence set, /t/ is original, having persisted in Tongan, Samoan, and Rarotongan, and having changed to /k/ in Hawaiian. Similarly, /k/ in Tongan and Rarotongan is considered to have been maintained from an earlier stage, and Samoan and Hawaiian /ʔ/ to be the result of a change. The sounds that are reconstructed as original are called PROTO-PHONEMES, and are marked with an asterisk. These proto-phonemes allow us to reconstruct the vocabulary of proto-languages; the proto-forms of this cognate set are *tapu 'forbidden', *tahi 'sea', *kalo 'dodge', and *aka 'root'. The changes that can be inferred from the dataset are: *t > k (Hawaiian); *h > ∅ (all but Tongan); *l > r (Samoan, Rototongan), and *k > ʔ (Samoan, Hawaiian). Incidentally, one can identify some relative chronology within these changes, as Proto-Polynesian /*k/ must have shifted to /ʔ/ in an ancestor of Hawaiian before Proto-Polynesian *t shifted to /k/; otherwise, all instances of /k/ would have shifted to /ʔ/.

Recently the comparative reconstruction approach outlined above has been implemented programmatically (Steiner et al. 2011), so that many of the time-consuming and redundant tasks of the historical linguist, e.g., inferring regular sound change, are automated (Bouchard-Côté et al. 2013; Hruschka et al. 2015). The resulting score of similarity from pairwise sets of words across all languages in

Table 1 Cognates in Polynesian (Austronesian)

Tongan	Samoaan	Rarotongan	Hawaiian	Gloss
tapu	tapu	tapu	kapu	‘forbidden’
pito	pute	pito	piko	‘navel’
tahi	tai	tai	kai	‘sea’
kalo	ʔaro	karo	ʔalo	‘dodge’
aka	aʔa	aka	aʔa	‘root’

Table 2 Sound correspondences in Polynesian (Austronesian)

Tongan	Samoaan	Rarotongan	Hawaiian
t	t	t	k
k	ʔ	k	ʔ
l	l	r	l
h	∅	∅	∅

a sample can help to identify cognates. Expert judgment is still needed, but tools (List and Moran 2013; List et al. 2018) and interfaces (List 2017) allow even the non-tech-savvy linguist to quickly identify cognates from masses of raw data, such as word lists from thousands of languages (Wichmann et al. 2017). Word lists that are coded for cognacy and phonetic similarity scores can be used as input for one of many algorithms that generate language family phylogenies. These language family trees can then be used as input to phylogenetic comparative methods that were developed by biologists for investigating the tree of life, which have now been adopted and adapted by linguists and evolutionary anthropologists to address research questions about ancient language structures, cultures, and population movements (e.g., Dediú 2010; Dunn et al. 2011, 2017; Verkerk 2015; Zhou and Bowern 2015).

A particularly ripe but under-researched area of historical linguistics is the cross-linguistic analysis of the evolution of sound systems (cf. Marsico 1999). Whereas data exist in both diachronic and synchronic forms, i.e., there are many detailed reconstructions of proto-languages and the sound systems of currently spoken languages are available online (e.g., Moran et al. 2014), what has been lacking so far is a digitally-accessible database of the sound systems of proto-languages.

Therefore, we extracted phonological inventories from the historical linguistics literature for a large number of ancient and reconstructed languages. In the next section, we describe our data extraction and aggregation pipeline and then describe the current language sample and the challenges that arise when comparing it with present-day spoken languages. Afterwards, we provide a case study that illustrates what BDPROTO has to offer in terms of research questions in historical and evolutionary linguistics.

3 The BDPROTO database

3.1 Primary data extraction

The phonological inventories in BDPROTO were extracted manually from source texts,¹ interpreted by experts, and then codified according to standardized Unicode conventions (Moran and Cysouw 2018) for the International Phonetic Alphabet (International Phonetic Association 1999). In practice, this means that we identified scientifically rigorous publications by experts in historical linguistics, who applied the historical-comparative method to synchronic datasets (e.g., word lists, phonological descriptions, grammars) and reconstructed the contrastive sound systems of proto-languages. Because linguists often use different phonetic transcription practices, we went through the source materials in detail and unified the phonetic practices into a well-defined notational convention.² This allows us to identify crosslinguistic comparative concepts for the domain of phonology (cf. Haspelmath 2010). We collected phonological inventory data into spreadsheets with separate tabs for the phonological inventories and for the metadata associated with each language.

Although each dataset (see Sect. 3.2) may have project-specific information (e.g., for the ancient languages we have descriptions of syllable structure), we specify a number of fields that are common to all input sources and are required for any additional resources that are added to BDPROTO. These fields include (their data types are given in parentheses):

- BdprotoID (int): primary key
- LanguageName (text): language name given in the source document (original resource)
- Glottocode (text): a valid Glottolog code denoting a languoid³
- LanguageFamily (text): name of top-level language family (e.g., Afro-Asiatic)
- TimeDepthYBP (int): inferred date of the proto-language in years before present
- TimeDepth (text): time depth as given in the source or elsewhere (if elsewhere, the source must be cited)
- Homeland (text): where the proposed homeland of the proto-language was (either in the original resource or in another source, but then it must be specified in the HomelandSource field)
- HomelandSource (text): source where the homeland data was taken from
- MetadataNotes (text): notes regarding the metadata file
- BibtexKey (text): BibTeX key
- Phoneme (text): the phonological segment
- PhonemeNotes (text): optional notes about the phonological segment

¹ <https://github.com/bdproto/bdproto/blob/master/sources.bib>.

² <http://phoible.github.io/conventions/>.

³ <http://cldf.clld.org/v1.0/terms.rdf#glottocode>.

The resulting datasets were put into a GitHub repository,⁴ with supplemental metadata for each inventory, including estimates for its age and the homeland where it was spoken. Both the time depth and the homeland of language families are hotly debated issues (see for example the discussion of the age and heartland of Indo-European, Bouckaert et al. 2012; Chang et al. 2015). Each data point in BDPROTO is also associated with a Glottolog language identifier, so that it is positioned within a language family phylogeny (Hammarström et al. 2017).⁵ Each inventory has one or more bibliographic citations, which are stored in a text-based BibTeX file, where the BibTeX key is given a transparent ID that is mapped to the source document, so that we can easily keep track of the digital documents (PDFs) from which the data were extracted. Lastly, an aggregation script was then written to integrate these independently collected datasets into a unified data dump for quantitative analysis (details are given in Sect. 3.2).

For data collection, aggregation, and dissemination, we use GitHub because it is a web-based hosting service with version control and issue tracking, both of which are useful for data curation and code maintenance. For example, when we spot a data or coding error, we can log it in the repository's issue tracker and assign the appropriate issue to either the data collectors or programmers, who subsequently fix and close the issue. Thus, we keep a record of how the data has been collected and what changes we have made to it. This method is in line with best practices in terms of reproducibility and the requirements of open data science. GitHub also allows us to tag stages of the data and code with version numbers that can then be converted into releases in the software engineering sense. Here we follow recommended practice by using semantic versioning, i.e., by following a major.minor.patch version numbering scheme.⁶ Thus, GitHub allows us to easily curate the BDPROTO data and code base and to integrate versions into releases, which bundle up the data and the code in into zipped file releases. These files are then archived on Zenodo,⁷ e.g., (<https://doi.org/10.5281/zenodo.3521639>, <https://doi.org/10.5281/zenodo.3522973>), an open-access repository by OpenAIRE⁸ and CERN.⁹ Zenodo provides a Digital Object Identifier (DOI)¹⁰ for each release, so that it can be cited, and importantly, so that each specific release is available for reproducibility of scientific results.

3.2 Data aggregation

The BDPROTO database is comprised of several individual datasets that have been collected by different researchers in different places at different times. The oldest is the original BDPROTO data from Marsico (1999). This dataset comprises 101

⁴ <https://github.com/bdproto/bdproto>.

⁵ <http://glottolog.org/>.

⁶ <https://semver.org/>.

⁷ <https://zenodo.org/>.

⁸ <https://www.openaire.eu/>.

⁹ <https://home.cern/>.

¹⁰ <https://www.doi.org/>.

phonological inventories of reconstructed proto-languages. Being collected in the 1990s, the data were stored in SQL tables in ISO 8859-1 encoding. We transformed the data into CSV files encoded in UTF-8 NFC with LF and no BOM, following the recommended linguistic data encoding practices put forth in Moran and Cysouw (2018). Given the legacy character encoding, we standardized character representations to current Unicode IPA standards (Moran and Cysouw 2018) and we follow the PHOIBLE IPA conventions (Moran 2012; Moran et al. 2014).¹¹

Two smaller sets of phonological inventories were collected in the Department of Comparative Linguistics at the University of Zurich. The first of these two included 15 inventories from reconstructed proto-languages. The second includes 21 inventories from ancient languages, by which we mean languages for which there are written records from which scholars have analyzed their phonologies, i.e., not languages which were reconstructed based on modern day sources. Both of these datasets were extracted from scientific publications and entered by hand into Excel spreadsheets and then exported as compliant Unicode Standard UTF-8. Each resource is kept separated in the `BDPROTO` GitHub repository and is updated periodically.

The most recently-collected dataset is part of an ongoing collaboration with the Department of Linguistics at the Hebrew University of Jerusalem. This dataset, which we refer to as `HUJI`, currently comprises 120 phonological inventories from reconstructed proto-languages extracted from expert sources.

To bring these resources together into a unified format, we aggregate their phonological inventory data and metadata with a script written in R Team (2013).¹² The aggregation script includes data sanity checks and it performs various cleaning procedures. For example, input data is notoriously messy when collected by hand, so we have several data checks that identify duplicated data or data in fields that do not follow our `BDPROTO` data specification. We use the R testing library `TESTTHAT` (Wickham 2011) to catch such errors.

The aggregation script then combines the inventory data from CSV files, joins in the additional linguistic and non-linguistic metadata (described above), and outputs the combined data sources as a denormalized CSV table and as an R data object (both are available in the GitHub repository). An example of the aggregated output is given in Table 3. After these sources have been combined, we extend the phoneme inventory data with vectors of binary distinctive features from the 37 phonetic features described in PHOIBLE (Moran et al. 2014). These include common articulatory phonetic descriptors, such as ‘labial’, ‘dental’, and ‘consonantal’, which allow us to categorize and compare the reconstructed sound systems of these proto-languages.

3.3 The language sample

In total, there are 257 phonological inventories in the combined `BDPROTO` sample; they represent 206 distinct reconstructed and ancient languages from over 100

¹¹ <http://phoible.github.io/conventions/>.

¹² <https://github.com/bdproto/bdproto/blob/master/src/aggregation.R>.

Table 3 Example of aggregated data from BDPROTO

BdprotoID	LanguageName	Glottocode	Phoneme	LanguageFamily	YBP
3	Proto-Balto-Slavic	balt1263	m	Indo-European	3000
3	Proto-Balto-Slavic	balt1263	p	Indo-European	3000
3	Proto-Balto-Slavic	balt1263	b	Indo-European	3000
3	Proto-Balto-Slavic	balt1263	w	Indo-European	3000
3	Proto-Balto-Slavic	balt1263	n	Indo-European	3000
3	Proto-Balto-Slavic	balt1263	t	Indo-European	3000
3	Proto-Balto-Slavic	balt1263	d	Indo-European	3000

different language families. The original BDPROTO sample was devised without duplicate entries and by considering the coherence of the proposed reconstructions and their relations to their modern daughter languages. The aggregation of the original BDPROTO sample with our more recent work results in duplicate data points, i.e., different reconstructions of phonological descriptions of the same proto-language or different DOCULECTS (Cysouw and Good 2013). We consider multiple entries a feature of our database, which allows users to explore and compare different reconstructions by different experts. Nevertheless, both the old and new BDPROTO language samples present novel challenges for comparative linguists. For example, how do we fairly assess the frequency of a particular phenomenon, e.g., reconstructed sounds, in our set of proto-languages?

3.4 Towards a typology of ancient phonological segment inventories

Studies in typology face several statistical biases when attempting to describe distributions of linguistic phenomena in the world's languages (Sherman 1975). For example, to say something meaningful about the distribution of sounds in today's spoken languages, one must consider both the BIBLIOGRAPHIC BIAS (Bakker 2011) (i. e., not all languages have been documented, and as such, there is a serious lack of available data, which is skewed towards larger well-described languages) and the GENEALOGICAL BIAS (languages fit into language families in a power-law-like distribution, i.e., a few languages families contain the majority of the world's languages, while nearly a third of all language families consist of a single language or its dialects, like Basque).¹³ Both biases mean that classical statistical methods for sampling are untenable for linguistic typology (Janssen et al. 2006).¹⁴

¹³ In fact, language isolates are pertinent to linguistic typology studies because they often display phonological and grammatical properties that are distinct from what we know about all other languages. Or in other words, language isolates contain a great deal of the worldwide linguistic diversity, and as such, they are particularly interesting for cross-linguistic study.

¹⁴ Another major bias is the GEOGRAPHICAL OR AREAL BIAS: languages that are spatially close are likely to influence each other, thereby making languages more similar to their neighbors and, potentially, more distinct from related languages; see Bickel (2007) on universal areality. There are additional biases that have been mentioned in the literature, such as CULTURAL BIAS (Perkins 1992), but addressing a full survey of issues related to sampling is beyond the scope of this work (see Moran 2019, 2012 for details and a historical overview).

While the *BDPROTO* sample is of course restricted by bibliographic bias (because there are relatively few language reconstructions currently available and the number of reconstructions is essentially non-exhaustive), it is also constrained by a form of genealogical bias that has not been identified in the literature, as far as we know. That is, the data points in the *BDPROTO* sample represent different stages of language families through time. This poses a problem for the direct comparison of proto-languages that have different ages, and for the direct comparison of proto-languages with present-day languages. For example, some data points in *BDPROTO* represent the root-level of a language family tree, also known as a language family *STOCK* (Nichols 1992). A stock is a genealogical group with considerable time depth, such as Indo-European. The stock is the highest reconstructible genealogical group.¹⁵ As such, stocks may exhibit regular sound correspondences that are not transparent to non-specialists and hence must be reconstructed by experts.

Other data points in *BDPROTO* are intermediate nodes in existing proposed phylogenies. For example, there are expert reconstructions of ancient Celtic, Germanic, Romance, and Balto-Slavic, each of which represents an intermediate node within the branches of the Indo-European tree, i.e., daughter languages of Indo-European but also parent languages of currently spoken languages. These genealogical classifications are referred to as *GENERA* (Dryer 2005). A language family *GENUS* is a genealogical classification whose time depth is not more than roughly 4000 years, and may be much younger. The relatedness of languages within these groups is fairly obvious without systematic comparative analysis.

Figure 1 lists 25 of the oldest languages in the *BDPROTO* sample and approximately when they were spoken. It contains both language family stocks and genera. Furthermore, some data points are nested within others, e.g., Finno-Permic is an intermediate node within the Uralic stock.

The fact that languages and families of various ages are nested recursively within larger families leads to a set of problems that we have called the *TEMPORAL BIAS*, which makes the comparison of ancient and proto-languages across time particularly problematic. First, the set of ancient and reconstructed languages is not temporally homogeneous. In other words, languages cannot be binned into “present-day” versus “ancient and reconstructed” categories. Each ancient or reconstructed language was spoken during a particular time period. In effect, *BDPROTO* (or any similar database) will include languages from 500 years ago, 2000 years ago, and, potentially, 8000 years ago. This means that from a purely chronological point of view, the blind comparison of such languages is akin to comparing an 8-year old, a 25-year old, and a 70-year old. A possible solution, in principle, would be to compare languages on the basis of their estimated ages, e.g., comparing only

¹⁵ It is generally agreed-upon that 10,000 years is the maximum time depth of reconstruction for the comparative method (Nichols 1992). Past this time depth, languages have simply had too much time to mutate in vocabulary through regular processes of sound change and it has not yet been discovered how to peer further back in time (although this is an active area of research, e.g., Pagel et al. 2013). Another tool of historical linguistics, *INTERNAL RECONSTRUCTION*, can be applied to proto-languages, thereby looking further back into linguistic prehistory; however, internal reconstruction is generally considered to be less reliable than reconstruction based on the comparative method (Ringe 2003). Finally, it has been suggested that certain structural properties of language may be deeply-conserved population markers, allowing linguists to reach even further back into human prehistory (Nichols 1994), but this is still controversial.

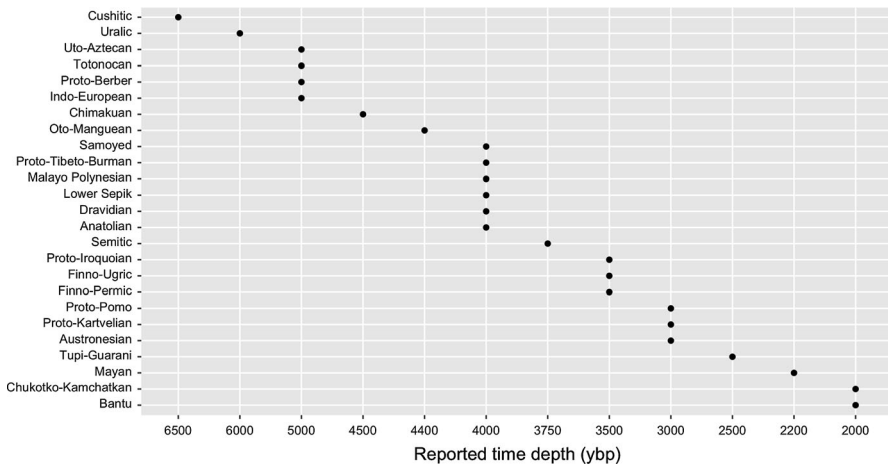


Fig. 1 Approximate age of 25 language families in the BDPROTO sample

languages that are estimated to have been spoken 1500–2000 years before present, and so on.

Secondly, however, language families are heterogeneous in terms of their size and diversification. For example, one stock may have many nodes (e.g., Indo-European), while another may have very few (e.g., Basque). This means that the blind comparison of ancient and reconstructed languages may be like comparing the branches of an angel oak tree with those of a birch tree (Fig. 2): at the same height from the ground, the number of nodes is very different, and, conversely, at the same taxonomic level (i.e., n nodes from the present), the height from the ground can be very different.

In order to make languages spoken at different time depths and at different nodes in a family tree structure comparable, one would either need to eliminate nodes from the more diversified families or to inflate the genealogical complexity of the less diversified families. Both of these would exacerbate the already-acute problem of data sparsity when existing genealogical diversity is taken into account. While most accounts estimate the number of distinct present-day languages as somewhere between 6000 and 7000, most also estimate the number of genealogically distinct stocks as relatively small. For example, Hammarström (2016) counts 424 such families, some of which are currently isolates or very small families. Such isolates are counted as stocks, even though they might be very recent or low-level nodes in families whose other members have been lost, which means that the genealogical diversity available to us for any particular time depth might be even less. On the other hand, some experts consider it plausible that some families are related to other families at an even deeper time depth than what can be confidently reconstructed using the comparative method. Nichols (1994) mentions Afroasiatic as a likely stock that cannot be confidently reconstructed, and points to plausible deeper relations between, e.g., Indo-European and Uralic, relations which cannot be proven by

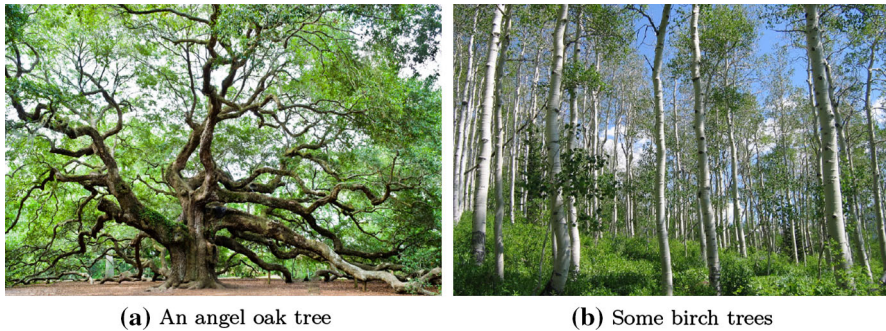


Fig. 2 Trees

currently available methods. In other words, even ‘independent’ families may represent earlier dependencies. To sum up, the total genealogical diversity represented by a present-day sample of languages, if we go back in time as far as current methods allow us, is simply not that great, which means that for statistical analysis, our data is very sparse.

All in all, the deep-time language tree structure, with recursively nested language family splits, means that classical and current approaches to dealing with genealogical bias in typological studies (e.g., language family as random effect structures in mixed effects models, Moran et al. 2012) are unlikely to be adequate, and the problem itself is potentially computationally intractable.¹⁶

Now consider, for example, if we want to compare the frequency distribution of reconstructed phonemes in proto-languages with sounds found in languages spoken today. Perhaps we want to ask whether the distribution of sounds in languages today are different from those from thousands of years ago. Figure 3 shows the 30 most frequent phonemes in the broadest sample of today’s languages (Moran et al. 2014) ($n = 1672$) in rank comparison with the frequency of reconstructed phonemes in BDPROTO.

We observe a general trend in the frequency of sounds cross-linguistically in modern spoken languages, i.e., they follow a power-law distribution in which articulatorily basic sounds, like ‘m’, ‘i’, ‘k’, ‘a’, are found across the majority languages sampled, but around half of all documented speech sounds are language-specific (as already noted by Maddieson 1984).¹⁷ The trend is similar in the BDPROTO sample; however, there are a number of differences. For example, whereas ‘m’ is the most frequent sound reported in today’s languages, in reconstructed languages it

¹⁶ There are two other aspects of the temporal bias that merit mentioning. One is methodological, i.e., the sounds of reconstructed languages tend to be normalized to ‘representative’ basic sounds, eliding phonetic detail available for present-day spoken languages. Another aspect of the temporal bias is the assumption that distributions of phonological segments is uniform across time, which may or may not be the case. As a result, phonological distributions should be visualized not only spatially, i.e., via maps, but also temporally, i.e., via a time-line.

¹⁷ We provide a more detailed report of general summary statistics and plots of the current data in an R markdown report at: https://github.com/bdproto/bdproto/blob/master/src/descriptive_stats.md.

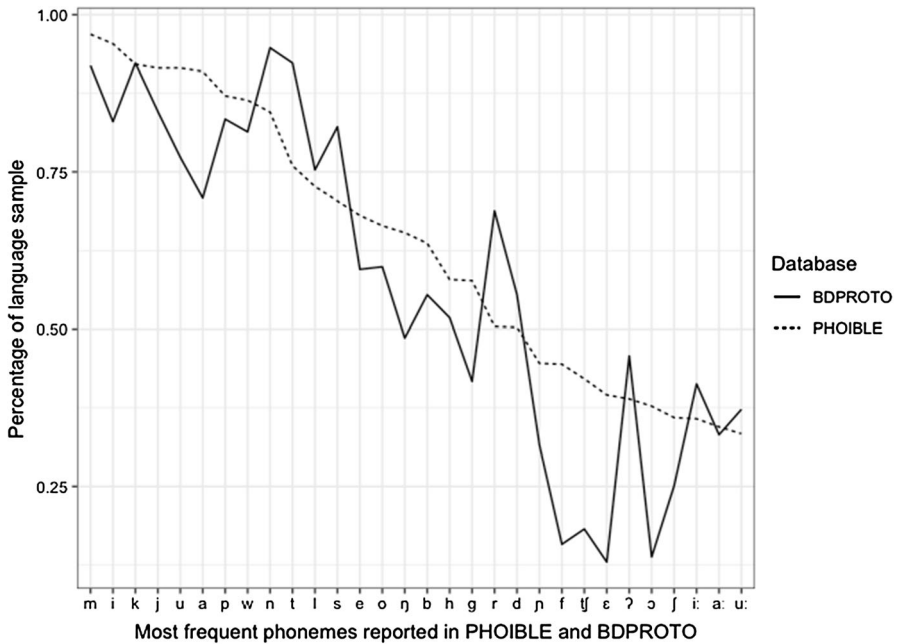


Fig. 3 The 30 most frequency phonemes reported in PHOIBLE compared with BDPROTO

is ‘n’. There are also sharp dips and rises in the reported frequency of sounds like ‘f’ and ‘r’. We can speculate on individual sounds. For example, the low frequency of ‘f’ illustrates the late emergence of labiodental sounds since the Neolithic due to changes in human bite configuration from food processing technology and agriculture (Blasi et al. 2019). The high frequency of ‘r’ likely reflects (along with other sounds including vowels) a tendency to normalize detailed synchronic sound descriptions into the basic phonological variants when using the historical-comparative method to project back in time. Even so, we need to keep in mind that the reconstruction of the actual phonetic values of ancestral proto-sounds is an unresolved problem (Coleman et al. 2015), although Bayesian computational phylogenetic approaches are being developed to try to tackle it (e.g., Blasi et al. 2019; Aston et al. 2012). We return to this and other issues in the discussion in Sect. 5.

We can at this point, however, avoid genealogical and temporal biases by investigating specific language families, regardless of their age, by comparing certain properties of proto-languages directly with the currently available data of its daughter languages. Moreover, although we cannot give precise phonetic values for reconstructed sounds from thousands of years ago, we can be fairly certain that proto-languages derived through the historical-comparative method provide us with a fairly accurate number of contrastive sounds in the phonological inventories that the parent languages would have had. Hence, we can compare whether ancient and

modern languages have shifted in their number of sounds, and in which direction, over the millennia.

4 Case study: evolutionary rates of consonant and vowel inventories

In a study of whether phonological inventories have become more or less complex over time, Marsico (1999) shows that languages dating back as far as 10,000 years are equally complex in terms of their number of segments, consonant/vowel ratio, average number of consonants and vowels, and frequency hierarchy of the segments. However, Marsico (1999) also notes that modern languages tend to have slightly more consonants today than their ancestors did in the past. The same does not apply to vowels. On average the number of consonants and vowels across proto-languages in BDPROTO are 18 and 8, respectively. In comparison, modern spoken languages have on average 22 consonants and 8 vowels (Maddieson 1984).¹⁸

This is in itself an interesting finding, because it is typically assumed that reconstructed phonological systems are likely to show *more* complexity than their daughter languages, due to inherent biases of the comparative method. For example, Fox (1995) argues that some reconstructions of the Proto-Indo-European consonant system contain more consonants than any of the daughter languages. If reconstructed languages indeed tend to have an artefactual inflation of inventory size, this would mean that the difference between reconstructed and present-day phonological inventories is *underestimated* in the present study.

Why is it that we observe more consonants in phonological inventories today than we see in reconstructed ancient languages of the past? We decided to test whether eight language families show greater rates of change in consonant inventory size as compared to vowel inventory size using phylogenetic comparative methods (see also Marsico et al. 2018, Moran and Verkerk 2018). Specifically, we use BayesTraits V3 (Meade and Pagel 2014), which implements a generalized least squares approach to modeling the evolution of continuously varying traits (Pagel 1997, 1999) and multistate traits (Pagel et al. 2004). We chose these eight language families because they have high-resolution expert-created phylogenies: Arawakan (language sample used for current study n=38; Walker and Ribeiro 2011), Austronesian (83; Gray et al. 2009), Bantu (114; Grollemund et al. 2015), Dravidian (19; Kolipakam et al. 2018), Indo-European (58; Bouckaert et al. 2012), Pama-Nyungan (134; Bowerman and Atkinson 2012), Tupi-Guarani (30; Michael et al. 2015), Turkic (19; Hruschka et al. 2015) and because they are in the BDPROTO sample.

Figure 4 shows box plots of the ranges of vowel and consonant inventory size in the language samples used for phylogenetic ancestral state estimations. P gives proto-language reconstruction from BDPROTO. R gives the ancestral state estimation

¹⁸ Note that these averages are not adjusted for phylogeny because so far there is a lack of high-resolution language phylogenies for most language families.

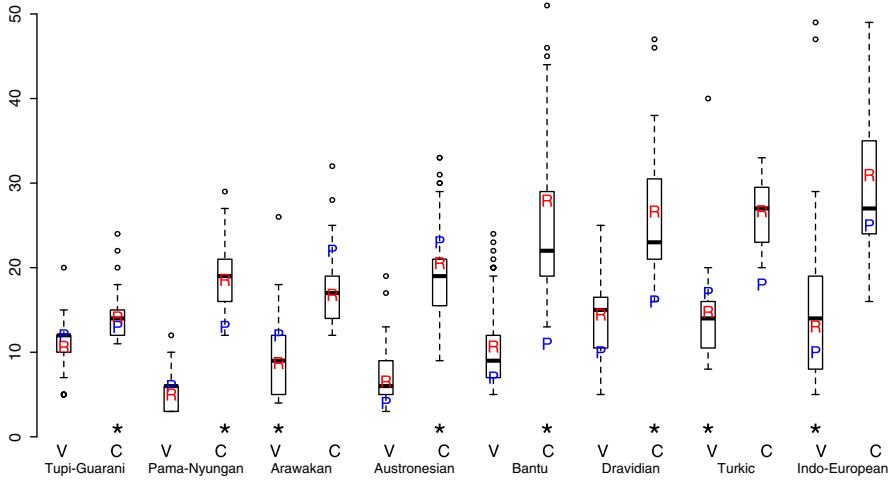


Fig. 4 Ranges, reconstructions, and ancestral state estimations of vowel and consonant inventory size

Table 4 Median rates of change in 1000s of years, continuous model

Language family	Consonants	Vowels
Pama-Nyungan	3.28 ± 0.18	0.74 ± 0.04
Tupi-Guarani	5.02 ± 0.60	4.32 ± 0.53
Austronesian	7.40 ± 0.41	2.46 ± 0.21
Turkic	12.91 ± 2.58	29.7 ± 5.66
Arawakan	17.42 ± 1.43	31.55 ± 2.94
Indo-European	29.38 ± 2.36	46.91 ± 3.96
Dravidian	51.41 ± 9.02	17.47 ± 3.69
Bantu	64.00 ± 3.01	8.53 ± 0.21

using BayesTrait V3’s analysis Continuous. An asterisk * indicates whether the rate of change of vowel or consonant inventory size is faster.¹⁹

Table 4 gives the mean rates of change of consonant and vowels on the branches of the listed phylogenetic tree sets by 1000s of years. Our results suggest a mixed picture for the acquisition of new consonants vs vowels over the last 10,000 years. In five out of eight language families sampled, the rate of change in consonants systems is greater. But in Turkic, Arawakan and Indo-European, vowel inventory

¹⁹ We also observe that ancestral state estimates of vowel and consonant inventory sizes are generally closer to the mean of the range than expert reconstructions of proto-languages. This means there is a difference between the well-worked historical-comparative method used by linguists to reconstruct proto-languages and the ancestral states generated through phylogenetic analysis. The latter optimize ancestral states by looking both at the ancestor as well as the daughter(s) of each node, which may have an effect of averaging out differences rather than picking up on (directional) trends. This observation warrants a closer evaluation using directional models of feature change.

Table 5 Median rates of change in 1000s of years, multistate model

Language family	Consonants	Vowels
Arawakan	37.08 ± 29.00	37.76 ± 28.93
Dravidian	39.36 ± 29.03	39.10 ± 29.60
Tupi-Guarani	39.73 ± 29.16	38.10 ± 29.60
Turkic	39.40 ± 29.67	41.08 ± 29.27
Indo-European	0.015 ± 23.92	26.17 ± 31.98
Austronesian	0.004 ± 0.0008	0.001 ± 0.0004
Bantu	0.037 ± 0.008	0.006 ± 0.001
Pama-Nyungan	0.002 ± 0.0003	0.0004 ± 0.00008

size changes faster than consonant inventory size. These three language families have in common a wider range of vowel inventory sizes as compared to the other families. However, if we take into account the mean and standard deviations of the rates given in Table 4, a high variance does not always entail a high rate of change and vice versa. Dravidian, for instance, has the second highest rate of change for consonants, but the Dravidian languages are less variable in their inventory size than Bantu and Indo-European.

While this is an interesting result, it is not clear a priori whether modeling change in phoneme inventory size should be done this way. BayesTraits V3's (Meade and Pagel 2014) Continuous analysis considers all numbers floats and can hence arrive at an ancestral inventory size of 8.5, for example, which is not in line with our understanding of how phoneme inventories evolve. An alternative approach would be using BayesTraits V3's Multistate analysis, which regards all inventory sizes as separate states, separate entities with no pre-defined relationship between them. However, this approach has its own problems. Bantu has 36 different consonant inventory sizes. As a result, to model these requires $36 \times 36 - 36 = 1260$ parameters (i.e., change from inventory size 6 to inventory size 7, from size 6 to inventory size 8, etc.). These have to be restricted in a meaningful way. In the following analyses, we restrict these changes to a single parameter so that we can compare the value of this parameter between consonant and vowel inventory sizes. Ideally, however, we would allow increases or decreases of, say, 2 or 3 phonemes at a time too, because if a particular feature is added to a phonological system (such as lengthening or nasalization) this may result in the instantaneous increase or decrease of the inventory across some number of affected phonemes. Such a more refined model is not possible within Multistate, which is agnostic with regard to such specificity.

Table 5 shows the median rate parameter values for modeling consonant and vowel inventory size in eight language families. The Multistate model seems to arrive at two solutions that are fundamentally different in nature: the first four families, Arawakan, Dravidian, Tupi-Guarani and Turkic, have high rates of change, with large standard deviations, both for consonants and for vowels. The last three families, Austronesian, Bantu, and Pama-Nyungan, have very low rates, and a difference between consonants and vowels, i.e., consonants evolve faster. Indo-European is somewhere in-between: for vowels we observe the first pattern, while

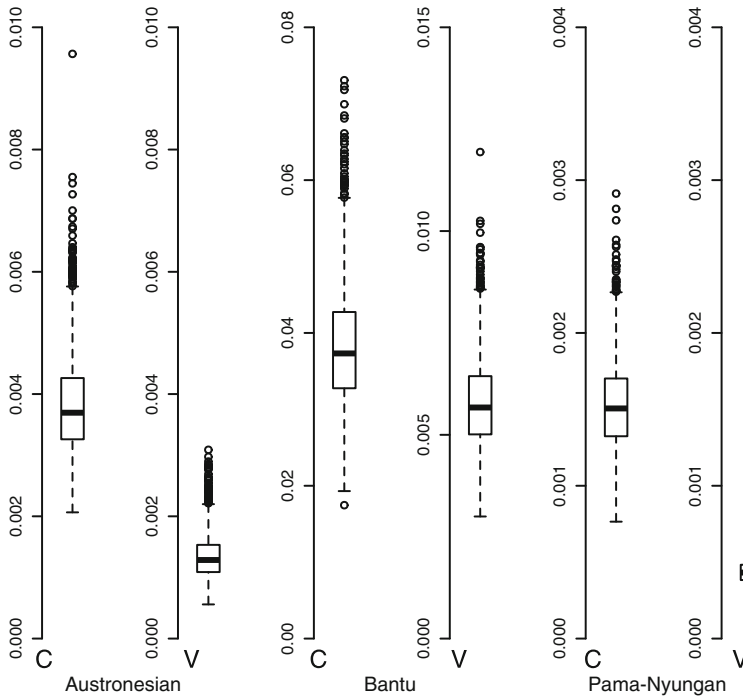


Fig. 5 Rates of change in multistate analyses for Austronesian, Bantu, and Pama-Nyungan language families

the analysis of the consonant inventories shows a bipartite distribution as it contains both very low and high rates (see the low median but the high standard deviation). The posterior distributions of the rate parameter values for Austronesian, Bantu, and Pama-Nyungan are displayed in Fig. 5.

The main difference between these two groups is the size of the family: Arawakan, Dravidian, Tupi-Guarani and Turkic are small in terms of number of daughter languages, Austronesian, Bantu, and Pama-Nyungan are big, Indo-European is in-between. The high rate typical for the first group signifies a poorly updated prior (uniform, 0–100) and thus indicates a poor fit of the data, presumably because of the small size of the datasets. The diverging results for the three bigger families show that the Multistate model is relevant for analyzing phoneme inventory size, given a large enough family. Figure 5 shows that in Austronesian, Bantu, and Pama-Nyungan, consonants evolve faster than vowels. This matches the results of the Continuous model for these three families (see Fig. 4), where it was Arawakan, Indo-European, and Turkic for which vowel inventories change faster than consonant inventories.

The results of the Continuous analyses suggest differential rates of change in consonants and vowels by language family, while the Multistate analyses are uninformative in this regard. This finding is surprising to us because the synchronic

data suggest that there is a diachronic pressure on languages to expand their consonant inventories at a greater rate than vowels, in line with the finding by Marsico (1999). For example, on average languages have more consonants than vowels, so we might expect phonological inventories to universally acquire consonants at a faster rate. Consonants are also more likely to be borrowed than vowels (Moran et al. 2014). Lastly, the synchronic data show more phonetic diversity in consonant inventories, suggesting a greater number of lexical contrasts made available by consonants. For example, there are three times as many contrastive consonants than vowels in the world's languages. Consonant inventories also range more in size, from 6 to 90 (Rotokas in Papua New Guinea vs the click language !Xu, spoken in Botswana and Namibia) than vowel qualities, which range from 2 to 14 (Maddieson 2013a, b). All in all, it is surprising that *any* language family shows a greater rate of change in vowel inventories than in consonant inventories. We end this section by highlighting the family-specific nature of patterns of change.²⁰ This finding is in contrast to the null hypothesis, i.e., that there is no relationship between family and rate of change in the domain of phonological segment inventories or parts thereof.

5 Discussion

Our findings warrant further research, but we might already speculate on where to look next. First of all, phonological segment inventories can be subject to a highly diverse range of potentially universal pressures (or biases or constraints). Such pressures can be related to the synchronic structure of phonological systems (e.g., Clements 2009), articulatory or perceptual properties of sounds (e.g., Ohala 1981, 2010), usage-based effects on languages (e.g., Bybee 2001, 2006), or the function of sounds as the basis of lexical contrasts. But they can also be related to socio-cultural, biological, or environmental factors, such as population size, genes, or humidity. Many such factors have been proposed in the literature (e.g., Pericliev 2004; Hay and Bauer 2007; Everett 2013; Greenhill 2014; Everett et al. 2015; Maddieson and Coupé 2015; Everett 2017; Benítez-Burraco and Moran 2018) and continue to be hotly debated (e.g., Moran et al. 2012; Donohue and Nichols 2011; Greenhill 2016; Roberts 2018; Mendiivil-Giró 2018). In the following, we mention just a few candidates for causal factors that shape phonological segment inventories by shaping the directionality of language change.

Inventories of both vowels and consonants can be extended through the use of secondary articulatory features. For example, a vowel space can be expanded straightforwardly by the contrastive features of length and nasalization. On the other hand, consonant inventories can be expanded by numerous features, such as length,

²⁰ Greenhill et al. (2018) caution against the assumption of uniform rates of change across all languages. They find differential rates of lexical change per language family, with some showing effects from population size, whereas others do not.

labialization, palatalization, velarization, and more.²¹ Mathematically, it may be the case that there is a greater number of dimensions for consonant inventories to expand, but that there are other constraints on how consonant or vowel inventories increase in size. Hence to create more and more vocabulary, increasing the number of contrastive sounds in the phonological inventory while keeping the number of distinctive phonetic features at a minimum is said to embody the principle of feature economy (Clements 2009). An example is given by Moran (2012), who shows that vowel systems tend to expand from the cardinal vowels through the highly economic features of length and nasalization before filling in the vowel space with peripheral vowels that require finer articulatory features and perceptual distinctions. Furthermore, there is an asymmetry of feature economy such that vowel inventories tend to be more economical than consonant inventories (Coupé et al. 2011). Thus, the articulatory and perceptual constraints that may govern the changes in phonological inventories over time must be incorporated into models of the evolution of spoken languages.

Another important consideration related to lexical contrasts and sound changes over time is that of FUNCTIONAL LOAD. Based on the premise that the primary function of a phonological system is to allow utterances to be distinguished, Hockett (1966) explored the idea that some phonological segments “do more of this job than others”. Martinet (1955, 1977) hypothesized that segments with a higher functional load would be more resistant to change. Interestingly, it has been found, albeit on the basis of a small cross-linguistic sample, that consonants tend to have a higher functional load than vowels (Oh et al. 2013). While Martinet’s hypothesis is not supported by all studies, the present study corroborates it—to the extent that the finding that consonants tend to have a higher functional load than vowels can be generalized—on the basis of a large cross-linguistic database.

Functional load, in turn, may be causally related to sound change. For example, it might be that vowels are more likely than consonants to be involved in mergers, thereby favoring the reduction of vowel inventories. Vowels may also be more likely to be involved in chain shifts, which could be taken as indicating a generally lower functional load. However, the empirical evidence does not allow us to make these generalizations, since there are no large-scale databases of sound changes, and those databases that do exist do not provide information about phonological change per se (e.g., mergers and splits). As such, we merely point to the possibility that vowels and consonants participate differentially in classical phonological sound changes.

Whatever the pressures or biases shaping phonological segment inventories may be, the findings of our case study point to several facts that warrant attention. First is the fact that the evolution of phonological inventories does not follow a universal pathway. Rather, individual families show particular biases of change. Secondly, all families show a clear directionality of change, i.e., there are no unbiased families. Bickel (2013) argues that situations like these, i.e., when there is a cross-family

²¹ Consider, for example, palatalization in Russian, which more or less doubles the number of possible lexical contrasts in Russian, while being as perceptually salient a feature as primary features like voicing (Kavitskaya 2006).

preference for change while individual families show particular preferences, point to (i) the operation of universal pressures leading to preferred pathways or directions of change, and (ii) the relative strength of these pressures. In the present case, the findings from our sample provide preliminary (and hence tentative) evidence for both pressures that lead to an expansion of consonant inventories and pressures that lead to an expansion of vowel inventories. These pressures must be relatively strong, since no unbiased families were found in our sample, but it is hard to say which pressures are stronger, because the respective proportions of vowel-increasing and consonant-increasing families in the sample are very close (5/8 vs. 3/8). Obviously, in order to further explore this avenue of research, a larger sample is needed. Likewise, much work remains to be done in order to investigate particular causal theories linking a particular pressure (e.g., functional load) with a particular change preference (e.g., higher rate of change).

However, it should also be pointed out that the results reported here may also stem at least in part from biases inherent to the data and the methods used. As we noted in Sect. 3.3, historical-comparative reconstructions of phonological inventories suffer from a number of empirical issues and sampling biases.

First, the reconstruction of actual phonetic values of ancestral proto-sounds is an unresolved problem (Blasi et al. 2019; Coleman et al. 2015; Aston et al. 2012). For many present-day languages, we have access to detailed phonetic descriptions (including audio recordings and their computerized acoustic analysis), which allow us to identify with some degree of certainty the primary phonetic realizations of particular segment types. For example, it is increasingly clear whether a voiceless stop is best analyzed as dental or alveolar in a particular language. However, reconstructed proto-languages do not allow for such detailed analysis, for obvious reasons. As a result, there is a tendency to normalize detailed synchronic sound descriptions into the basic phonological variants when using the historical-comparative method to project back in time. In practice, this means that it is likely that, e.g., taps, trills, and flaps are all normalized to /t/, or close-mid and open-mid vowels are normalized to /e/, and so on. Additionally, reconstructing sounds relies on the principle of maximum parsimony, which fails to take into account sound change reversals, i.e., a sound change that takes place, but then reverses back to its previous sound. Presently, these problems limit the overall accuracy of statements about the worldwide distribution of sounds and their precise phonetic values in ancient languages. This is a ripe area for future research (cf. Blasi et al. 2019; Coleman et al. 2015; Aston et al. 2012).

Second, there are several biases regarding statistical sampling that are problematic for cross-linguistic language data in general, and BDPROTO in particular. These include the bibliographic bias (not all languages are documented, so our data samples are incomplete) and the genealogical bias (language families contain an uneven and skewed distribution of languages due to the effects of history). In addition to these well-described biases, we propose a novel problem for the comparative study of ancient and modern languages in terms of language evolution, i.e., the temporal bias. The temporal bias is due to language family trees being recursively nested data structures, which in turn encode multiple levels of language

family splits, through time.²² The temporal bias presents a potentially computationally intractable problem for current statistical approaches that address genealogical bias in comparative linguistic studies. Namely, to account for autocorrelation, which arises from the inter-relatedness of languages within language families, the confound ‘language family’ is typically treated as a random structure in mixed effects models. However, approaches that can deal with nested temporal data points must be found if we are to describe in more detail the differences in phonological systems of ancient and modern languages cross-linguistically.

Hence, another avenue for future research is to investigate the phonetic and phonological features involved in creating lexical contrasts in languages. One approach is taken for example by Nikolaev and Grossman (2018), who count places of articulation, rather than phonemes. They argue that adding new secondary articulations is easier than adding new places of articulation, so their model presupposes that places are more stable than individual segment counts. The investigation of broad phonetic and phonological features and how they may have evolved in phonological systems over time is another area for further research using BDPROTO.

6 Conclusion

In this article, we present an expanded version of BDPROTO (Marsico et al. 2018), an open-access database of phonological inventories from a sample of 257 ancient and reconstructed languages.²³ BDPROTO provides a rich resource for investigating historically reconstructed languages and whether they show any significant changes with languages spoken today. After an initial brief overview of the historical-comparative method, we describe the data extraction and aggregation techniques that we used to create the BDPROTO database. Finally, in a case study we use phylogenetic methods to show that the evolution of consonant and vowel systems have differential rates of change – an unexpected observation given what we know about the ancient and reconstructed languages in the BDPROTO sample and their modern descendants. We hope that this resource will be used by both traditional philologists and computational linguist alike, and everybody in-between. Since there are often several reconstructions for a given proto-language, collecting these in one place will allow for critical evaluation and further study.

This study also identifies a type of bias in typological studies that has not yet been identified, namely, temporal bias. This bias stems from the fact that languages are nested within families that are nested within larger families, which are in turn nested within larger families. This problem is likely to be computationally intractable, at least with respect to traditional methods of treating family as a random effect. This is in line with the finding of Piantadosi and Gibson (2014), who

²² Language splits are, in fact, simplifications of the diversification of languages, and as such, they are typically modeled as binary-branching language family trees.

²³ <https://github.com/bdproto/bdproto/>.

estimate that one would need hundreds of independent languages in order to demonstrate a linguistic universal above and beyond genealogical and areal effects, and are generally pessimistic about the sample size of present-day languages and families providing an empirical basis for such inferences.

A way forward may be found in approaches that directly target language change. At present, such approaches focus on estimating the probabilities of particular changes (e.g., the Family Bias Method, Bickel 2013). However, there may be a limit to what one can do with increasingly sophisticated statistical analyses applied to currently available data. Ultimately, it is likely that in order to directly model language change, resources much richer than those currently available will be necessary. For making progress in research on the evolution of phonological systems, a suite of resources—of which BDPROTO would be part of—can be envisioned. Other such future resources include (i) extensive cross-linguistic databases of sound change, both of pathways (represented, minimally, as input-output processes in particular contexts) and of classical types of phonological change, such as merger and split; (ii) cross-linguistic databases of contact-induced phonological change, such as the borrowing of segments and other phonological properties (such as tone); and (iii) databases of expert-established and computationally-derived phylogenies that include best estimates of time depths of the various nodes of families.

Acknowledgements Open access funding provided by Max Planck Society. Thanks to Egidio Marsico, Sebastien Flavier, Ian Maddieson, and Joël Brogniart, for data collection, curation, and for making openly available the original BDPROTO data. Balthasar Bickel and Paul Widmer provided the ancient languages data. Special thanks to Hanan Amouyal, Patrick Haller, Gali Katsir, Júda Ronén, Lilja Sæbø, Layla Schwartz, and Yoav Yosef for recent data entry. This work was inspired by Swiss National Science Foundation grant/award number PCEFP1_186841 (Steven Moran, PI).

Author contributions SM designed the research. SM, EG collected and curated the data. SM designed and developed the database. AV, SM designed and implemented the case study. SM, AV, EG wrote the paper.

Funding Open access funding provided by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aston, J. A., Buck, D., Coleman, J., Cotter, C. J., Jones, N. S., Macaulay, V., et al. (2012). Phylogenetic inference for function-valued traits: Speech sound evolution. *Trends in Ecology and Evolution*, 27(3), 160.
- Bakker, D. (2011). Language sampling. In J. J. Song (Ed.), *Handbook of linguistic typology*. Oxford: Oxford University Press.
- Benítez-Burraco, A., & Moran, S. (2018). Editorial: The adaptive value of languages: Non-linguistic causes of language diversity. *Frontiers in Psychology*, 9, 1827. <https://doi.org/10.3389/fpsyg.2018.01827>.
- Bickel, B. (2007). Typology in the 21st century: Major current developments. *Linguistic Typology*, 11, 239.
- Bickel, B. (2013). Distributional biases in language families. In B. Bickel, L. A. Grenoble, D. A. Peterson, & A. Timberlake (Eds.), *Language typology and historical contingency* (pp. 415–444). Amsterdam: Benjamins.
- Blasi, D.E., Moran, S., Moisiuk, S.R., Widmer, P., Dediu, D., & Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363(6432). <https://doi.org/10.1126/science.aav3218>.
- Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39), 201605782.
- Bouchard-Côté, A., Hall, D., Griffiths, T. L., & Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11), 4224.
- Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., et al. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097), 957.
- Bowern, C., & Atkinson, Q. D. (2012). Computational phylogenetics and the internal structure of Pama-Nyungan. *Language*, 88, 817.
- Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.
- Bybee, J. (2006). *Frequency of use and the organization of language*. Oxford: Oxford University Press.
- Campbell, L. (2004). *Historical linguistics: An introduction*. Edinburgh: Edinburgh University Press. <https://books.google.co.uk/books?id=EjXrrOJhex8C>.
- Chang, W., Cathcart, C., Hall, D., & Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1), 194.
- Clements, G.N. (2009). The role of features in phonological inventories. In E. Raimy & C.E. Cairns (Eds.) *Contemporary views on architecture and representations in phonology* (pp. 19–68). Cambridge: MIT Press.
- Coleman, J., Aston, J., & Pigole, D. (2015). Reconstructing the sounds of words from the past. In *The Scottish Consortium for ICPHS 2015, Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: The University of Glasgow.
- Coupé, C., Marsico, E., & Philippson, G. (2011). How economical are phonological inventories?. In *Proceedings of the 17th International Congress of Phonetic Sciences*.
- Crowley, T., & Bowern, C. (2010). *An introduction to historical linguistics*. Oxford: Oxford University Press.
- Cysouw, M., & Good, J. (2013). Languoid, doculect and Glossonym: Formalizing the notion ‘Language’. *Language Documentation and Conservation*, 7, 331.
- Dediu, D. (2010). A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society B*, 278, 474.
- Donohue, M., & Nichols, J. (2011). Does phoneme inventory size correlate with population size? *Linguistic Typology*, 15(2), 161.
- Dryer, M. S. (2005). Genealogical language list. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The World Atlas of Language Structures* (pp. 584–644). Oxford: Oxford University Press.
- Dunn, M., Dewey, T. K., Arnett, C., Eythórsson, T., & Barðdal, J. (2017). Dative sickness: A phylogenetic analysis of argument structure evolution in Germanic. *Language*, 93(1), e1.
- Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345), 79.

- Everett, C. (2013). Evidence for direct geographic influences on linguistic sounds: The case of ejectives. *PLoS ONE*, 8(6), e65275.
- Everett, C. (2017). Languages in drier climates use fewer vowels. *Frontiers in Psychology*, 8, 1285. <https://doi.org/10.3389/fpsyg.2017.01285>.
- Everett, C., Blasi, D. E., & Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences*, 112(5), 1322.
- Fortson, B.W. (2004). *Indo-European language and culture: An introduction*. No. 13 in Blackwell Textbooks in Linguistics. Oxford: Blackwell.
- Fox, A. (2000). On simplicity in linguistic reconstruction. In J.C. Smith, & D. Bentley (Eds.), *Historical linguistics 1995. Selected papers from the 12th international conference on historical linguistics, Manchester, August 1995* (pp. 99–110). Amsterdam: John Benjamins.
- Gray, R. D., Drummond, A. J., & Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, 323(5913), 479.
- Greenhill, S.J. (2014). Demographic correlates of language diversity. In: *The Routledge handbook of historical linguistics*. London: Routledge.
- Greenhill, S. J. (2016). Overview: Debating the effect of environment on language. *Journal of Language Evolution*, 1(1), 30. <https://doi.org/10.1093/jole/lzv007>.
- Greenhill, S. J., Hua, X., Welsh, C. F., Schneemann, H., & Bromham, L. (2018). Population size and the rate of language evolution: A test across Indo-European, Austronesian, and Bantu languages. *Frontiers in Psychology*, 9, 576. <https://doi.org/10.3389/fpsyg.2018.00576>.
- Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., & Pagel, M. (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, 112(43), 13296.
- Hammarström, H. (2016). Linguistic diversity and language evolution. *Journal of Language Evolution*, 1(1), 19. <https://doi.org/10.1093/jole/lzw002>.
- Hammarström, H., Forkel, R., & Haspelmath, M. (2017). Glottolog 3.0. <http://glottolog.org>. Accessed 15 Jan 2019.
- Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3), 663.
- Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, 83, 388.
- Hockett, C.F. (1966). The quantification of functional load: A linguistic problem. Tech. rep., Rand Corporation.
- Hruschka, D. J., Branford, S., Smith, E. D., Wilkins, J., Meade, A., Pagel, M., et al. (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, 25(1), 1.
- International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.
- Janssen, D., Bickel, B., & Zúñiga, F. (2006). Randomization tests in language typology. *Linguistic Typology*, 10, 419.
- Kavitskaya, D. (2006). Perceptual salience and palatalization in Russian. *Laboratory Phonology*, 8, 589.
- Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., et al. (2018). A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, 5, 171504.
- List, J. M. (2017). A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations* (pp. 9–12). Valencia: Association for Computational Linguistics. <http://edictor.digling.org>.
- List, J. M., Greenhill, S., Tresoldi, T., & Forkel, R. (2018). *Lingpy: A python library for historical linguistics. version 2.6.4*. <https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy>. <http://lingpy.org>.
- List, J.M., & Moran, S. (2013). An open source toolkit for quantitative historical linguistics. In *ACL (Conference System Demonstrations)*. pp. 13–18.
- Maddieson, I. (1984). *Pattern of sounds*. Cambridge: Cambridge University Press.
- Maddieson, I. (2013a). Consonant inventories. In M.S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/1>. Accessed 15 Jan 2019.

- Maddieson, I. (2013b). Vowel quality inventories. In M.S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/2>. Accessed 15 Jan 2019.
- Maddieson, I., & Coupé, C. (2015). Human spoken language diversity and the acoustic adaptation hypothesis. *The Journal of the Acoustical Society of America*, 138(3), 1838.
- Marsico, E. (1999). What can a database of proto-languages tell us about the last 10,000 years of sound changes? In *Proceedings of the XIVth International Congress of Phonetic Sciences* (pp. 353–356). San Francisco.
- Marsico, E., Flavier, S., & Moran, S. BDPROTO 1.0 (version 1.1). <https://doi.org/10.5281/zenodo.3521639>.
- Marsico, E., Flavier, S., Verkerk, A., Moran, S. (2018). BDPROTO: A database of phonological inventories from ancient and reconstructed languages. In N.C.C. chair, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).
- Martinet, A. (1955). *Économie des changements phonétiques. Traité de phonologie diachronique*. Francke.
- Martinet, A. (1977). Some basic principles of functional linguistics. *La Linguistique*, 13(1), 7.
- Meade, A., & Pagel, M. (2014). Bayestrains v2. <http://www.evolution.rdg.ac.uk/BayesTraitsV3/BayesTraitsV3.html>. Accessed 15 Jan 2019.
- Mendívil-Giró, J. L. (2018). Why don't languages adapt to their environment? *Frontiers in Communication*, 3, 1. <https://doi.org/10.3389/fcomm.2018.00024>.
- Michael, L., Chousou-Polydouri, N., Keith, B., Donnelly, E., Meira, S., Wauters, V., et al. (2015). A Bayesian phylogenetic classification of Tupí-Guaraní. *LIAMES - Línguas Indígenas Americanas*, 15, 193.
- Moran, S. (2012). Phonetics information base and lexicon. Ph.D. thesis, University of Washington. <https://digital.lib.washington.edu/researchworks/handle/1773/22452>.
- Moran, S. (2019). Phonological inventories. *Oxford Research Encyclopedia of Linguistics*. <https://doi.org/10.1093/acrefore/9780199384655.013.112>.
- Moran, S., & Cysouw, M. (2018). *The Unicode cookbook for linguists*. Berlin: Language Science Press.
- Moran, S., & Grossman, E. BDPROTO 1.1 (version 1.1). <https://doi.org/10.5281/zenodo.3522973>.
- Moran, S., McCloy, D., & Wright, R. (2014). *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://phoible.org/>. Accessed 15 Jan 2019.
- Moran, S., McCloy, D., & Wright, R. (2012). Revisiting population size vs. phoneme inventory size. *Language*, 88(4), 877.
- Moran, S. & Verkerk, A. (2018). Differential rates of change in consonant and vowel systems. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII)*. April 16–19, Toruń, Poland. <http://evolang.org/torun/proceedings/papertemplate.html?p=98>.
- Nichols, J. (1992). *Linguistic diversity in space and time*. Chicago: University of Chicago Press.
- Nichols, J. (1994). The spread of language around the pacific rim. *Evolutionary Anthropology: Issues, News, and Reviews*, 3(6), 206. <https://doi.org/10.1002/evan.1360030607>.
- Nikolaev, D., & Grossman, E. (2018). Areal sound change and the distributional typology of affricate richness in Eurasia. *Studies in Language*, 42(3), 562.
- Oh, Y.M., Pellegrino, F., Coupé, C., & Marsico, E. (2013). Cross-language comparison of functional load for vowels, consonants, and tones. In *Proceedings of the Annual Conference of the International Speech Communication Association*. (pp. 3032–3036). Graz: INTERSPEECH.
- Ohala, J.J. (1981). The listener as a source of sound change. In C. Masek, R. Hendrik & M. Miller (Eds.) *Papers from the Parasession on language and behavior* (pp. 178–203). Chicago: Chicago Linguistic Society.
- Ohala, J.J. (2010). The relation between phonetics and phonology. *The handbook of phonetic sciences* (2nd ed., pp. 653–677).
- Pagel, M. (1997). Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, 26(4), 331.
- Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756), 877.
- Pagel, M., Atkinson, Q. D., Calude, A. S., & Meade, A. (2013). Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*, 110(21), 8471.

- Pagel, M., Meade, A., & Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*, 53, 673–684.
- Pericliev, V. (2004). There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology*, 8(3), 376.
- Perkins, R. D. (1992). *Deixis, grammar, and culture*. Amsterdam: Benjamins.
- Piantadosi, S. T., & Gibson, E. (2014). Quantitative standards for absolute linguistic universals. *Cognitive Science*, 38(4), 736.
- R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org/>.
- Ringe, D. (2003). Internal reconstruction. In B. D. Joseph & R. D. Janda (Eds.), *The Handbook of Historical Linguistics* (pp. 244–261). Oxford: Blackwell.
- Roberts, S. G. (2018). Robust, causal, and incremental approaches to investigating linguistic adaptation. *Frontiers in Psychology*, 9, 1. <https://doi.org/10.3389/fpsyg.2018.00166>.
- Sherman, D. (1975). Stop and fricative systems: A discussion of paradigmatic gaps and the question of language sampling. In *Working Papers on Language Universals*, vol. 17. Stanford University.
- Steiner, L., Stadler, P. F., & Cysouw, M. (2011). A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1), 89.
- Verkerk, A. (2015). Where do all the motion verbs come from?: The speed of development of manner verbs and path verbs in Indo-European. *Diachronica*, 32(1), 69.
- Walker, R. S., & Ribeiro, L. A. (2011). Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1718), 2562.
- Wichmann, S., Holman, E. W., & Brown, C. H. (2017). *The ASJP database (version 17)*. <http://asjp.cldd.org/>. Accessed 15 Jan 2019.
- Wickham, H. (2011). testthat: Get started with testing. *The R Journal*, 3, 5.
- Zhou, K., & Bowerman, C. (2015). Quantifying uncertainty in the phylogenetics of Australian numeral systems. *Proceedings of the Royal Society B*, 282(1815), 20151278.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.