Journal of Language Evolution, 2023, **8**, 149–168
https://doi.org/10.1093/jole/lzad011
Advance access publication 24 November 2023
Research article

OXFORD

Variation in phoneme inventories: quantifying the problem and improving comparability

Cormac Anderson,[1,*] Tiago Tresoldi,[2] Simon J. Greenhill,[1,3] Robert Forkel,[1] Russell Gray,[1,4] and Johann-Mattis List[1,5]

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, Germany
[2]Institutionen för lingvistik och filologi, Box 635, Uppsala University, 751 26 Uppsala, Sweden
[3]School of Biological Sciences, The University of Auckland, Private Bag 92019, Auckland Mail Center, Auckland 1142, New Zealand.
[4]School of Psychology, The University of Auckland, Auckland, Private Bag 92019, Auckland Mail Center, Auckland 1142, New Zealand
[5]Chair of Multilingual Computational Linguistics, University of Passau, Dr.-Hans-Kapfingerstr. 14d, 94032 Passau, Germany.

[*]Corresponding author: Department of Linguistic and Cultural Evolution, Max-Planck-Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, Germany. E-mail: cormacanderson@gmail.com

For over a century, the phoneme has played a central role in linguistic research. In recent years, collections of phoneme inventories, originally designed for cross-linguistic purposes, have increasingly been used in comparative studies involving neighbouring disciplines. Despite the extended application of this type of data, there has been no research into its comparability or tests of its reliability. In this study, we carry out a systematic comparison of nine popular phoneme inventory collections. We render them comparable by linking them to standardised formats for the handling of cross-linguistic datasets, develop new measures to test both size and similarity, and release the organised data in supplementary material. We find considerable differences in inventories supposedly representing the same language variety, both in terms of size and transcriptional choices. While some of these differences appear to be predictable, reflecting design decisions in the different collections, much of the observed variation is unsystematic. These results should sound a note of caution for comparative studies based on phoneme inventories, which we suggest need to take the question of comparability more seriously. We make a number of proposals for improving the comparability of phoneme inventories.

Keywords: phenome; inventories; language comparison.

1. Introduction

For a century now, the phoneme has been the most frequently employed means of representing sound in linguistic descriptions. Phoneme inventories are the sets of phonemes associated with a given language variety and are frequently included in grammatical descriptions and phonological illustrations, typically in the form of charts such as those used to present the International Phonetic Alphabet (IPA, Handbook of the IPA 1999). The widespread availability of phoneme inventories and their highly conventionalised presentation makes them an ideal target for large typological collections, whose compilation dates back to the late 1970s (Crothers et al., 1979). These collections, whether regionally focused or global in scope, collect and display inventories aggregated from the work of multiple authors and sources in a unified form. The number of these collections has increased over time, particularly with the quantitative turn in comparative linguistics since the beginning of the 21st century (Levinson and Evans 2010), which has led to an increased use of quantitative methods and the compilation of large typological databases such as the World Atlas of Language Structures Online (https://wals.info, Dryer and Haspelmath 2013) and Grambank (https://grambank.clld.org/, Skirgård et al. 2023).

Originally, these collections of phoneme inventories were compiled for applications in phonological typology, generating, and testing hypotheses about the nature of human sound systems (e.g. Crothers

1978; Maddieson 1984). This research has continued up to the present day (e.g. Maddieson 2016; Dediu and Moisik 2019; Easterday 2019; Johansson et al., 2020), but since most collections are freely accessible (Donohue et al., 2013 being an exception) and include a large range of different language varieties, recent years have seen an increase in studies attempting to correlate phoneme inventory data with other variables, including social factors such as community size (Pericliev 2004; Trudgill 2004; Atkinson 2011; contra Donohue and Nichols 2011; Moran et al., 2012), subsistence strategies (e.g. Blasi et al., 2019; Everett and Chen 2021), and ecology (Everett et al., 2015; Maddieson and Coupé 2015; Everett 2017). In some cases, phoneme inventories have been used in studies of gene-culture coevolution (Creanza et al., 2015), or as a shortcut for language comparison, with application to human prehistory (Atkinson 2011; Ceolin et al., 2020). This has led at times to very far-fetched claims, such as the attempt by DeMille et al. (2018) to correlate specific regulatory genes with phoneme distributions, or the claim by Georgiou and Kilani (2020) that the transmission of COVID-19 was favoured by the presence of aspirated consonants in a sample of a few dozen languages.

In spite of this integral role that phoneme inventories have played in studies on such a broad range of topics, reflection on the overall comparability of phoneme inventory data has been lacking. It is rather taken for granted that phoneme inventory data is reliable, or at least sufficiently so to put the results of these investigations on a firm footing. This high trust in the robustness of phoneme inventories as data is surprising in the light of phonological theory, where scholars have long argued that phoneme inventories cannot be extracted from spoken languages in the same way in which one might measure physical entities such as mass and temperature. A phoneme inventory is the product of a linguistic analysis in which the practice of individual linguists can impact results in a nontrivial way. These differences in linguistic analysis are liable to affect the results of secondary studies which make use of phoneme inventory data, especially when these treat phonemes as independent characters, or which compare inventories on the basis of size. This has led numerous phonologists to criticise the use of phoneme inventories as a measure (e.g. Simpson 1999; Coupé et al., 2009; Deutscher 2009; Ohala 2009).

The phoneme concept underwent extensive theoretical elaboration by linguists from the 1920s onwards and different schools of structuralism held sometimes quite varying positions on the nature of the phoneme. Key disagreements included whether the phoneme was a physical or mental entity (see Twaddell 1935; Halle 1963); whether it can be considered to have positive substantive content or is defined in relation to other terms within a system (see the discussion in Anderson 1985). There was also a vivid methodological debate on the problems of establishing what is phonemic (e.g. Sapir 1933; Chao 1934) and how the phoneme system of a language should be determined (e.g. Harris 1951; Reformatsky 1970).

These debates are relevant to the question of comparability. If the phoneme is to be defined negatively, in terms of its relationship to other terms within a system, as was the position of the most influential European structuralists (Saussure 1916; Trubetzkoy, 1939; Hjelmslev 1943), then individual phonemes in different languages are not strictly commensurable (Simpson 1999), and it is rather only systems that can be validly compared (see also Sapir 1925). The opposing position is rather that the phoneme is positively defined as having phonetic substance, a view held by a good number of structuralists of different schools (Baudouin de Courtenay 1894; Bloomfield 1933; Jones 1950) and advocated for by Maddieson (1984: 160), who states that 'phonological segments can (and should) be characterised by phonetic attributes'. It should be noted that most contemporary approaches to phonology do not privilege the phoneme as a representational technology and there has been some recent debate on how phonological comparison might proceed without using the classical phoneme (i.a. Vaux 2009; Kiparsky 2018).

For the purposes of this paper, we assume that the terms of phoneme inventories can be compared in theory. Our goal is rather to investigate the robustness of this comparison in practice. However, some of the insights of the structuralists are relevant to our study. Since Chao (1934), linguists have recognised that multiple phonemic solutions are possible for any given language, distinguishing between overanalysis, whereby a phonetically simple span is represented using multiple symbols, and underanalysis, whereby a phonetically compound span is represented using a single symbol. This distinction is particularly pertinent for the analysis of entities such as those that, if analysed as unit phonemes, are termed long vowels and diphthongs, geminate consonants, affricates, prenasalised segments, and segments with secondary localisation. Analytical decisions over whether to overanalyse or underanalyse in a given instance can lead to considerable differences between inventories.[1]

In order to compare phoneme inventories across different language varieties, it is first necessary to assume that phoneme inventories compiled from different sources for the same language variety do not show significant variation. Cross-linguistic studies rest on the implicit assumption that data for individual varieties are robust and reliable. However, given these known theoretical problems of phonemic analysis, it is worth

testing the extent to which this assumption holds. In order to evaluate the degree of variation in phoneme inventories for individual varieties, we have applied standardisation and normalisation procedures to nine datasets derived from large phoneme inventory collections. These techniques allow us to evaluate the robustness of inventory data compiled by different authors and derived from various sources.

In discussing the comparability of phoneme inventories, we have to distinguish two separate, though related, types of comparability. Firstly, there is what we might call systemic comparability: to what extent do two inventories reflect the same underlying analysis of the sound system of a given language? Secondly, there is what we might call graphemic comparability: to what extent are the actual symbols used for cognate characters in two inventories the same?

Comparing the systemic properties of phoneme inventories is not a straightforward enterprise. The system of oppositions underlying the phonemic system of different language varieties cannot easily be compared, as they reflect different underlying patterns of allophony and underspecification. The same grapheme may be used to represent phonemes with radically different ranges of realisation, while different graphemes may be used to represent phonemes with similar function (List 2019), at the same 'point in the pattern' to use the terminology of Sapir in the very first issue of Language (Sapir 1925). For this reason, we have relied primarily on inventory sizes as a proxy for systemic comparability. While we recognise the limitations of this, it is also the case that many secondary studies using phoneme inventory data use inventory size as a variable, so this comparison is anyway relevant. As we will show, phoneme inventory sizes vary quite widely in our sample even for very well-described languages.

While graphemic comparability is more tractable, in that it involves comparing individual symbols and not systemic properties, it nevertheless presents a considerable practical challenge. There is great diversity with respect to the number of graphemes used in the larger collections (see the overview in Anderson et al., 2018). In order to investigate graphemic similarity, we link all of the phoneme data in our study to the Cross-Linguistic Transcription Systems (CLTS, https://clts.clld.org, List et al., 2021), which allows us to compute statistics on the use of phonemes and phonological classes in the different datasets under consideration. As a measure, we use a strict similarity measure based on Jaccard similarity (see Batagelj and Bren 1995), treating inventories as sets of graphemes.

In the following, we will describe the datasets used in this study, explain how they were normalised and standardised, and illustrate how we use them to measure phoneme inventory comparability for supposedly identical language varieties. We then present our results and discuss some of the reasons behind them. We conclude by proposing some new ideas regarding the compilation, application, and interpretation of phoneme inventories in the linguistic literature and beyond.

2. Materials and methods

2.1 Materials

For our study, we extracted nine datasets from published phoneme inventory collections. Four of these datasets have global scope, while a further five have a regional focus. This section explains the provenance and extraction of these datasets.

JIPA provides a collection of phoneme inventories for 147 languages extracted from the well-known Illustrations of the IPA series in the Journal of the International Phonetic Association, coded by Baird et al. (2021). LAPSyD, the Lyon-Albuquerque Phonological Systems Database (Maddieson et al., 2013, https://lapsyd.huma-num.fr/lapsyd/) consists of 584 sound inventories which we extracted from the original database. Additional datasets were extracted from PHOIBLE, Phonetic Information Base, and Lexicon (Moran and McCloy 2019, https://phoible.org), which provides a unified presentation for a number of distinct collections, including earlier ones (such as the Stanford Phonology Archive, Crothers et al., 1979) and collections with an areal focus. From this, we extracted UPSID, the UCLA Phonological Segment Inventory Database (Maddieson 1984, expanded by Maddieson and Precoda 1990), which is a widely cited phoneme collection of 450 inventories and was a constituent part of the first version of PHOIBLE (see Moran 2012 for details). We also extracted a fourth global dataset from PHOIBLE, derived from three distinct collections. These are labelled in the original resource as PH, collected by Steven Moran for his 2012 thesis, UZ, material collected by Moran while working at the University of Zurich with the aim of filling genealogical gaps in the existing dataset, and GM, material collected by Christopher Green and Moran with the goal of attaining pan-Africa coverage (see Moran et al., 2014 for details). In total, these three subsets contain 899 language varieties and they are treated as a single resource here, labelled Phoible (distinct from uppercase PHOIBLE for the entire collection) in what follows. The justification for this is that they all have the same primary coder, Steven Moran, and follow the same coding principles. Other collections in PHOIBLE come from different sources using correspondingly different coding principles.

Of these other collections, we extracted five datasets with a regional focus. From the Alphabets of Africa

collection (AA: Hartell 1993; Chanard 2006) we considered inventories for 194 languages of Africa. From the database of Eurasian phonological inventories (EA: Nikolaev et al., 2015) we took inventories for 285 Eurasian languages. Within this same area, but with a narrower geographic focus, we took data for 95 languages of India (RA: Ramaswami 1999). For languages in Australia, we extracted data for 329 languages (ER: Round 2015). From the South American phonological inventory database (SAPhon: Michael et al. 2012), we took data for 339 languages of that region.

Table 1 provides a summary of the datasets used in this study. As the table shows, the datasets differ substantially in terms of their size and their number of transcribed sounds. Although offered in digital form, substantial efforts in terms of standardisation were required to make them comparable. These efforts are described in the following section.

2.2 Methods

2.2.1 Standardising phoneme inventory data

Having access to different datasets in digital form does not guarantee that they can be directly compared and this is true also for the phoneme inventory datasets we selected for our study. The main obstacles to data comparison we encountered included (1) the harmonisation of transcription systems, (2) the identification of language varieties across datasets, and (3) the identification of individual sources from which the individual phoneme inventories were derived. In order to guarantee that the original data is truthfully reflected, it is furthermore important to preserve the original format.

The most challenging part of the data standardisation was the harmonisation of transcription systems. Although all nine datasets we selected make use of the International Phonetic Alphabet (IPA), the concrete use of the IPA can differ drastically at times, ranging from differences in Unicode normalisation, via inherent ambiguities of the IPA itself, up to different interpretations of how the IPA should be used (see Moran and Cysouw 2018). Thus, while there may be only one way to write a nasalised unrounded open front vowel [ã], there are two ways to encode it on a computer, one consisting of two symbols < a > (U + 0061) and the diacritic marker <◌̃> (U + 0303), and one where there is a sole codepoint <ã> (U + 00E3). Additionally, the IPA offers two diacritics to indicate breathiness, <◌̤> (U + 0324) and <◌ʱ> (U + 02B1), which means that < dʱ> and < d̤> essentially refer to the same sound. Last but not least, the IPA does not provide a preferred order by which multiple diacritics should be combined. As a result, it is not clear if one should write a sound such as a labialised aspirated velar stop as < kʷʰ> or as < kʰʷ>, and on occasion one can encounter both variants in the same dataset.

In order to address these very practical problems with phonetic transcription systems such as the IPA, the Cross-Linguistic Transcription Systems project (CLTS, https://clts.clld.org, [si: ɛl ti: ɛs] List et al., 2021, see Anderson et al., 2018 for a description of the main goals) provides a reference catalogue for the handling of speech sounds. CLTS also offers a specific version of the IPA, labelled B(road)IPA, which directly addresses these inconsistencies and describes more than 8,000 different possible speech sounds with the help of a pragmatic feature system derived from the features inherently used in the IPA. In the BIPA transcription of the CLTS, <ã> is always represented by two codepoints, <◌̃> is the only diacritic allowed to indicate breathiness on consonants, and <◌ʷʰ> is the prescribed order of the diacritics <◌ʷ> and <◌ʰ>. CLTS does not only define these conventions but also offers code and web-based applications that help scholars to check if their data conforms to the standards defined by the reference catalogue. In addition, CLTS offers explicit mappings between the transcription

Table 1. Overview of the datasets used in this study.

Dataset	Description	Varieties	Sounds
JIPA	Illustrations of the IPA series in the Journal of the IPA	147	791
LAPSyD	Lyon-Albuquerque Phonological Systems Database	584	796
UPSID	UCLA Phonological Segment Inventory Database (via PHOIBLE)	450	847
Phoible	Core data collected for PHOIBLE.	899	1,442
AA	Alphabets of Africa (via PHOIBLE)	194	284
EA	Eurasian phonological inventories (via PHOIBLE)	362	1,218
RA	Inventories for languages of India (via PHOIBLE)	98	221
ER	Inventories for languages of Australia (via PHOIBLE)	391	104
SAPhon	South American phonological inventory basis (via PHOIBLE)	355	302

Table 2. Statistics for mapping the three data collections to the CLTS reference catalogue.

Dataset	Graphemes	CLTS Sounds	Mapped	Unmapped	Coverage	Normalised size
JIPA	957	791	938	19	0.980	0.843
LAPSyD	810	796	810	0	1.00	0.983
UPSID	919	848	918	1	0.999	0.924
Phoible	1,760	1442	1,712	48	0.973	0.842
AA	310	284	302	8	0.974	0.940
EA	1,402	1,218	1,363	39	0.972	0.894
RA	235	221	234	1	0.996	0.944
ER	105	104	105	0	1.0	0.991
SAPhon	303	302	303	0	1.0	0.997

data employed in various phoneme inventory datasets and the BIPA transcription that lies at the core of the initiative.

Despite its relative completeness, the CLTS reference catalogue could not be used directly to compare the datasets in our study, since the mapping to the three phoneme inventory collections from which the nine datasets were derived still showed too many gaps. In order to cope with this, we therefore expanded the current CLTS catalogue considerably, adding and refining mappings, and also modifying the code base which is needed to test the underlying data. These refinements were then added to the CLTS project. Table 2 provides an overview of the results of the mapping procedure. More information on the improved version of the CLTS data can be found in Appendix A1.

At this point, we extracted seven distinct datasets from the PHOIBLE collection: the global UPSID and Phoible (PH-UZ-GM) datasets and the regional AA, EA, RA, ER, and SAPhon datasets. In order to allow for a direct comparison of phoneme inventories compiled independently for the same language varieties, we used the Glottocodes (https://glottolog.org, Hammarström et al., 2022) provided in the PHOIBLE database and derived Glottocodes from the ISO codes provided along with the LAPSyD and the JIPA datasets.

The identification of common sources was more demanding since sources are often provided in different formats and styles. Data in PHOIBLE usually provides sources in BibTeX format, and for the JIPA resources, it was straightforward to retrieve the data in this form, since all articles from which the inventories were derived have their distinct Digital Object Identifier. For the LAPSyD resource, the references had to be converted to BibTeX format in a semi-automated way. Having converted a source to BibTeX does not guarantee direct comparability, since scholars may differ in the way in which articles are quoted or the author is referred to. Nevertheless, since BibTeX offers unified fields for typical roles such as Author, Year, and Journal, having access to bibliographic data in standardised form greatly eases the qualitative comparison of phoneme inventories across different datasets. Our datasets differ somewhat in their use of sources. While Phoible and JIPA have one source per inventory, describing a distinct doculect, UPSID and LAPSyD often rely on multiple sources, with inconsistencies normalised by the coder (in the latter case often quite explicitly in the associated notes).

Since in our improved CLTS datasets we were still left with some graphemes that could not be harmonised across all datasets, we selected those inventories for our study for which all graphemes could be fully represented in the improved CLTS system. From the remaining inventories, we considered only segmental phonemes (i.e. consonants and vowels; including diphthongs), ignoring tone and other suprasegmental features. Table 3 summarises the number of languages included in our datasets.

Providing this consistent mapping of language varieties and transcriptions for the nine datasets was a necessary prerequisite to directly comparing them. In order to allow for a convenient access to these mappings via software packages, we furthermore converted the JIPA and the LAPSyD data to the standard formats proposed by the Cross-Linguistic Data Formats initiative (CLDF, https://cldf.clld.org, Forkel et al., 2018). The PHOIBLE data did not need conversion, since it was already available in CLDF format. CLDF proposes standard table formats to render cross-linguistic datasets of various types, including, among others, wordlists, structural datasets, and dictionaries. Phoneme inventory data is treated as a structural dataset in which the parameters are provided in the form of features. In order to convert the data into the CLDF format, the CLDFBench package (https://github.com/cldf/cldfbench, Forkel and List 2020) was used. Once a dataset has been converted to CLDF, it

Table 3. Summary of key features of our datasets after lifting and standardisation.

Dataset	Number of inventories	With Glottocode	Mapped to CLTS	Distinct languages	Excluded inventories
JIPA	159	155	147	144	8
LAPSyD	584	584	584	584	0
UPSID	451	451	450	450	1
Phoible	932	932	899	841	33
AA	203	203	194	185	9
EA	390	390	362	285	28
RA	100	100	98	95	2
ER	392	391	391	339	0
SAPhon	355	355	355	329	0

can be conveniently used in a uniform way from within any modern programming language. Appendix A2 provides more information on all CLDF datasets used in this study along with additional explanations and information on their structure and how they can be accessed from Python.

2.2.2 Comparing phoneme inventories

By mapping all nine datasets to the standards proposed by the CLTS project, we increase the comparability of the phoneme inventories and allow for direct access to the feature system covering the sounds of the improved CLTS catalogue. We can then directly compute certain basic statistics, such as inventory sizes for vowels and consonants, but also more complex statistics based on the presence or absence of certain features.

We first computed the overall inventory sizes by counting the number of graphemes for each inventory. Inventory size is a useful, if imperfect, proxy for systemic similarity, as similar phonemic systems will tend to have a similar number of phonemes, even if the graphemes used in these may differ. Using the feature system of CLTS we also computed consonant inventory sizes and vowel inventory sizes for each inventory, and the number of long consonants, long vowels, and diphthongs in each inventory, given that varying approaches to these segments are often the cause of differences in phoneme inventories.

All other things being equal, differences in mean inventory size between datasets can thus be considered a good indicator of different coding strategies. However, datasets may differ in their underlying language sample in such a way that also has an impact on mean inventory size. In order to tease apart the effects of the language sample, we can compare the mean inventory size for a given dataset (its global mean) with the actual mean inventory size of the subset of inventories in that dataset which are directly compared to those of another dataset.

This can be illustrated with an example comparison between the LAPSyD dataset and the ER dataset, which is discussed further in 3.2.3. The mean inventory size in the 584 inventories of LAPSyD is 31.29, while that of the 391 inventories in ER is 24.04. A priori, this might point to a considerable difference in coding strategies between these two datasets. However, when we consider only the 46 inventories that can be directly compared between LAPSyD and ER, the actual mean inventory size of the 46 LAPSyD inventories is only 24.41 segments, while that of the ER inventories is 23.63. The difference in the global means between the two datasets is thus an effect of different language samples, while the coding strategies are likely to be rather similar, at least in terms of the size of inventories they produce.

There is, of course, the possibility that only looking at mean inventory size can obscure considerable variation between datasets. For this reason, we compute also the mean difference in inventory size (the 'delta') between datasets. If dataset A has 30 phonemes in its inventory for language y and 30 phonemes in that of language z, while dataset B has 20 phonemes for language y and 40 phonemes for language z, the mean inventory size is 30 for both datasets, but the mean delta is 10. The mean delta is thus a useful ancillary metric for the investigation of systemic comparability.

While the computation of inventory sizes is straightforward, it is not a very exact measure, given that two inventories can be the same size but still contain completely different sounds. It can serve as a proxy for systemic comparability but tells us nothing about graphemic comparability. A more refined measure to compare similarity is thus needed and for this we measured the Jaccard similarity (Batagelj and Bren 1995) between two phoneme inventories. This is defined as the division of the number of common elements in two sets by the number of unique elements in total. If two sets are identical, the Jaccard similarity is 1, if they

have no element in common, the Jaccard similarity is 0. Computing the Jaccard similarity is straightforward and can be done both for the graphemes originally extracted and for those graphemes mapped to the CLTS catalogue. As this metric accepts as similar only those sounds which are identical in their graphemic representations, we call it a strict similarity metric. Frequency statistics for each grapheme in each dataset are also available in the supplementary files that accompany this paper.

For both inventory size and grapheme distribution, we also use Spearman's correlation (p and r). This is a nonparametric measure used to gauge the strength and direction of association between two ranked variables. It is denoted by the correlation coefficient rr, which ranges from −1 to 1. A value of 1 signifies a perfect positive relationship, −1 indicates a perfect negative relationship, and a value close to 0 indicates no relationship. The *P*-value tests the null hypothesis that there is no correlation between the two variables. A low *P*-value (typically $P < 0.05$) rejects the null hypothesis and suggests a statistically significant correlation.

Appendix A3 illustrates how one can compute strict and approximate similarities from individual sound inventories in Python, while Appendix B1 discusses further metrics that are included in the supplementary files but not reported in the main text.

2.2.3 Inspecting phoneme inventories

To allow for convenient inspection of the phoneme inventories, we created an interactive JavaScript application that can be used in standard web browsers. The application allows users to search for language varieties by their Glottocode or by their name. Once selected, it displays all language varieties that have been assigned the same Glottocode in one of the nine datasets, along with the original graphemes and their standardised CLTS counterparts. If more than one variety can be identified, the varieties are furthermore compared by our strict inventory similarity measure which is applied both to the original graphemes and to the standardised CLTS sounds. In order to allow for a convenient qualitative inspection of the data, matching and non-matching sounds are displayed in tables, illustrating also the impact of our CLTS normalisation on the comparability of phoneme inventories. Additional information for the different language varieties sharing the same Glottocode is displayed in tabular form. Important in this context is the source information, which cannot be trivially compared automatically, but which allows human experts to quickly check to which degree inventories were drawn from the same or different sources. Figure 1 shows a screenshot of the web application. The web application has been submitted along with the supplemental material accompanying this study and can be used by unpacking the archive and opening the file index.html in a web browser.

2.2.4 Comparing phoneme inventories across datasets

In order to compare phoneme inventories for the same language variety across different datasets, we first have to identify the number of varieties that can be compared with each other for each of the nine datasets. Since we use Glottocodes to identify identical language

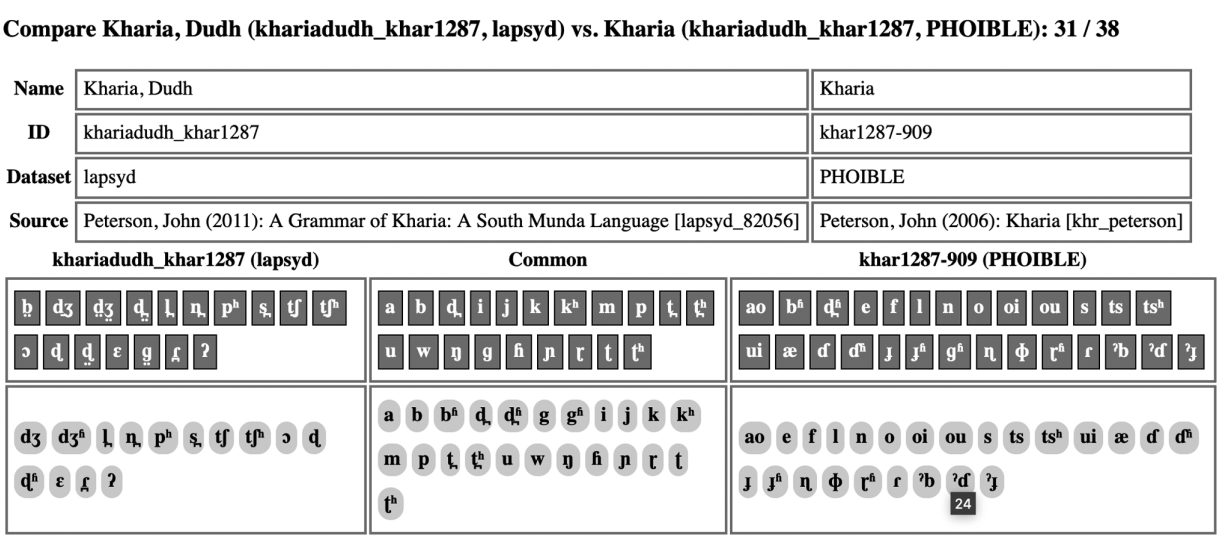

Figure 1. Interactive application for inspecting phoneme inventories across sources and datasets, showing here the LAPSyD and Phoible inventories for Kharia. The graphemes in dark grey (upper layer) are graphemes before normalisation, while those in light grey are graphemes after normalisation.

varieties across datasets, it is important to handle those cases where a given dataset has two or more inventories for the same Glottocode. When computing general systemic statistics (inventory sizes for sounds, vowels, and consonants) from the inventory datasets, we decided to aggregate the data by using the median value, rather than including all possible pairings for individual language varieties in two datasets, or excluding these data points. The former might result in an overcounting of outliers, while the latter would reduce our comparative basis. When comparing direct strict and approximate similarities among datasets, we compare all varieties corresponding to one Glottocode in one dataset with all varieties corresponding to the same Glottocode in the other dataset and then calculate the mean of the similarity score. Table 4 provides general mutual coverage statistics for the comparison of the nine datasets.

Some of the differences in mutual coverage are to be expected. There is no overlap between the regional datasets, except between EA, which targets Eurasia, and RA, which focuses on India. As the LAPSyD dataset is a development of the UPSID dataset with the same primary coder (Ian Maddieson) and draws on many of the same sources, it is no surprise that we find a high degree of mutual coverage between these datasets. On the other hand, the low mutual coverage between UPSID (and by extension LAPSyD) and Phoible is likely to derive at least in part from the stated aim of Phoible to fill gaps in the UPSID coverage. As for the high mutual coverage between Phoible and JIPA, this is likely a result of the compilers of Phoible targeting the Illustrations of the Journal of the IPA, given their high quality and accessibility as a source of phonemic inventories. Further discussion of imbalances in the coverage of the global datasets is discussed in section 3.1.

2.2.5 Implementation

The normalisation of the datasets was carried out with the help of the CLDFBench software package in individual repositories. The mapping to the BIPA transcription system of CLTS was implemented in the improved version of CLTS, accompanied by an improved software package to curate and analyse the data. The supplementary material accompanying this study contains the newly contained datasets along with the code to convert the data into CLDF packages, as well as additional code to carry out all analyses reported in this study. It has been published on Zenodo, where it is freely available to download at https://zenodo.org/records/10005821. The appendix accompanying this study contains additional information on the data and the code.

3. Results

In the following, we report results from the four global datasets in our study in 3.1, and then from the five regional datasets in 3.2. Full results are given in the supplementary material accompanying this study. The structure and contents of the supplementary material are laid out in Appendix B2.

Our study includes nine datasets, four of which have global coverage, and five of which have a regional focus. We report statistics for these datasets and compare them in terms of their systemic comparability (particularly inventory size) and graphemic comparability. The four global datasets are discussed first, in 3.1, while the five regional ones are examined in 3.2.

3.1 Global datasets

In this study, we consider four distinct datasets from published segment inventory collections with global scope: JIPA, LAPSyD, UPSID, and Phoible. We begin

Table 4. Mutual coverage statistics for the nine datasets. The values show the number of distinct language varieties per dataset and the number of Glottocode matches between them. Cells in grey show the total number of varieties covered in each dataset.

	JIPA	LAPSyD	UPSID	Phoible	AA	EA	RA	ER	SAPhon
JIPA	144	56	38	88	5	48	6	3	6
LAPSyD	56	584	303	121	24	77	25	46	97
UPSID	38	303	450	56	31	61	15	22	59
Phoible	88	121	56	839	39	70	16	11	36
AA	5	24	31	39	185	0	0	0	0
EA	48	77	61	70	0	285	30	0	0
RA	6	25	15	16	0	30	95	0	0
ER	3	46	22	11	0	0	0	339	0
SAPhon	6	97	59	36	0	0	0	0	329

by presenting results for inventory size, which is a common metric of comparison in phonological typology and is frequently used by studies that make use of segment inventory data. As a first test, we computed inventory sizes from the data and then tested the Spearman's rank correlation between all four datasets. The results of this analysis are given in Table 5.

As can be seen from the table, there are positive correlations with respect to inventory size between all datasets. In fact, the figures for the correlations show a considerable degree of uniformity, being within a very narrow range of around 0.84, except for the comparisons of UPSID with JIPA and Phoible, which are much lower than this, at 0.49 and 0.62, respectively. These lower correlations are likely to result partially from the use of different (and more modern) source material in JIPA and Phoible, but also from differences in coding strategy, evidenced by the fact that the genealogically related LAPSyD and UPSID datasets have a higher correlation.

Differences in coding strategies become clearer when we look at the mean deltas (i.e. the mean differences in inventory sizes) between the four datasets. The figures in Table 5 suggest that inventories in JIPA and Phoible are generally considerably bigger than those of LAPSyD and UPSID. This is confirmed in Table 6, which presents figures for both the global mean inventory size for each dataset and the actual mean inventory size for the comparisons between these four datasets.

There are quite considerable differences in mean inventory size for each of the four datasets. The mean inventory size in UPSID is just under three-quarters of that of that in JIPA. Further, we can observe quite surprising variations in values between the overall mean size for given datasets and actual mean inventory size in the comparisons. This is particularly clear in the comparisons with JIPA, where the three other datasets have an actual mean that is higher than their global mean (3.57 phonemes higher in the case of LAPSyD, 5.59 in the case of UPSID, and 2.89 in the case of Phoible). However, we find considerable variation also for some of the other compared pairs, such as the comparisons with UPSID, where the actual means of the JIPA and Phoible datasets are also rather higher than expected (3.98 and 4.61 phonemes respectively), and the comparisons with Phoible, where those of LAPSyD and UPSID are somewhat higher (2.26 and 2.38 phonemes, respectively).

Some of the motivation for this variation is likely to be due to differences in the underlying source material. In particular, the JIPA language sample is based on descriptions from the Journal of the International Phonetic Association, and thus not geographically well-balanced. Areal variation in mean inventory size in the descriptions underlying these datasets will affect headline mean inventory sizes. Table 7 and Figure 2 show the areal distribution of inventories in each of the four datasets across the six macroareas defined by Glottolog 4.7 (Hammarström et al., 2022): Africa, Eurasia, Australia, Papuanesia, North America, and South America.

It is clear from Table 7 that all of the global datasets are regionally skewed with respect to Glottolog 4.7 in one way or another. To an extent, this is to be

Table 5. Sample size of compared datasets, correlations in inventory size for compared inventories (all highly significant with *P*-values below 0.05), and mean difference in size between compared inventories (positive numbers indicate that the first dataset is most often larger, negative ones that the second dataset tends to be larger).

	JIPA–LAPSyD	JIPA–UPSID	JIPA–Phoible	LAPSyD–UPSID	LAPSyD–Phoible	UPSID–Phoible
Languages	56	38	88	303	121	56
Size correlation	0.84	0.49	0.84	0.85	0.84	0.62
Delta $\bar{x}$	5.32	9.05	1.82	2.14	−4.01	−8.67

Table 6. Global mean inventory size for global datasets and actual mean inventory size for comparisons. Shaded cells indicate the global mean inventory size for the dataset as a whole, while unshaded cells show actual mean size for compared languages.. Brackets indicate the number of languages under consideration.

	vs JIPA	vs LAPSyD	vs UPSID	vs Phoible
JIPA	41.63 (144)	40.31 (56)	45.61 (38)	41.84 (88)
LAPSyD	34.86 (56)	31.29 (584)	31.18 (303)	33.55 (121)
UPSID	36.55 (88)	29.03 (303)	30.96 (450)	33.34 (56)
Phoible	39.97 (88)	37.91 (121)	41.69 (56)	37.08 (839)

Table 7. Macroarea information for the four global datasets and for Glottolog 4.7. No. refers to the number of languages associated with each macroarea for each dataset, while % expresses this as a percentage of the total number of languages in that dataset.

	JIPA		LAPSyD		UPSID		Phoible		Glottolog	
	No.	%	No.	%	No.	%	No.	%	No.	%
Africa	25	17	102	17	107	24	526	59	2,367	28
Eurasia	85	58	127	22	121	27	129	14	2,004	23
Australia	3	2	52	9	25	6	12	1	388	5
Papuanesia	15	10	111	19	65	14	121	13	2,212	26
N. America	12	8	89	15	71	16	70	8	791	9
S. America	7	5	103	18	61	14	41	5	716	8
Total	147	100	584	100	450	100	899	100	8,572	100

expected, as global linguistic datasets tend to prioritise genealogical balance at least as highly as areal balance, which will mean sampling some regions more heavily than others.

That said, this consideration does not suffice to explain the degree of skewing in some of the datasets. It is clear that JIPA oversamples Eurasia with respect to other regions (58% of languages in this dataset). Similarly, the Phoible data is strongly biased towards Africa (59% of languages in this dataset). The figures for LAPSyD and UPSID are much more balanced across the different macroareas: while some areas are still comparatively overrepresented with respect to Glottolog (notably North and South America) and others are comparatively underrepresented (Papuanesia and in LAPSyD also Africa), this can be explained (at least in part) by the demands of balancing genealogical and areal sampling.

To examine the extent to which regional imbalances in language sampling contribute to differences in mean inventory size for each dataset, we compute figures for inventory size on a regional basis in Table 8, which presents mean inventory sizes for regional subsets of each dataset. This data is also visualised in Figure 3.

The figures in Table 8, visualised in Figure 3, show a clear areal signal for mean inventory size. For all four datasets, mean inventory size in Eurasia and Africa is above (and often well above) the global mean, while it is close to it in North America (a little higher in LAPSyD and UPSID, a little lower in Phoible), and well below it in South America, Australia, and Papuanesia. Further, LAPSyD, UPSID, and Phoible all agree in showing the highest mean inventory sizes for Eurasia, the next highest in Africa, followed by North America, then South America, then Australia, then Papuanesia. JIPA differs from this ranking only in having Africa above Eurasia and having South America below Australia and Papuanesia, but this dataset is both considerably smaller than the other three and is geographically heavily skewed (only three inventories from Australia and seven from South America).

The relative tendencies with respect to mean inventory size in the four global datasets under consideration hold also when the regional subsets are examined. In every region except South America, the highest mean inventory size is found in JIPA, followed by Phoible, then LAPSyD, and then UPSID. This strongly indicates that differences in mean inventory size across the four datasets result from different coding strategies and not just from areal imbalances in the underlying language samples.

One clear way in which different coding strategies are likely to influence these figures is in the treatment of different segment types. We computed the mean number of segments, consonants, and vowels, for each of the four datasets, as well as the proportions of language varieties with long consonants, long vowels, and with diphthongs. The results of this computation are presented in Table 9.

Beyond the differences with respect to mean total inventory size, these figures show considerable differences between the four datasets in the treatment of consonant and vowel length. The proportion of inventories in UPSID with long vowels is only 0.11, far lower than that of any of the three other datasets, which all have similar proportions of long vowels. Also notable is the extent to which JIPA includes long consonants (0.17 of inventories) and diphthongs (0.31 of inventories), whereas the other datasets describe these segments much more rarely. It is unsurprising that there are varying approaches to coding segments such as these, seeing as disputes over whether a certain span of sound should be analysed as one item or as two items are extremely common in phonemic analysis (see Section 1). These differences in coding strategy affect mean inventory sizes and the comparability of inventories across datasets.

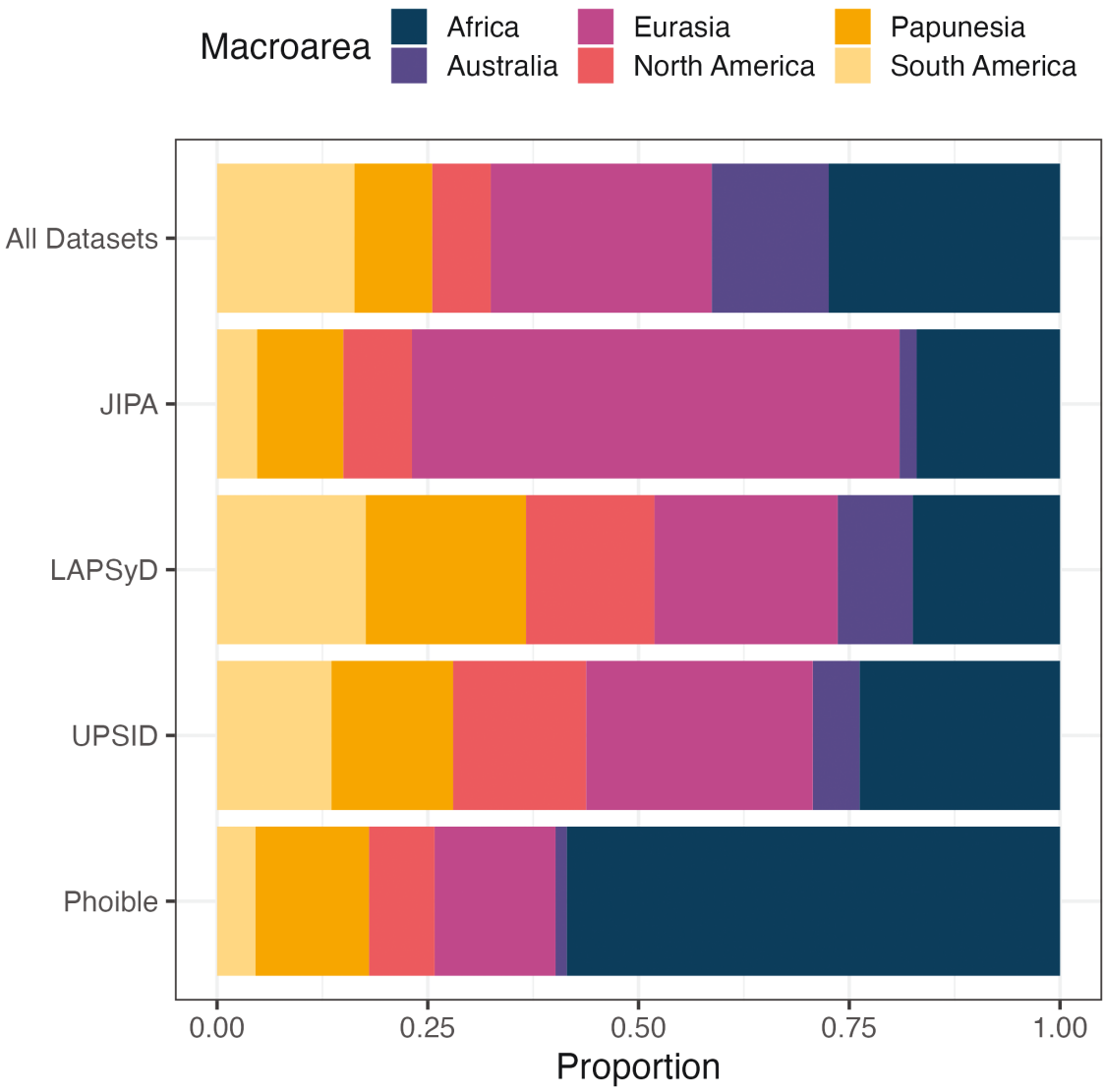

Figure 2. Proportion of inventories by Glottolog macroarea for all datasets considered in the study and for the four global datasets.

Table 8. Actual mean inventory sizes for regional subsets of each dataset and these means expressed as a proportion of global means.

	JIPA		LAPSyD		UPSID		Phoible	
	$\bar{x}$	/global	$\bar{x}$	/global	$\bar{x}$	/global	$\bar{x}$	/global
Global	41.63	1	31.29	1	30.96	1	37.08	1
Africa	50.4	1.21	36.22	1.16	34.35	1.11	40.21	1.08
Eurasia	43.01	1.03	38.01	1.21	36.4	1.18	41.59	1.12
Australia	33.67	0.81	24.27	0.78	22.16	0.72	25.42	0.69
Papuanesia	27.53	0.66	24.18	0.77	21.86	0.71	24.74	0.67
North America	42.25	1.01	32.89	1.05	31.82	1.03	34.54	0.93
South America	26	0.62	27.94	0.89	26.52	0.86	26.90	0.73

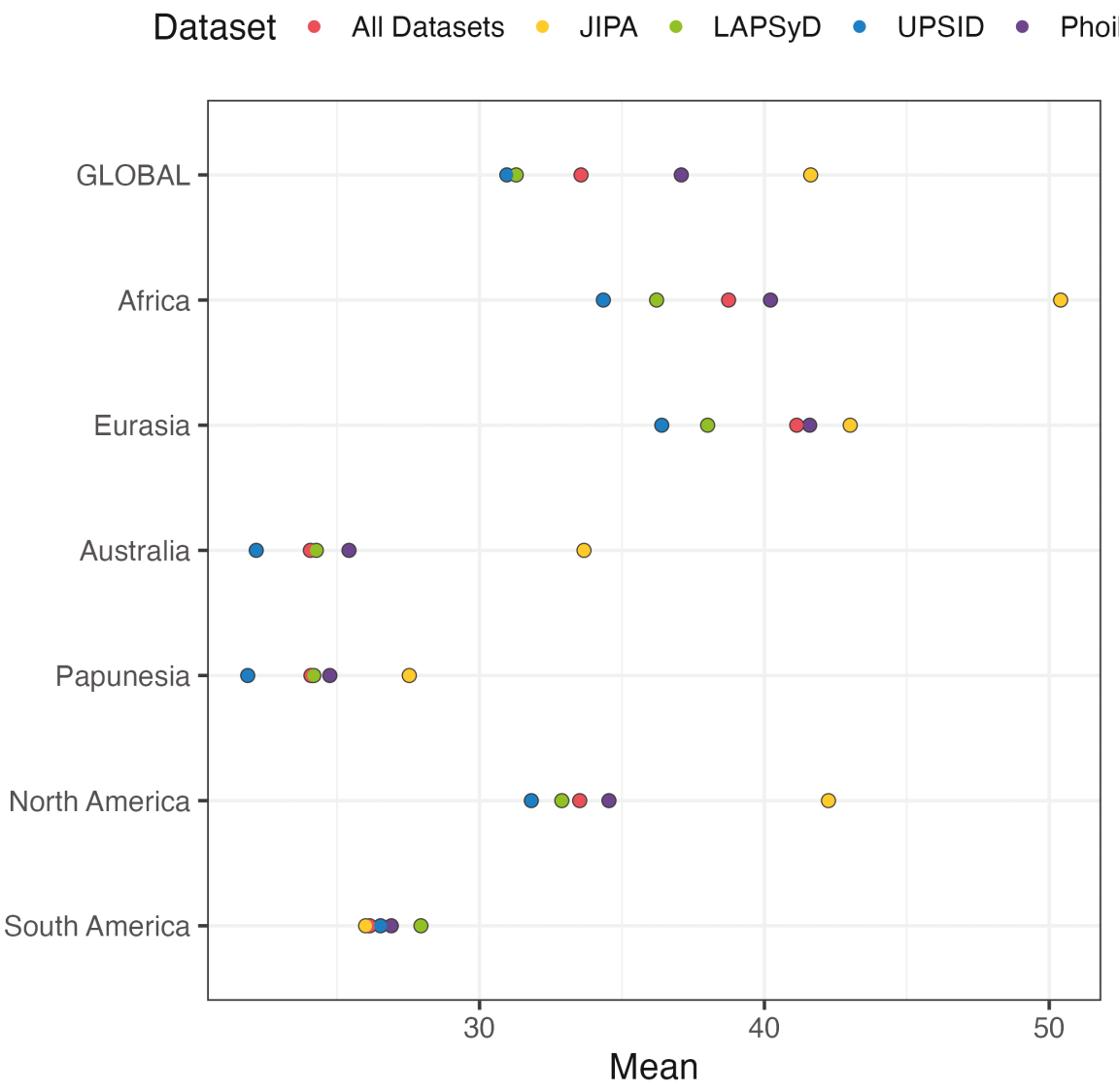

Figure 3. Mean inventory size in each of the Glottolog macroareas for each of the four global datasets.

Of course, different coding strategies go beyond inventory size but also involve differences in the analysis of certain segment types. We can thus expect to observe differences in what graphemes are used to represent the segments of an inventory. To investigate this aspect of coding strategy in the global datasets, we computed the mean strict similarity scores for all segments in the comparisons of the four global datasets under investigation. We also computed the Spearman's rank correlation for grapheme distribution. These figures are provided in Table 10.

The figures in Table 10 show considerable differences in how segment inventories for supposedly identical language varieties have been coded. We find a similar trend with respect to those pairings that showed low correlations for inventory size in Table 5, particularly reflected in the low similarity scores and correlations for the comparisons of UPSID with JIPA and Phoible. In general, the correlations here are much lower than for inventory size, suggesting substantial differences in the graphemes used in the different datasets. Among the comparisons, the relatively higher correlation between LAPSyD and UPSID is likely to stem from their close genealogical relationship.

Differences in coding strategy with respect to graphemic similarity are discussed further in Section 4, but our attention now turns to results from the regional datasets.

Table 9. Mean numbers of segments, consonants, and vowels per inventory in each of the four global datasets, as well as the proportions of language varieties containing long consonants, long vowels, and diphthongs.

	Mean sizes			Proportions		
	Segments	Consonants	Vowels	Long C	Long V	Diphthongs
JIPA	41.63	28.63	11.06	0.17	0.42	0.31
LAPSyD	31.29	21.42	9.34	0.01	0.4	0.12
UPSID	30.96	22.43	8.06	0.03	0.11	0.11
Phoible	37.08	25.93	10.4	0.08	0.39	0.12

Table 10. Strict similarity scores and Spearman's rank correlation for all segments in compared inventories in the four global datasets.

	JIPA–LAPSyD	JIPA–UPSID	JIPA–Phoible	LAPSyD–UPSID	LAPSyD–Phoible	UPSID–Phoible
Strict similarity	0.64	0.42	0.68	0.67	0.62	0.48
Grapheme correlation	0.45	0.19	0.49	0.59	0.5	0.27

Table 11. Summary statistics for AA dataset and comparisons with four global datasets. Correlations which are not highly statistically significant (with a *P*-value below 0.05) are put in parentheses. For the delta $\bar{x}$ positive values indicate that AA tends to be higher, negative values that rather the compared dataset tends to be higher.

	AA	vs JIPA	vs LAPSyD	vs UPSID	vs Phoible
Languages	194	5	24	31	39
Size correlation	1	(0.82)	0.55	(0.33)	0.6
Delta $\bar{x}$	0	−4.3	−1.7	1.15	−3.57
Global $\bar{x}$	37.05	41.63	31.29	30.96	37.08
Actual $\bar{x}$ vs AA	37.05	40.6	35.21	33.84	38.31
Strict similarity	1	0.69	0.72	0.63	0.66
Char. correlation	1	0.72	0.73	0.48	0.59

3.2 Regional datasets

Our study includes five regional datasets, covering four of the macroareas defined by Glottolog 4.7. The AA dataset covers languages of Africa, the EA dataset of Eurasia, the RA dataset of the Indian subset of Eurasian languages, the ER dataset of the languages of Australia, and the SAPhon those of South America. In the following sections, we present results for each of these datasets in turn.

3.2.1 Alphabets of Africa (AA)

Table 11 presents summary results for the AA dataset and comparisons between it and the four global datasets.

While the AA dataset has 194 inventories, the number of inventories which can be directly compared to those in the global datasets is rather low. The correlations between AA and the global global datasets are highly significant only in the case of Phoible, so must be viewed with some caution. The mean differences in inventory size between AA and each of the global datasets, between 1.15 and 4.3 segments, are broadly comparable to those that hold among JIPA, LAPSyD and Phoible.

Mean inventory size in AA is intermediate between that of LAPSyD and UPSID on the one hand and that of Phoible and JIPA on the other. Except for JIPA, which shares only five inventories with AA, actual mean inventory sizes of the subsets of these datasets that can be compared to AA are rather higher than their global means (3.92 for LAPSyD, 2.88 for UPSID, and 1.23 for Phoible). This gives further, if limited, support to the generalisation that the mean inventory size

of languages in the Africa macroarea is higher than the global norm.

As for the strict similarity scores, these are broadly similar to those that hold between the global datasets, while the correlations of AA inventories with those of the global datasets are rather high when compared to the equivalent figures between the global datasets. This may stem from the fact that these are in origin alphabets, and consequently the graphemes that appear in AA tend not to have a high degree of phonetic specification and have a relatively high overall frequency.

3.2.2 Eurasian phonological inventories (EA)

The EA dataset covers the whole of Eurasia and also includes nine inventories from the Glottolog macroarea of Papuanesia. Summary statistics for this dataset are presented in Table 12.

A first observation concerns the mean size of these inventories. For every comparison with EA, the actual mean inventory size of each of the global datasets in comparison to EA is higher, often considerably higher, than the global means for these datasets. This lends additional support to the contention that the languages of Eurasia, as a whole, tend to have inventories larger than the global norm (see Table 8).

As for mean inventory size in EA itself, these figures show that the inventories in EA tend to be somewhat smaller than the JIPA ones, but larger than those of the other datasets. This tendency is relatively consistent, especially with LAPSyD, where the size correlation is quite high (0.87), with the EA inventories being on average larger than the LAPSyD ones.

The correlations in grapheme distribution are, on the other hand, considerably lower than those for inventory size and the strict similarity scores are also rather low. Whatever the size of the inventories coded in EA, it seems that the actual graphemes used to code these inventories are often quite different from those used in the other datasets under consideration. Differences in coding long consonants, long vowels, and diphthongs do not appear to drive this variation (figures can be seen in the supplementary data files). A more likely explanation lies in the high degree of phonetic specification used in coding this dataset. EA uses 1218 grapheme types to code the inventories of 285 languages in Eurasia, while LAPSyD only uses 796 types for 584 languages from across the world. Inevitably, this reduces graphemic similarity in the comparisons between EA and other datasets.

3.2.3 Languages of India (RA)

The second dataset focused on the Eurasia macroarea is RA, which concentrates on the languages of India. This means that as well as comparisons with the global datasets, a comparison with EA is also possible for this dataset. Summary statistics for RA are given in Table 13.

In general, the size correlation between RA and the global datasets is rather high. The mean size of inventory in RA tends to be larger inventories than in the global datasets in the compared pairs. Mean inventory sizes in the comparisons of the global datasets with RA are similar to the regional means of these datasets given in Table 8, suggesting that mean inventory size in India is in line with that of the Eurasia macroarea as a whole.

The strict similarity scores and correlations in grapheme distribution between RA and other datasets are quite low. Part of the explanation for this may lie in the high proportion of long vowels (0.69) in this dataset. However, it should also be considered that RA is focused on a well-defined region whose languages tend to share phonological features, and in coding these languages, a quite restricted set of 98 grapheme types is used. This makes the languages within this dataset more comparable among themselves, but somewhat

Table 12. Summary statistics for EA dataset and comparisons with four global datasets. Correlations which are not highly statistically significant (with a *P*-value below 0.05) are put in parentheses. For the delta $\bar{x}$ positive values indicate that EA tends to be higher, negative values that rather the compared dataset tends to be higher.

	EA	vs JIPA	vs LAPSyD	vs UPSID	vs Phoible
Languages	285	48	77	61	70
Size correlation	1	0.71	0.87	0.68	0.69
Delta $\bar{x}$	0	−2.56	2.88	2.84	0.74
Global $\bar{x}$	42.53	41.63	31.29	30.96	37.08
Actual $\bar{x}$ vs EA	42.53	45.48	37.21	36.57	39.68
Strict similarity	1	0.55	0.49	0.39	0.49
Char. correlation	1	0.35	0.48	0.39	0.25

Table 13. Summary statistics for RA dataset and comparisons with four global datasets. Correlations which are not highly statistically significant (with a *P*-value below 0.05) are put in parentheses. For the delta $\bar{x}$ positive values indicate that RA tends to be higher, negative values that rather the compared dataset tends to be higher.

	RA	vs JIPA	vs LAPSyD	vs UPSID	vs Phoible	vs EA
Languages	95	6	25	15	16	30
Size correlation	1	(0.66)	0.8	0.82	0.84	0.72
Delta $\bar{x}$	0	–11	3.73	8.19	1.19	3.06
Global $\bar{x}$	41.98	41.63	31.29	30.96	37.08	42.53
Actual $\bar{x}$ vs RA	41.98	60.5	39.52	36.33	43.56	42.16
Strict similarity	1	0.39	0.46	0.44	0.44	0.46
Char. correlation	1	0.25	0.52	0.57	0.4	0.43

less directly comparable to the global datasets, where the comparative view is much broader.

3.2.4 Languages of Australia (ER)

The ER dataset focuses on languages of Australia. Summary statistics for ER and its comparison with the global datasets are given in Table 14.

The correlation in size among ER, LAPSyD, and Phoible is quite high, while the figures with UPSID are relatively lower. On balance, however, the mean size of the ER inventories is quite similar to that of the global datasets (except for JIPA, but the comparison in this case includes only three languages).

Interesting, however, is the fact that this overall similarity in inventory size is not matched by an overall similarity in the graphemes used. The strict similarity scores between ER and the global datasets are not particularly high and the correlations in grapheme distribution are very low. Part of the explanation for this lies in the relatively restricted character set used in ER: only 104 distinct grapheme types encode inventories for 395 languages. In contrast, SAPhon uses almost three times as many distinct graphemes (302) to encode inventories for fewer languages (339). The comparison with EA, which uses 1218 graphemes to encode inventories for 285 languages, is even more striking.

Round (2019) justifies the small inventory set in ER, by noting that 'cross-doculectal variation in phoneme labels may reflect more a variation among analysts' necessary choices between multiple, defensible options, than empirical differences among languages'. Indeed, this is likely to be true more broadly, but the languages of Australia are known to share a number of common features (e.g. Butcher 2012), diverging in certain respects from the global norm. The ER dataset is designed to facilitate comparison between the phonologies of the languages of this region, not broader global comparisons, and thus for some segment types, it uses conventions that reflect contrasts frequently found in Australian languages, but not elsewhere. This is particularly striking for coronal places of articulation, where Australian languages 'contrast either one or two "apical" places of articulation, where the primary constriction is formed with the tongue tip; and either one or two "laminal" places of articulation, with constriction formed by the tongue blade' (Round 2019). Characters reflecting these contrasts have a high relative frequency in ER: laminal prepalatal stops or sonorants are represented in ER with a palatal curl (i.e. the base characters <t̠ n̠ l̠>) and comprise over 11% of all tokens in the dataset, but only 0.1% of tokens in the rest of the inventories combined. The apical diacritic, used when there is no alveolar-retroflex contrast in apical stops, occurs in 6.6% of tokens in ER, but in only 0.8% of tokens elsewhere.

3.2.5 South American phonological inventory database (SAPhon)

The SAPhon dataset focuses on the phonologies of languages of South America. Table 15 presents summary statistics for this dataset.

Overall, the correlations in inventory size between SAPhon and the global datasets are moderately high, and the mean difference in inventory size between SAPhon and these are rather low. In fact, mean inventory sizes are also very consistent between SAPhon and the global datasets, within a narrow range of 24.47–26.6. This tallies with the mean inventory sizes for the South American inventories of the global datasets, given in Table 8, which range from 26 to 27.94. All the evidence thus points towards mean inventory size in South America being considerably lower than the global average.

Graphemic similarity between SAPhon and the global datasets is quite reasonable when compared to the figures for most of the other regional datasets. This

Table 14. Summary statistics for ER dataset and comparisons with four global datasets. Correlations which are not highly statistically significant (with a *P*-value below 0.05) are put in parentheses. For the delta $\bar{x}$ positive values indicate that ER tends to be higher, negative values that rather the compared dataset tends to be higher.

	ER	vs JIPA	vs LAPSyD	vs UPSID	vs Phoible
Languages	329	3	46	22	11
Size correlation	1	(0.5)	0.83	0.59	0.87
Delta $\bar{x}$	0	–3	–0.64	0.73	2.83
Global $\bar{x}$	24.04	41.63	31.29	30.96	37.08
Actual $\bar{x}$ vs ER	24.04	33.67	24.41	22.27	25.55
Strict similarity	1	0.61	0.5	0.41	0.41
Char. correlation	1	0.53	0.18	–0.03	0.09

Table 15. Summary statistics for SAPhon dataset and comparisons with four global datasets. Correlations which are not highly statistically significant (with a *P*-value below 0.05) are put in parentheses. For the delta $\bar{x}$ positive values indicate that SAPhon tends to be higher, negative values that rather the compared dataset tends to be higher.

	SAPhon	vs JIPA	vs LAPSyD	vs UPSID	vs Phoible
Languages	339	6	97	59	36
Size correlation	1	0.97	0.73	0.66	0.7
Delta $\bar{x}$	0	–0.83	–0.86	0.73	–0.48
Global $\bar{x}$	25.47	41.63	31.29	30.96	37.08
Actual $\bar{x}$ vs SAPhon	25.47	25.5	26.6	26.6	26.4
Strict similarity	1	0.83	0.63	0.52	0.72
Char. correlation	1	0.78	0.67	0.4	0.69

is likely due to the fact that SAPhon uses 302 discrete grapheme types for the inventories of 339 languages, which is a middle way between the highly diverse strategy of EA (1218 discrete graphemes for 285 languages) and the highly constrained strategy of ER (104 discrete grapheme types for 329 languages).

4. Discussion

Our study compared inventories along two different axes of comparability. The first axis is systemic comparability, the extent to which two different inventories reflect the same underlying analysis of a given language. We used mean inventory size as a proxy for this, with mean difference in inventory size as an ancillary measure. The second axis is graphemic comparability, the extent to which the same graphemes are used to represent equivalent 'points in the pattern', where we relied primarily on a strict similarity measure based on the Jaccard index. Our results show sometimes quite striking variation on both axes.

For the common metric of mean inventory size, the datasets under consideration here show a very clear regional pattern. Mean inventory size is highest in Eurasia, followed closely by Africa, around the global mean in North America, and well below it in South America, Australia, and Papuanesia. While it is possible that some of this variation stems from different descriptive practices in the different regions, it is highly likely that much of it also stems from actual linguistic differences.

In spite of this clear regional pattern, regionally biased language samples do not explain differences in mean inventory size between datasets. When we take the actual mean inventory size in pairwise comparisons between datasets, considerable differences remain. Among the global datasets, JIPA consistently has larger inventories than Phoible, which in turn has larger inventories than LAPSyD and UPSID.

Some of the causes for this variation are explored in section C of the appendix. In some cases, variation results from the simple fact that inventories have

been coded on the basis of the same source material (Appendix C1). On the one hand, it is not infrequent to find differences in inventories coded from the same source material (Appendix C2), reflecting the varying conventions, assumptions, and interpretative frameworks of different coders. Particularly frequent are differences in whether vowel and consonant length are considered phonemic (Appendix C3).

In general, the results for mean inventory size are somewhat more similar than those for graphemic comparability, where we find sometimes quite considerable differences between datasets. One reason for this is the degree of phonetic specification used by different coders. This is explored for the global datasets in Appendix C4, but is quite striking for some of the regional datasets. EA, SAPhon, and ER are quite similar in terms of the number of languages they cover: 285, 329, and 339, respectively. However, they differ drastically in the number of grapheme types they use to encode these languages: EA uses 1218, SAPhon 302, and ER only 104. These differences cannot be explained by the difference in mean inventory size, but rather reflect quite radically different approaches to inventory coding between the three datasets.

In all, our results call into question the reliability of the results of a considerable body of work based on phoneme inventory data. It is unclear how well some of the claims which have been made about phoneme inventory size and other variables hold up given the variation which we have found in this study. An obvious target of further research is to test some of the existing claims in the literature against principled subsets of the data we have assembled here, or controlling for the variation that we have observed in this study.

5. Conclusion

Scholarly work has used phoneme inventories as evidence for studies across a wide range of topics, with an implicit assumption that they constitute a reliable and robust source of data. However, our results sound a note of caution for these studies. We tested inventories for the same languages in different datasets and found a high degree of variation. While some of this variation reflects differences in the source material used to draw up these inventories, much of it is driven by differences in the interpretation of these sources and by the different coding policies used in the collections under investigation.

We believe that our study points to serious ramifications for all secondary work that is based on phoneme inventories. Different analyses of the phonology of any given language are not only possible, but common, and different choices in what is considered phonemic, and what graphemes are used to represent these make inventory comparison a difficult enterprise. Inventory size varies considerably between different inventories for the same language, making this a dubious statistic for cross-linguistic comparison. Many studies refer to the presence or absence of certain phonemes across different languages, but the low scores for strict similarity we found mean that the robustness of inferences based on presence or absence of given graphemes must be questioned. The question of the comparability of phoneme inventories needs to be taken seriously and more consideration given to how these can be rendered more comparable in a principled way. Our study provides tools that would allow existing and new studies to be tested against a variety of datasets.

On this last point, we feel capable of offering some words of hope. As our results show, the various normalisation attempts that we have carried out, specifically those based on our enhanced version of the Cross-Linguistic Transcription Systems originally proposed by Anderson et al. (2018) have been shown to play a very significant role in increasing the comparability of inventories in our study. We hope that this study may help to raise awareness among scholars of some of the problems of using aggregated phoneme inventory data without reflection on the theoretical and practical issues surrounding its comparability.

Acknowledgements

The authors would like to thank three anonymous reviewers and Erich Round for feedback on this paper.

Data availability

The appendix to this study is available at *Journal of Language Evolution Journal* online. This study is accompanied by supplementary materials published on Zenodo and freely available to download at https://zenodo.org/records/10005821. This repository includes the data used in this study as well as the code needed to replicate the experiments carried out and instructions how to do so. Also included as a file in this repository is the appendix referred to in the text (Appendix.pdf).

Funding

J.M.L. was supported by the ERC Consolidator Grant "ProduSemy" (Grant No. 101044282, see https://doi.org/10.3030/101044282) and the MPG Research Grant "CALC³" (https://calc.digling.org). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union

or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

Conflict of interest statement

The authors declare no competing interests.

Notes

1. An example from our data illustrating overanalysis and underanalysis comes from the treatment of the Journal of the International Phonetic Association descriptions of Lizu and Ersu (Chirkova and Chen 2013; Chirkova et al., 2015) in Baird et al. (2021). Prenasalised segments are overanalysed in Lizu, being treated as clusters of/N/ plus obstruent, while in Ersu they are underanalysed, being treated as unit phonemes,/Npʰ, Nb, Ntʰ etc./. While these languages are closely related and have similar phonologies, the result of these analytical choices means that Lizu has 48 phonemes and Ersu 62.

References

Anderson, C. et al. (2018) 'A Cross-Linguistic Database of Phonetic Transcription Systems', *Yearbook of the Poznań Linguistic Meeting*, 4: 21–53. https://doi.org/10.2478/yplm-2018-0002

Anderson, S. R. (1985) *Phonology in the Twentieth Century*. Chicago: University of Chicago.

Atkinson, Q. D. (2011) 'Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa', *Science*, 332: 346–49. https://doi.org/10.1126/science.1199295

Baird, L., Evans, N., and Greenhill, S. J. (2021) 'Blowing in the Wind: Using North Wind and Sun Texts to Sample Phoneme Inventories', *Journal of the IPA*, 52: 453–494. https://doi.org/10.1017/S002510032000033X

Batagelj, V., and Bren, M. (1995) 'Comparing Resemblance Measures', *Journal of Classification*, 12: 73–90.

Baudouin de Courtenay, J. (1894) 1972. 'Próba Teorii Alternacji Fonetycznych', in E. Stankiewicz (ed) *A Baudouin de Courtenay Anthology. The Beginnings of Structural Linguistics*, Bloomington: Indiana University Press.

Blasi, D. E. et al. (2019) 'Human Sound Systems Are Shaped by Post-Neolithic Changes in Bite Configuration', *Science*, 363: 1–10. https://doi.org/10.1126/science.aav3218

Bloomfield, L. (1933) *1973. Language*. London: Allen & Unwin.

Butcher, A. (2012) 'On the Phonetics of Long, Thin Phonologies', in *Quantitative Approaches to Problems in Linguistics*, pp. 133–154. LINCOM EUROPA.

Ceolin, A. et al. (2020) 'Formal Syntax and Deep History', *Frontiers in Psychology*, 11: 2384. https://doi.org/10.3389/fpsyg.2020.488871

Chanard, C. (2006). *Systèmes Alphabétiques Des Langues Africaines*.

Chao, Y. R. (1934) 'The Non-Uniqueness of Phonemic Solutions of Phonetic Systems', *Bulletin of the Institute of History and Philology*, 4: 363–97.

Chirkova, K. et al. (2015) 'Ersu', *Journal of the International Phonetic Association*, 45: 187–211. https://doi.org/10.1017/s0025100314000437

Chirkova, K., and Chen, Y. (2013) 'Lizu', *Journal of the International Phonetic Association*, 43: 75–86. https://doi.org/10.1017/s0025100312000242

Coupé, C., Marsico E., and Pellegrino F. (2009) 'Structural Complexity of Phonological Systems', in *Pellegrino, Francesco, Egidio Marsico, Ioana Chitoran and Christophe Coupé eds. Approaches to Phonological Complexity*, pp. 141–70. Berlin: De Gruyter Mouton.

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

Crothers, J. H. (1978) 'Typology and Universals of Vowel Systems', in J. H. Greenberg, C. A. Ferguson, E. A. Moravcsik (eds.) *Universals of Human Language*, Vol. 2: *Phonology*, pp. 93–152. Stanford: Stanford University Press.

Crothers, J. H. et al. (1979). *Handbook of Phonological Data from a Sample of the World's Languages. A Report of the Stanford Phonology Archive*. Stanford: Department of Linguistics, Stanford University.

Dediu, D., and Moisik, S. (2019) 'Pushes and Pulls from Below: Anatomical Variation, Articulation and Sound Change', *Glossa*, 4: 1–33.

DeMille, M. M. C. et al. (2018). 'Worldwide distribution of the DCDC2 READ1 regulatory element and its relationship with phoneme variation across languages', *Proceedings of the National Academy of Sciences United States of America*, 115: 4951–56. https://doi.org/10.1073/pnas.1710472115. Erratum in: Proc. Natl. Acad. Sci. U.S.A. 2018 May 21.

Deutscher, G. (2009) '"Overall Complexity": A Wild Goose Chase?' in G. Sampson, D. Gil, and P. Trudgill (eds) *Language Complexity as an Evolving Variable*, 243–52. Oxford University Press.

Donohue, M. et al. (2013) *World Phonotactics Database*. Canberra: Department of Linguistics. The Australian National University.

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.

Dryer, M. S., and Haspelmath, M., eds. (2013) *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Easterday, S. (2019) *Highly complex syllable structure: A typological and diachronic study*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.3268721.

Everett, C. (2017) Languages in Drier Climates Use Fewer Vowels. *Frontiers in Psychology* 8. https://doi.org/10.3389/fpsyg.2017.01285.

Everett, C., Blasi, D. E., and Roberts, S. G. (2015) 'Climate, Vocal Folds, and Tonal Languages: Connecting the Physiological and Geographic Dots', *Proceedings of the National Academy of Sciences*, 112: 1322–27. https://doi.org/10.1073/pnas.1417413112

Everett, C., and Chen, S. (2021). 'Speech Adapts to Differences in Dentition Within and Across Populations', *Scientific Reports*, 11: 1–10. https://doi.org/10.1038/s41598-020-80190-8

Forkel, R. et al. (2018) 'Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics', *Scientific Data*, 5: 1–10. https://doi.org/10.1038/sdata.2018.205

Forkel, R., and List, J-.M. (2020) *'CLDFBench. Give Your Cross-Linguistic Data a Lift'*. In Proceedings of the Twelfth International Conference on Language Resources and Evaluation, 6997–7004. Luxembourg: European Language Resources Association (ELRA).

Georgiou, G. P., and Kilani, A. (2020) 'The Use of Aspirated Consonants During Speech May Increase the Transmission of Covid-19', *Medical Hypotheses*, 144: 109937. https://doi.org/10.1016/j.mehy.2020.109937

Halle, M. (1963) 'Phonemics'. In Soviet and East European Linguistics, in T. Sebeok (ed.) *Current Trends in Linguistics 1*, pp. 5–21. Amsterdam; New York: De Gruyter Mouton. https://doi.org/10.1515/9783110814620

Hammarström, H. Forkel, R., Haspelmath, M., and Sebastian Bank. (2022) *Glottolog. Version 4.7*. Jena: Max Planck Institute for the Science of Human History. https://glottolog.org.

Harris, Z. (1951) *Structural Linguistics*. Chicago: Phoenix.

Hartell, R. L. (ed.) (1993) *Alphabets des langues africaines*. Dakar: UNESCO and Société Internationale de Linguistique.

Hjelmslev, L. (1943) *Omkring Sprogteoriens Grundlæggelse*. København: Akademisk forlag.

International Phonetic Association. (1999) *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Johansson, N. E. et al. (2020) 'The Typology of Sound Symbolism: Defining Macro-Concepts via Their Semantic and Phonetic Features', *Linguistic Typology*, 24: 253–310. https://doi.org/10.1515/lingty-2020-2034

Jones, D. (1950) *The Phoneme, Its Nature and Use*. Cambridge: Heffer.

Kiparsky, P. (2018) 'Formal and Empirical Issues in Phonological Typology', in L. M. Hyman and F. Plank (eds) *Phonological Typology*. Berlin: De Gruyter Mouton. 54–106.

Levinson, S., and Evans, N. (2010) 'Time for a Sea-Change in Linguistics', *Response to Comments on "the Myth of Language Universals"*, 120: 2733–58. https://doi.org/10.1016/j.lingua.2010.08.001

List, J. -M. (2019) 'Beyond Edit Distances: Comparing Linguistic Reconstruction Systems', *Theoretical Linguistics*, 45: 247–58. https://doi.org/10.1515/tl-2019-0016

List, J. -M. et al. (2021) *Cross-Linguistic Transcription Systems. Version 2.2.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5583682

Maddieson, I. (1984) *Patterns of Sounds*. Cambridge; New York: Cambridge University Press.

Maddieson, I. et al. (2013) *'LAPSyD: Lyon-Albuquerque Phonological Systems Database'*. Proceedings of the14th Interspeech Conference, Lyon, France, 25–29 August 2013.

Maddieson, I. (2016) 'Word Length Is (in Part) Predicted by Phoneme Inventory Size and Syllable Structure', *Journal of the Acoustical Society of America*, 139: 2218. https://doi.org/10.1121/1.4950645

Maddieson, I., and Coupé, C. (2015) 'Human Spoken Language Diversity and the Acoustic Adaptation Hypothesis', *Journal of the Acoustical Society of America Am*, 138: 1838–38. https://doi.org/10.1121/1.4933848

Maddieson, I., and Precoda, K. (1990) *'Updating UPSID'. UCLA Working Papers in Phonetics,* pp. 104–111. UCLA : Department of Linguistics. https://doi.org/10.1121/1.2027403

Michael, L., Stark, T., and Chang, W. (2012) *South American Phonological Inventory Database*. University of California. http://linguistics.berkeley.edu/saphon/en/

Moran, S. (2012). 'Phonetics Information Base and Lexicon'. *PhD dissertation*, University of Washington.

Moran, S., and Cysouw, M. (2018) *The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles*. Berlin: Language Science Press. http://langsci-press.org/catalog/book/176

Moran, S., McCloy, D., and Wright, R. (2012) 'Revisiting Population Size vs Phoneme Inventory Size', *Language*, 88: 877–893. https://doi.org/10.1353/lan.2012.0087

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

Moran, S., McCloy, D., and Wright, R., eds. (2014) *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://phoible.org/.

Nikolaev, D., Nikulin, A., and Kukhto, A. (2015) *The Database of Eurasian Phonological Inventories*.

Ohala, J. J. (2009). 'Languages' Sound Inventories: The Devil in the Details', in F. Pellegrino, E Marsico, I. Chitoran and C. Coupé (eds) *Approaches to Phonological Complexity*, pp. 47–58. Berlin: De Gruyter Mouton.

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: 376. https://doi.org/10.1515/lity.2004.8.3.376

Ramaswami, N. (1999) *Common Linguistic Features in Indian Languages: Phonetics*. Mysore: Central Institute of Indian Languages.

Reformatsky, A. (1970) *Iz Istorii Otečestvennoj Fonologii [on the History of Russian Phonology]*. Moscow: Nauka.

Round, E. (2015) *Phonemic Inventories of Australia*. https://doi.org/10.5281/zenodo.3464333

Round, E. (2019) *Australian Phonemic Inventories Contributed to Phoible 2.0: Essential Explanatory Notes*. https://doi.org/10.5281/zenodo.3464333

Sapir, E. (1925) 'Sound Patterns in Language', *Language*, 1: 37–51. https://doi.org/10.2307/409004

Sapir, E. (1933) 'La Réalité Psychologique Des Phonèmes', *Journal de Psychologie Normale et Pathologique*, 30: 247–65.

Saussure, F. de. (1916) Cours de Linguistique Générale. Paris: Payot.

Simpson, A. P. (1999) *'Fundamental problems in comparative phonetics and phonology: does UPSID help to solve them'*. Proceedings of the 14th International Congress of Phonetic Sciences. Vol. 1. Berlin: De Gruyter.

Skirgård, H. et al. (2023) Grambank Reveals the Importance of Genealogical Constraints on Linguistic Diversity and Highlights the Impact of Language Loss, *Science Advances*, 9: 1-15. https://doi.org/10.1126/sciadv.adg6175

Trubetzkoy, N. S. (1939) *Grundzüge der Phonologie*. Prague: Travaux du Cercle Linguistique de Prague 7.

Trudgill, P. (2004) 'Linguistic and Social Typology: The Austronesian Migrations and Phoneme Inventories', *Linguistic Typology*, 8: 305–320. https://doi.org/10.1515/lity.2004.8.3.305

Twaddell, W. F. (1935) 'On Defining the Phoneme', *Language*, 11: 5–62.

Vaux, B. (2009) 'The Role of Features in a Symbolic Theory of Phonology', in E. Raimy and C. E. Cairns (eds) *Contemporary Views on Architecture and Representations in Phonology*, Cambridge, Ma: MIT Press. pp. 75–97.