# Measuring variation in phoneme inventories

**Cormac Anderson** ( cormacanderson@gmail.com )
Max Planck Institute for Evolutionary Anthropology, Leipzig
**Tiago Tresoldi**
Uppsala University
**Simon J. Greenhill**
Max Planck Institute for Evolutionary Anthropology, Leipzig
**Robert Forkel**
Max Planck Institute for Evolutionary Anthropology, Leipzig
**Russell D. Gray**
Max Planck Institute for Evolutionary Anthropology, Leipzig
**Johann-Mattis List**
Max Planck Institute for Evolutionary Anthropology, Leipzig

**Research Article**

**Keywords:** phoneme, inventory, phonology, cross-linguistic, phonological typology

**Posted Date:** September 10th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-891645/v1

# Abstract

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 four popular phoneme inventory collections. We render them comparable by linking them to standardised formats for the handling of cross-linguistic datasets and develop new measures to test both size and similarity. 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 predic, 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.


# 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 (Levison 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).

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; Easterday 2019; Dediu and Moisik 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 (Trudgill 2004, Atkinson 2011; contra Periclev 2004, Donohue and Nichols 2011; Moran et al. 2012), subsistence strategies (e.g. Blasi et al. 2019; Everett and Chen 2021), and ecology (Maddieson and Coupé 2015; Everett et al. 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 non-trivial way. These differences of 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.

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; Hjelmslev 1943; Trubetzkoy 1939), 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]

A prerequisite for the comparison of phoneme inventories across different language varieties is some certainty 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 four 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 varies 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, which can easily reach more than 1000 or 2000 (see the overview in Anderson et al. 2018). Firstly we use a *strict similarity* measure based on the Jaccard similarity (see Batagelj and Bren 1995), treating inventories as *sets of*

*graphemes*. Secondly, we use an *approximate similarity measure*, which aims to capture cases in which it is likely that a high degree of system comparability has been obscured by different transcriptional practice, e.g. the use of symbols for alveolar stops /t, d/ and dental stops /t̪, d̪/ for functionally equivalent phonemes. This measure determines approximate matches between the sounds in two inventories and assigns them distinct similarity scores derived from their feature values.

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.

[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, /Np̪, Nb, Nt̪ 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.

## 2 Materials And Methods

# 2.1 Materials

For our study, we extracted four distinct datasets from published phoneme inventory collections with global scope. JIPA provides a collection of 131 phoneme inventories 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, http://www.lapsyd.ddl.cnrs.fr/lapsyd/) consists of 584 sound inventories which we extracted from the original database. Two 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 449 inventories and was a constituent part of the first version of PHOIBLE (see Moran 2012 for details). It should be noted that UPSID and LAPSyD have the same primary coder, Ian Maddieson. We also extracted a fourth 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 872 language varieties and they are treated as a single resource here, labelled PH-UZ-GM

in what follows. The justification for this is that they all have the same primary coder, Steven Moran, and follow the same coding principles, while other collections in PHOIBLE come from different sources using correspondingly different coding principles.

As can be seen from the comparison of the data in Table 1, 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 are described in the following section.

Table 1
Overview of the datasets used in this study

| Dataset | Description | Varieties | Sounds | Source |
|---|---|---|---|---|
| **UPSID** | UCLA Phonological Segment Inventory Database as provided by PHOIBLE | 451 | 848 | Maddieson 1984 and Maddieson and Precoda 1990, provided by Moran et al. 2014 |
| **LAPSyD** | Lyon-Albuquerque Phonological Systems Database. | 583 | 810 | Maddieson et al. 2013 |
| **PH-UZ-GM** | Core data collected for PHOIBLE. | 872 | 1445 | Moran and McCloy 2019 |
| **JIPA** | Illustrations of the IPA series in the Journal of the IPA. | 159 | 957 | Baird et al. 2021 |

# 2.2 Methods

## 2.2.1 Standardising phoneme inventory data

Having access to different datasets in digital form does not guarantee that one can directly compare them and this is true also for the phoneme inventory datasets we selected for our study. The main obstacles for 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 four 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 write 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 < d> and < d> 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 < k> or as < k>, 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  I ti  s] 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 8000 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 <> is the prescribed order of the diacritics <> and <>. 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 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 four 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. The refined reference catalogue is submitted along with the supplemental material accompanying this study. In the future, we plan to offer our refinements 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 the Appendix A.

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

| Dataset | Graphemes | CLTS Sounds | Mapped | Unmapped | Coverage |
|---|---|---|---|---|---|
| **LAPSyD** | 894 | 840 | 894 | 0 | 100% |
| **PHOIBLE** | 3182 | 2756 | 3094 | 86 | 97% |
| **JIPA** | 962 | 826 | 942 | 20 | 98% |

At this point, we extracted two distinct datasets from the PHOIBLE collection: UPSID and PH-UZ-GM. 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. 2020) 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 (DOI). 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 the 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 UZ-PH-GM 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.

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

| Dataset | Number of varieties | With Glottocode | Mapped to CLTS | Distinct Glottocodes | Excluded |
|---|---|---|---|---|---|
| **UPSID** | 451 | 451 | 450 | 450 | 1 |
| **LAPSyD** | 584 | 584 | 584 | 584 | 0 |
| **PH-UZ-GM** | 932 | 932 | 899 | 841 | 33 |
| **JIPA** | 159 | 155 | 147 | 144 | 12 |

Providing this consistent mapping of language varieties and transcriptions for the four 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 can be conveniently used in a uniform way from within Python applications. Appendix B 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 four datasets to the standards proposed by the CLTS project, we increase the comparability of the phoneme inventories in our four datasets 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 firstly computed the overall *inventory sizes* by counting the number of graphemes for each inventory. This is also a useful 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, with the latter including diphthongs.

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. A more refined measure to compare similarity is thus needed and for this we measured the Jaccard similarity (Batagelj and Brem 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.

When considering only the sounds mapped to the CLTS catalogue, more sophisticated similarity measures can be computed. For this purpose, we defined a *weighted Jaccard similarity*, which takes into account the similarity of sounds with respect to their representation in the CLTS feature system. The procedure starts by selecting one sound of the first inventory and then iterating over all sounds in the other inventory, searching for the sound with the highest similarity. Once this sound has been identified, it is removed from the second inventory, and the similarity between the two sounds is stored in a list. Remaining sounds that cannot be further matched between the two inventories are given a similarity rating of 0 and also added to the list. The overall similarity between two inventories is then computed by taking the mean similarity attested in the list. Similarity between individual sounds is computed by taking the Jaccard similarity of the CLTS feature values of the two sounds. Since the score may depend on the order by which inventories are selected, we compute two scores, one starting from the first, and one starting from the second inventory and then take the average similarity score. Since this kind of inventory similarity uses approximate similarities between individual sounds based on their feature values, we call it *approximate similarity*.

In order to contrast the differences between strict and approximate inventory similarities, consider two fictitious inventories, one consisting of the phonemes /a e i u p t k/ and one consisting of the phonemes /a e i u b d g/. Since both inventories have four phonemes in common, i.e. /a e i u/, and the total number of distinct phonemes is ten, we divide 4 by 10 and receive a strict similarity score of 0.4. For the approximate similarity, all four phonemes which the two systems have in common are added to the list with a similarity score of 1. The remaining three phonemes can be matched with each other, with the closest matches consisting of the pairing of /p/ with /b/, /t/ with /d/, and /k/ with /g/. Each of these is defined by four features (type, place, manner, voicing), and each pair differs only with respect to voicing. As a result, we have a total of five distinct feature values and three matches, which means that the similarity is 3/5 = 0.6. We thus add three times 0.6 to our list of similarity scores [1, 1, 1, 1, 0.6, 0.6, 0.6]. Taking the average of the values on these lists yields a similarity score of 0.83. Appendix C illustrates how one can compute strict and approximate similarities from individual sound inventories in Python.

## 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 four 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 or can be accessed online at https://phonemeinventory.netlify.app.

## 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 four datasets. Since we use Glottocodes to identify identical language 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 four datasets.

Table 4
Mutual coverage statistics for the four datasets. The values show the number of distinct language varieties per dataset and the number of Glottocode matches between them

| | UPSID | LAPSyD | PH-UZ-GM | JIPA |
|---|---|---|---|---|
| **UPSID** | 450 | | | |
| **LAPSyD** | 304 | 584 | | |
| **PH-UZ-GM** | 54 | 119 | 841 | |
| **JIPA** | 37 | 56 | 87 | 144 |

The differences in mutual coverage are to be expected. 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, so 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 PH-UZ-GM 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 PH-UZ-GM and JIPA, this is likely a result of the compilers of PH-UZ-GM targeting the *Illustrations of the Journal of the IPA*, given their high quality and accessibility as a source of phonemic inventories. It should be noted also that the JIPA dataset is not very well balanced in terms of its genealogical coverage (see Baird et al. 2021), with a very high proportion of Indo-European languages, while the UPSID and LAPSyD datasets have a more balanced genealogical sample. This regional bias, alongside the fact that many of the source materials for JIPA had not yet been published by the time that UPSID was completed, may account for the relatively low mutual coverage between these datasets.

# 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 submitted to the Open Science Framework and can be downloaded from https://osf.io/25k7f/?view_only=e57456a73c8a4aed8c2d899ca7d7e3dd. The appendix accompanying this study contains additional information, the data, and the code.

# 3 Results

As a first test, we computed inventory sizes from the data, and then tested the Spearman 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. However, there is a considerable degree of variation, with a range from 0.48 and 0.84 for all sounds, from 0.62 to 0.90 for consonants, and from 0.63 to 0.87 for vowels (including diphthongs). These correlations are visualised in Fig. 2. The correlations are in general relatively low and while there is some variation in the correlations for consonant and vowel inventories considered separately, no consistent trend can be observed.

The high correlation in inventory sizes between UPSID and LAPSyD likely results in part from the genealogical relationship between these two datasets. UPSID inventories also differ considerably from those of the other two datasets however, while the correlations between LAPSyD, JIPA, and PH-UZ-GM are broadly comparable. All in all, these results are rather discouraging: while there is a correlation between consonant and vowel inventory sizes for the same language varieties across the different datasets, this correlation is lower than one might have expected.

Table 5
Correlations in size for complete inventory, consonant inventory, and vowel inventory. All correlations are highly significant with *p*-values below 0.005 (see Appendix E2)

| Type | UPSID – LAPSyD | UPSID – PH-UZ-GM | UPSID – JIPA | LAPSyD – PH-UZ-GM | LAPSyD – JIPA | PH-UZ-GM – JIPA |
|---|---|---|---|---|---|---|
| **Sample** | 304 | 54 | 37 | 119 | 56 | 87 |
| **All sounds** | 0.84 | 0.62 | 0.48 | 0.84 | 0.84 | 0.84 |
| **Consonants** | 0.90 | 0.62 | 0.65 | 0.87 | 0.84 | 0.80 |
| **Vowels** | 0.72 | 0.69 | 0.64 | 0.70 | 0.87 | 0.85 |

As a further way to visualise the differences found, Fig. 3 provides geographic maps of the calculated differences between individual varieties in all six combinations of the four datasets. As can be seen from this figure, it is hard to detect clear trends in the comparison. No obvious geographical tendency emerges. All in all, however, we can see that the differences in inventory sizes for individual language varieties are disappointingly high in all comparisons.

As a next step, we investigated the similarities of phoneme inventories across datasets. Here, we provide the strict and approximate similarities in Table 6, computed for all phonemes in the samples. We can find a similar trend with respect to those pairings that showed low correlations in Table 5, especially reflected in the low strict and approximate similarity scores of pairings involving the UPSID database, especially so with JIPA and PH-UZ-GM. Given the notable differences in strict and approximate similarities, we can also see the direct benefits of harmonising the transcription systems by mapping them to CLTS. All in all, these results are much lower than one might have expected. We note quite drastic differences in the way in which phoneme inventories for supposedly identical language varieties have been coded.

Table 6
Strict and approximate similarities for all pairs of language varieties in the datasets

| Similarity | Type | UPSID – LAPSyD | UPSID – PH-UZ-GM | UPSID – JIPA | LAPSyD – PH-UZ-GM | LAPSyD – JIPA | PH-UZ-GM – JIPA |
|---|---|---|---|---|---|---|---|
| strict | Sounds | 0.67 | 0.49 | 0.43 | 0.64 | 0.64 | 0.71 |
| | Consonants | 0.77 | 0.56 | 0.50 | 0.65 | 0.65 | 0.74 |
| | Vowels | 0.50 | 0.37 | 0.32 | 0.67 | 0.69 | 0.72 |
| approximate | Sounds | 0.83 | 0.70 | 0.68 | 0.79 | 0.80 | 0.86 |
| | Consonants | 0.87 | 0.74 | 0.72 | 0.81 | 0.81 | 0.87 |
| | Vowels | 0.73 | 0.60 | 0.60 | 0.79 | 0.82 | 0.85 |

In order to check if any systematic differences in the datasets could be observed, we computed the average size for the total inventory, 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 in all datasets. The results of this computation are given in Table 7. Here, we see striking differences between the datasets, with respect to the average sizes, with a range between an average of 30.95 sounds per inventory in UPSID to an average of 41.63 in JIPA. In general, LAPSyD and UPSID have many fewer sounds than PH-UZ-GM and JIPA. In fact, the average inventory in UPSID is less than three quarters the size of the average inventory in JIPA. The differences in size hold for both consonants and vowels.

Table 7
Mean numbers of sounds, consonants, and vowels per inventory in each of our four datasets, as well as the proportions of language varieties containing long consonants, long vowels, and diphthongs.

| Dataset | Average Sizes | | | Proportions | | |
|---|---|---|---|---|---|---|
| | Sounds | Consonants | Vowels | Long consonants | Long vowels | Diphthongs |
| UPSID | 30.95 | 22.44 | 8.5 | 0.03 | 0.11 | 0.11 |
| LAPSyD | 31.29 | 21.44 | 9.85 | 0.01 | 0.4 | 0.12 |
| PH-UZ-GM | 37.01 | 25.91 | 11.11 | 0.08 | 0.39 | 0.12 |
| JIPA | 41.63 | 28.62 | 13 | 0.17 | 0.42 | 0.31 |

The comparison of the proportions of varieties with long consonants and vowels as well as diphthongs also show considerable differences between the four datasets. In particular, the inventories in USPID and LAPSyD very rarely include long consonants, while these are considerably more frequent in PH-UZ-GM and, especially, in JIPA. Also, long vowels are much less frequent in UPSID than in the other datasets. Finally, diphthongs are much more common in JIPA than in the other three datasets.

It is clear that these differences are likely to contribute to the relatively low correlations for inventory size and inventory similarity. They are discussed further below.

# 4 Discussion

We aimed to compare inventories along two different axes of comparability. The first axis is systemic comparability, or the extent to which two different inventories reflect the same underlying analysis of a given language. The second axis is graphemic comparability, the extent to which the same graphemes are used to represent equivalent 'points in the pattern'. Our results show sometimes quite striking variation on both axes. On the one hand there is considerable variation in inventory size for different inventories for the same language, suggesting low systemic comparability. On the other hand, even when inventories for a given language reflect the same underlying analysis, we find low graphemic comparability, as shown by the strict similarity scores we have reported.

The following sections discuss some of the reasons for the variation we find across our dataset. Some of this variation appears to result from the use of diffierent source material and is discussed in Sect. 4.1. However, further variation stems from differing interpretation of the same source material, i.e. the same sources are used, but different inventories are coded. This is discussed in 4.2. A particularly important source of variation in this regard is the treatment of consonant and vowel length, which is dealt with in Sect. 4.3. Section 4.4 examines rather graphemic variation, especially variation resulting from the use of graphemes of different phonetic specification to represent the phonemes of the same underlying analysis.

# 4.1 Systemic variation: different source material

The phoneme is first and foremost a technology of linguistic description, and phoneme inventories have the primary function of presenting, in a well-ordered way, information about the sound system of a specific language variety. Individual linguists vary in terms of their theoretical proclivities and aesthetic preferences and consequently differ in their analytical choices when drawing up phonemic descriptions. For this reason alone one might expect a degree of systemic variation in the inventories given in published sources for any given language: different sources may present different phonemic systems.

Different phonemic analyses in source material may be faithfully reflected in our datasets. For Bemba (bemb1257) there is a series of prenasalised consonants in the JIPA inventory that are absent in the UZ-PH-GM one, reflecting analytical choices in the original sources for these inventories (Hamann and Kula 2015 and Kula 2002 respectively). Even though there are no further differences between the inventories, this pushes their strict similarity score down to 0.69. Similarly, for Ukrainian (ukra1253), the JIPA inventory includes a palatalised series of consonants that is not present in the UZ-PH-GM one, which is drawn from a different source. Combined with different transcriptional choices for other segments, this pushes the strict similarity as low as 0.37.

Impressionistically, differences in original source material are likely to be especially pertinent to the divergence of UPSID from the other datasets, seeing as UPSID was completed in 1984 and is by now a legacy dataset, while the JIPA dataset comes from very recent work, and LAPSyD and the sections of PHOIBLE extracted for the current study are under continual development. This means that the compilers of the latter three datasets were able to draw on a wider range of sources, and often more detailed and comprehensive ones, than were available during the development of UPSID.

There is a high degree of mutual coverage between the UPSID and LAPSyD, as these have the same principal coder, and some differences in inventories for the same language in these datasets are driven by the fact that he was able to draw on a larger range of sources in developing LAPSyD. The inventories for San Miguel el Grande Mixtec (sanm1295) illustrate a case in which UPSID draws on more limited and earlier source material. Here the UPSID inventory can draw on only a single journal article (Hunter and Pike 1969), while the LAPSyD one has recourse to a more comprehensive grammatical description (Macaulay 1996) and an in-depth phonetic study led by the same author (Macaulay and Salmons 1995). This leads to very different resulting inventories: the strict similarity score between the UPSID and LAPSyD inventories for this language is only 0.24.

## 4.2 Systemic variation: different interpretation of the same source material

However, a high degree of systemic variation can occur even when the same source material is used. One cause of variation arises from different interpretations of what should be considered phonemic. This often leads to inventories for the same language, drawn from the same source material, in which one dataset has an entire series of vowels or consonants lacking in another, leading often to considerable divergence in inventory size. Some of this variation appears to be quite irregular.

Sometimes, the inventory in one dataset has an entire series of consonants lacking in another. Examples are not difficult to find. For Mambay (mamb1294) JIPA has a series of pharyngealised vowels, lacking in UZ-PH-GM; while for Tamil (tami1289) LAPSyD and UZ-PH-GM have a voiced series of consonants not found in the JIPA inventory. The inclusion of an extra series of consonants can result in dramatically different inventory sizes, particularly if it concerns secondary localisation. For example, for Arrernte (east2379) JIPA has a full labialised series of consonants, and some palatalised ones, having thus 52 phonemes, while UZ-PH-GM does not, having only 31; for Tangale (nucl1696) the 58 phonemes in UZ-PH-GM includes a labialised series of consonants, absent in LAPSyD, which has only 41 phonemes; while for Siwi (siwi1239) it is rather LAPSyD that includes a labialised series, alongside some palatalised ones, for 38 consonants in total, while UZ-PH-GM, without these, has only 26 consonants.

While the examples above show variation between datasets due to different source material, we find also considerable variation when datasets share the same source. Even between LAPSyD and UPSID, examples are numerous, including Achuwami (achu1247), where LAPSyD includes 39 phonemes and UPSID 23, with a strict similarity of 0.32; Khalkha (halh1238), where LAPSyD has 53 phonemes and UPSID 33, with a strict similarity of 0.31; Kurukh (kuru1302), where LAPSyD has 52 phonemes and UPSID

32, with a strict similarity of 0.35. Although the LAPSyD inventories are usually larger than their UPSID counterparts, the situation may be reversed: such is the case with Wichita (wich1280), where LAPSyD has 16 phonemes and UPSID 29, with a strict similarity of only 0.18.

Our results also found some predictable variation between the datasets, both in terms of the average numbers of phonemes per language and in differing treatment of consonant length, vowel length, and diphthongs. With respect to long consonants, LAPSyD has them in only 1.2% of inventories and UPSID in only 2.67% of inventories, while they occur much more frequently in the other two datasets: in 8.01% of inventories in PH-UZ-GM and in 17.01% of those in JIPA. Careful perusal of the data shows that the relationship between LAPSyD, PH-UZ-GM and JIPA with respect to long consonants can be stated implicationally: if LAPSyD marks consonant length, both PH-UZ-GM and JIPA do too, and if PH-UZ-GM marks consonant length then JIPA also does.

With respect to long vowels, our results showed that UPSID more rarely marks vowel length, doing so in only 11.33% of inventories, while the corresponding figures for the other datasets are broadly comparable: 39.72% for LAPSyD, 38.93% for PH-UZ-GM, and 42.18% for JIPA. This being the case, it is impossible to predict when one of these three datasets includes long vowels and another does not. For Central Sama (cent2092) JIPA has long vowels, while PH-UZ-GM does not; for Slovene (slov1268) PH-UZ-GM has long vowels, while LAPSyD and JIPA do not; and for Comanche (coma1245) LAPSyD has long vowels, whereas PH-UZ-GM does not.

# 4.3 Systemic variation: consonant and vowel length

Differing treatments of consonant and vowel length is a particularly common source of systemic variation in our dataset. Parsing long consonants and diphthongs (and in some cases long vowels as well) as unit phonemes is an example of underanalysis, i.e. considering a phonetically complex span to be a single entity, e.g. the long consonant /t□/, or the diphthong /ai/. The alternative is overanalysis, treating such spans as combinations of more than one entity, e.g. /t/ + /t/ or /a/ + /i/. Broadly speaking, PH-UZ-GM and especially JIPA tend towards underanalysis, while UPSID and LAPSyD tend towards overanalysis, especially when it comes to the treatment of consonant length.

Some of the most striking examples of variation due to consonant and vowel length, as well as the treatment of diphthongs come from the JIPA and UZ-PH-GM inventories for Estonian and Danish, drawing from the same source in each case. For Estonian (esto1258), there is only 0.3 similarity between the JIPA and UZ-PH-GM inventories. Vowel length is considered phonemic in the JIPA inventory, which has 70 vowels and 25 consonants, while it is rather consonant length that is understood as phonemic in the UZ-PH-GM one, with 9 vowels and 45 consonants. For Danish (dani1285), the JIPA and UZ-PH-GM inventories have a strict similarity score of 0.33. The former has 47 phonemes, including pharyngealised consonants and vowels but no diphthongs, while the latter does not have these pharyngealised phonemes but rather 34 diphthongs. While these languages have admittedly created difficulties in analysis for phonologists (see Kuznetsova 2018) they are also very well-studied and documented, proving that the problems we identify here do not derive from a lack of scholarly investigation.

The decision to include long consonants in an inventory frequently has the effect of doubling, or nearly so, the size of the consonant inventory. For example, while the JIPA, UZ-PH-GM, and LAPSyD inventories for Kunama (kuna1268) are all drawn from the same source, JIPA includes long consonants, whereas the other two inventories do not. This means that it includes 53 phonemes, and has a strict similarity of 0.4 with the UZ-PH-GM inventory (31 phonemes) and of 0.49 with the LAPSyD inventory (35 phonemes), whereas the UZ-PH-GM and LAPSyD pair, both lacking long consonants, have a higher strict similarity of 0.61. Similar examples are relatively frequent: the JIPA inventory for Hungarian (hung1274), with 68 phonemes, has a strict similarity of only 0.33 with the LAPSyD inventory, which has only 40; the UZ-PH-GM inventory for Tsamai (tsam1247) has 28 long consonants, whereas that of LAPSyD has only 2, with a strict similarity of 0.5; the UZ-PH-GM inventory for Krongo (kron1241) has 19 long consonants, with none in LAPSyD, giving a strict similarity is 0.54. In all these cases, the inventories were drawn from the same source.

# 4.3 Graphemic variation: degree of phonetic specification

Even where inventories reflect the same systemic analysis, or a very similar one, there can be considerable graphemic variation: different graphemes may be used for each ‘point in the pattern’. The underlying problem here is that phonemes are cover-symbols for a wide number of differing realisations depending on a very wide variety of factors. Most people collecting phoneme inventories are committed to using a transcription representing the *basic allophone* of a given phoneme. However, it is by no means always clear what that actually is and different researchers vary in their policies. While each of them may be principled and coherent in their own practice, these practices themselves may differ.

A common problem concerns the degree of phonetic specification of phonemes in the different datasets. By and large, the graphemes used in PH-UZ-GM tend to carry more phonetic information than those of JIPA or LAPSyD. This type of variation is quite common with respect to the marking of aspiration on voiceless stops in languages in which this occurs, and here LAPSyD tends to omit the aspiration. Another common pattern, although an apparently unsystematic one, is that one inventory presents graphemes without a dental diacritic, e.g. /t d n/, while another has graphemes with the diacritic, e.g. /t̪ d̪ n̪/.

Differences in the degree of phonetic specification in grapheme choice are illustrated by Standard Arabic (stan1318), where UZ-PH-GM marks retracted tongue root on consonants, while JIPA does not; by Ega (egaa1242) where the inclusion of diacritics on vowels in UZ-PH-GM, but not in JIPA, means that only one of the nine vowels in the language are strictly similar between the two inventories; and by Basque (basq1248), where UZ-PH-GM uses both apical and dental diacritics, while LAPSyD uses only dental diacritics.

A subset of this type of variation is the variable marking of aspiration on voiceless stops in languages in which this occurs. A common pattern observed is for the aspiration to be included in UZ-PH-GM but absent in LAPSyD, e.g. in German (stan1295), Persian (west2369), Irish (iris1253), Tuvinian (tuvi1240), Kabiye (kabi1261).

A further subset concerns the variable presence of dental diacritics on coronal consonants. This appears to be unsystematic, so for French (stan1290) UZ-PH-GM and LAPSyD have the diacritic but JIPA does not; for Bardi (bard1255) UZ-PH-GM has the diacritic while JIPA and LAPSyD do not; while for Ersu (ersu1241) JIPA has the diacritic while UZ-PH-GM does not. Sometimes, the presence or absence in this diacritic can have quite significant bearing on comparability: while the UZ-PH-GM and JIPA inventories for Jicarilla Apache are nearly the same size (53 vs 52 phonemes), 15 phonemes are found with the dental diacritic in the former inventory and without it in the latter, being the major contributing factor in pushing the strict similarity score down to 0.5.

# 5 Conclusion

Scholarly work has used phoneme inventories as evidence for studies in a rather wide range of topics, with an implicit assumption that they constitute a reliable and robust source of data. For these studies, our results sound a note of caution. 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 rather 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.

## Declarations

# Funding information

As part of the CLLD project (cf. https://clld.org) and the Glottobank project (cf. https://glottobank.org), the work was supported by the Max Planck Society, the Max Planck Institute for the Science of Human History, and the Royal Society of New Zealand (Marsden Fund grant 13-UOA-121). JML was funded by the ERC Starting Grant 715618 *Computer-Assisted Language Comparison* (cf. https://digling.org/calc/). SJG was supported by the Australian Research Council's Discovery Projects funding scheme (project number DE 120101954) and the ARC Center of Excellence for the Dynamics of Language grant (CE140100041).

# Competing interests

The authors declare no competing interests.

## References

Anderson, Cormac, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. "A Cross-Linguistic Database of Phonetic Transcription Systems." *Yearbook of the Poznań Linguistic Meeting* 4 (1): 21–53.

Anderson, Stephen R. 1985. *Phonology in the Twentieth Century*. Chicago: University of Chicago.

Atkinson, Quentin D. 2011. "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa." *Science* 332 (6027): 346–49.

Baird, Louise, Nicholas Evans, and Simon J. Greenhill. 2021. Blowing in the wind: Using North Wind and Sun texts to sample phoneme inventories. *Journal of the IPA*. 1–42. https://doi.org/10.1017/S002510032000033X

Batagelj, Vladimir, and Matevz Bren. 1995. "Comparing Resemblance Measures." *Journal of Classification* 12: 73–90.

Baudouin de Courtenay, Jan. (1894) 1972. "Próba Teorii Alternacji Fonetycznych." In *A Baudouin de Courtenay Anthology. The Beginnings of Structural Linguistics*, translated by Edward Stankiewicz. Bloomington: Indiana University Press.

Blasi, Damián E., Steven Moran, Scott R. Moisik, Paul Widmer, Dan Dediu, and Balthasar Bickel. 2019. "Human Sound Systems Are Shaped by Post-Neolithic Changes in Bite Configuration." *Science* 363 (1192): 1–10. https://doi.org/10.1126/science.aav3218.

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

Ceolin, Andrea, Cristina Guardiano, Monica Alexandrina Irimia, and Giuseppe Longobardi. 2020. “Formal Syntax and Deep History.” *Frontiers in Psychology* 11: 2384. https://doi.org/10.3389/fpsyg.2020.488871.

Chao, Yuen Ren. 1934. “The Non-Uniqueness of Phonemic Solutions of Phonetic Systems.” *Bulletin of the Institute of History and Philology* 4 (4): 363–97.

Chirkova, Katia, and Yiya Chen. 2013. “Lizu.” *Journal of the International Phonetic Association* 43 (1): 75–86. https://doi.org/10.1017/s0025100312000242.

Chirkova, Katia, Dehe Wang, Yiya Chen, Angélique Amelot, and Tanja Kocjančič Antolík. 2015. “Ersu.” *Journal of the International Phonetic Association* 45 (2): 187–211. https://doi.org/10.1017/s0025100314000437.

Creanza, N., M. Ruhlen, T. J. Pemberton, N. A. Rosenberg, M. W. Feldman, and S. Ramachandran. 2015. “A Comparison of Worldwide Phonemic and Genetic Variation in Human Populations.” *Proc. Natl. Acad. Sci. U.S.A.* 112 (5): 1265–72.

Crothers, John H., James P. Lorentz, Donald A. Sherman, and Marilyn M. Vihman. 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.

Crothers, John H. 1978. “Typology and universals of vowel systems”. In Greenberg, Joseph H., Charles A. Ferguson, E. A. Moravcsik (eds.), *Universals of human language, Vol. 2: Phonology*. Stanford: Stanford University Press, 93–152.

de Saussure, Ferdinand. 1916. *Cours de Linguistique Générale*. Edited by Charles Bally. Lausanne: Payot.

Dediu, Dan, and Scott Moisik. 2019. “Pushes and Pulls from Below: Anatomical Variation, Articulation and Sound Change.” *Glossa* 4 (1): 1–33.

DeMille, Melissa M. C., Kevin Tang, Chintan M. Mehta, Christopher Geissler, Jeffrey G. Malins, Natalie R. Powers, Beatrice M. Bowen, Andrew K. Adams, Dongnhu T. Truong, Jan C. Frijters, and Jeffrey R. Gruen. Worldwide distribution of the DCDC2 READ1 regulatory element and its relationship with phoneme variation across languages. *Proc. Natl. Acad. Sci. U.S.A*. 2018 May 8; 115 (19): 4951–4956. https://doi.org/10.1073/pnas.1710472115. Erratum in: *Proc. Natl. Acad. Sci. U.S.A.* 2018 May 21.

Donohue, Mark, Rebecca Hetherington, James McElvenny, and Virginia Dawson. 2013. *World Phonotactics Database*. Canberra: Department of Linguistics. The Australian National University. http://phonotactics.anu.edu.au.

Donohue, Mark, and Johanna Nichols. 2011. “Does Phoneme Inventory Size Correlate with Population Size?” *Linguistic Typology* 15 (2): 161–70, https://doi.org/10.1515/lity.2011.011.

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

Everett, Caleb, Damian E. Blasi and Seán G. Roberts. 2015. “Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots.” *Proc. Natl. Acad. Sci. U.S.A.* 112 (5) 1322–1327. 112 (5) 1322–1327. https://doi.org/10.1073/pnas.1417413112.

Everett, Caleb, and Sihan Chen. 2021. “Speech Adapts to Differences in Dentition Within and Across Populations.” *Scientific Reports* 11 (1066): 1–10. https://doi.org/10.1038/s41598-020-80190-8.

Forkel, Robert, and Johann-Mattis List. 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).

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. “Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.” *Scientific Data* 5 (180205): 1–10.

Georgiou, Georgios P., and Ahmad Kilani. 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, Morris. 1963. “Phonemics.” In *Soviet and East European Linguistics*, edited by Thomas Sebeok, 5–21. Current Trends in Linguistics 1. Amsterdam; New York: De Gruyter Mouton. https://doi.org/10.1515/9783110814620.

Hamann, Silke and Kula, Nancy Chongo. 2015. Bemba. In *Journal of the International Phonetic Association* 45: 61–69.

Hammarström, Harald, Martin Haspelmath, Robert Forkel, and Sebastian Bank. 2020. *Glottolog. Version 4.3*. Jena: Max Planck Institute for the Science of Human History. https://glottolog.org.

Harris, Zellig. 1951. *Structural Linguistics*. Chicago: Phoenix.

Hjelmslev, Louis. 1943. *Omkring Sprogteoriens Grundlæggelse*. København: Akademisk forlag.

Hunter, Georgia G. and Pike, Eunice V. 1969. The Phonology and Tone Sandhi of Molinos Mixtec. *Linguistics* 47. 24–40.

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

Johansson, Niklas Erben, Andrey Anikin, Gerd Carling, and Arthur Holmer. 2020. “The Typology of Sound Symbolism: Defining Macro-Concepts via Their Semantic and Phonetic Features.” *Linguistic Typology* 24 (2): 253–310. https://doi.org/10.1515/lingty-2020-2034.

Jones, Daniel. 1950. *The Phoneme, Its Nature and Use*. Cambridge: Heffer.

Kiparsky, Paul. 2018. Formal and empirical issues in phonological typology. In L. M. Hyman & F. Plank eds. *Phonological typology*. De Gruyter Mouton. 54–106.

Kula, Nancy Chongo. 2002. *The Phonology of Verbal Derivation in Bemba*. Utrecht: LOT.

Kuznetsova, Natalia. 2018. What Danish and Estonian can show to a modern word-prosodic typology. In Rob Goedemans, Jeffrey Heinz & Harry van der Hulst (eds.), *The study of word stress and accent: theories, methods and data* (Conceptual Foundations of Language Science), 102–143. Cambridge: Cambridge University Press.

Levinson, Steven, and Nicholas Evans. 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, Johann-Mattis. 2019. “Beyond Edit Distances: Comparing Linguistic Reconstruction Systems.” *Theoretical Linguistics* 45 (3–4): 1–10.

List, Johann-Mattis, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. *Cross-Linguistic Transcription Systems. Version 2.0.0*. Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.3515744.

Macaulay, Monica. 1996. *A Grammar of Chalcatongo Mixtec*. University of California Press, Berkeley and Los Angeles.

Macaulay, Monica & Joe Salmons. 1995. The phonology of glottalization in Mixtec. In *International Journal of American Linguistics* 61: 38–61.

Maddieson, Ian. 1984. *Patterns of Sounds*. Cambridge; New York: Cambridge University Press.

Maddieson, Ian. 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, Ian and Precoda, Kristin. 1990. “Updating UPSID.” UCLA Working Papers in Phonetics, 104–111. Department of Linguistics, UCLA.

https://doi.org/10.1121/1.2027403.

Maddieson, Ian, and Christophe Coupé. 2015. “Human Spoken Language Diversity and the Acoustic Adaptation Hypothesis.” *Journal of the Acoustical Society of America* Am.138: 1838–8.

https://doi.org/10.1121/1.4933848.

Maddieson, Ian, Sébastien Flavier, Egidio Marsico, Cristophe Coupé, and François Pellegrino. 2013. “LAPSyD: Lyon-Albuquerque Phonological Systems Database.” *Proc. of 14th Interspeech Conference, Lyon, France, 25–29 August 2013*.

Moran, Steven. 2012. “Phonetics Information Base and Lexicon.” PhD, University of Washington.Moran, Steven, Daniel McCloy and Richard Wright. 2012. "Revisiting population size vs. phoneme inventory size." *Language* 88.4: 877–893.

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

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

Moran, Steven, Daniel McCloy, and Richard Wright, eds. 2014. *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://phoible.org/.

Pericliev, Vladimir. 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): 376–83, https://doi.org/10.1515/lity.2004.8.3.376.

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

Sapir, Edward. 1925. “Sound Patterns in Language.” *Language* 1 (2): 37–51. https://doi.org/10.2307/409004.

Sapir, Edward. 1933. “La Réalité Psychologique Des Phonèmes.” *Journal de Psychologie Normale et Pathologique* 30: 247–65.

Simpson, Adrian P. 1999. “Fundamental problems in comparative phonetics and phonology: does UPSID help to solve them.” *Proc. of the 14th International Congress of Phonetic Sciences*. Vol. 1. Berlin: De Gruyter.

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

Trudgill, Peter. 2004. “Linguistic and Social Typology: The Austronesian Migrations and Phoneme Inventories.” *Linguistic Typology* 8 (3): 305–20, https://doi.org/10.1515/lity.2004.8.3.305.

Twaddell, W.Freeman. 1935. “On Defining the Phoneme.” *Language* 11 (1): 5–62.

Vaux, Bert. 2009. “The role of features in a symbolic theory of phonology”. In E. Raimy & C. E. Cairns eds. *Contemporary Views on Architecture and Representations in Phonology*. 75–97.

## Supplementary Files

This study is accompanied by supplementary materials submitted to the Open Science Framework at https://osf.io/thukn/?view_only=35a40d1e4df3463ea54a08243947aeec. 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).

The interactive web application for inspecting the phoneme inventories for this study can be accessed at https://phonemeinventory.netlify.app.

## Figures

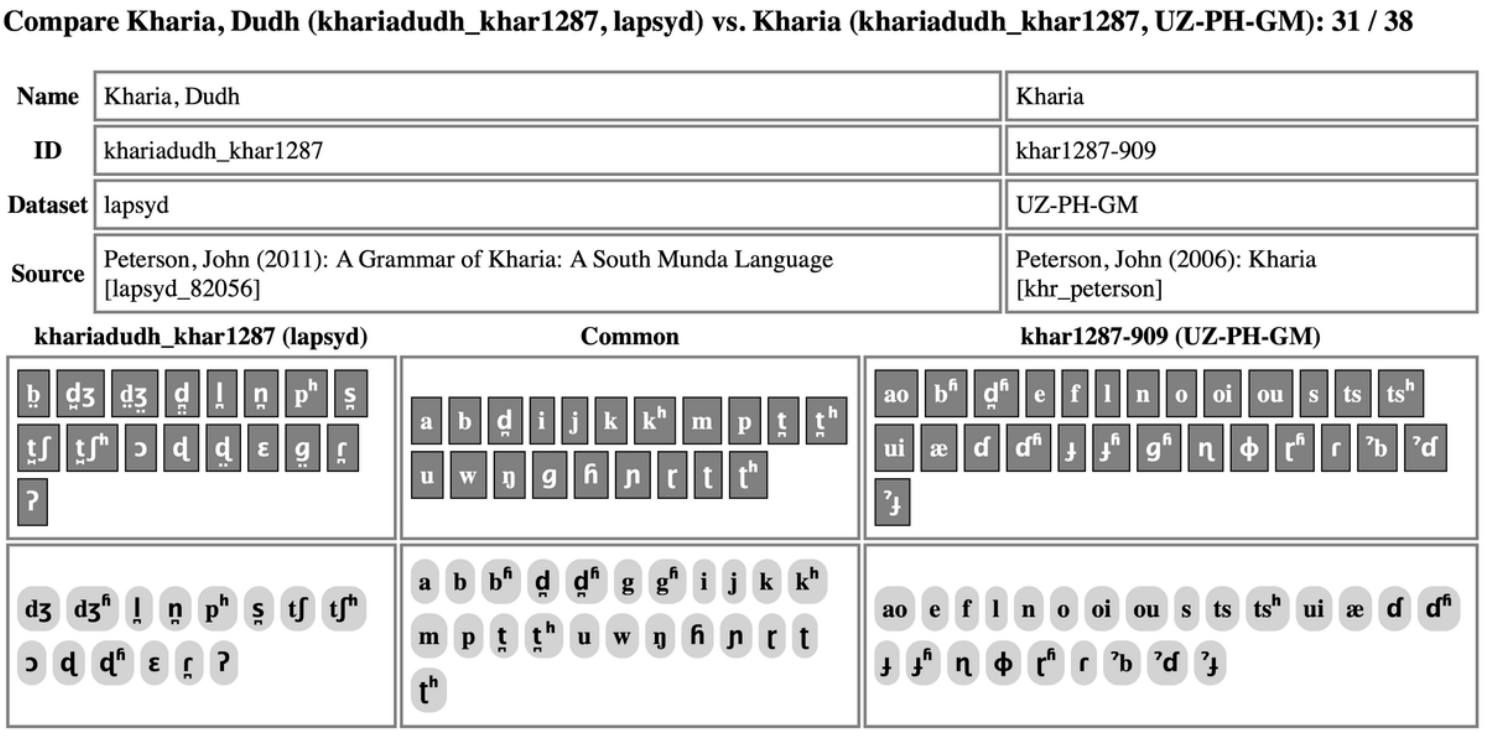
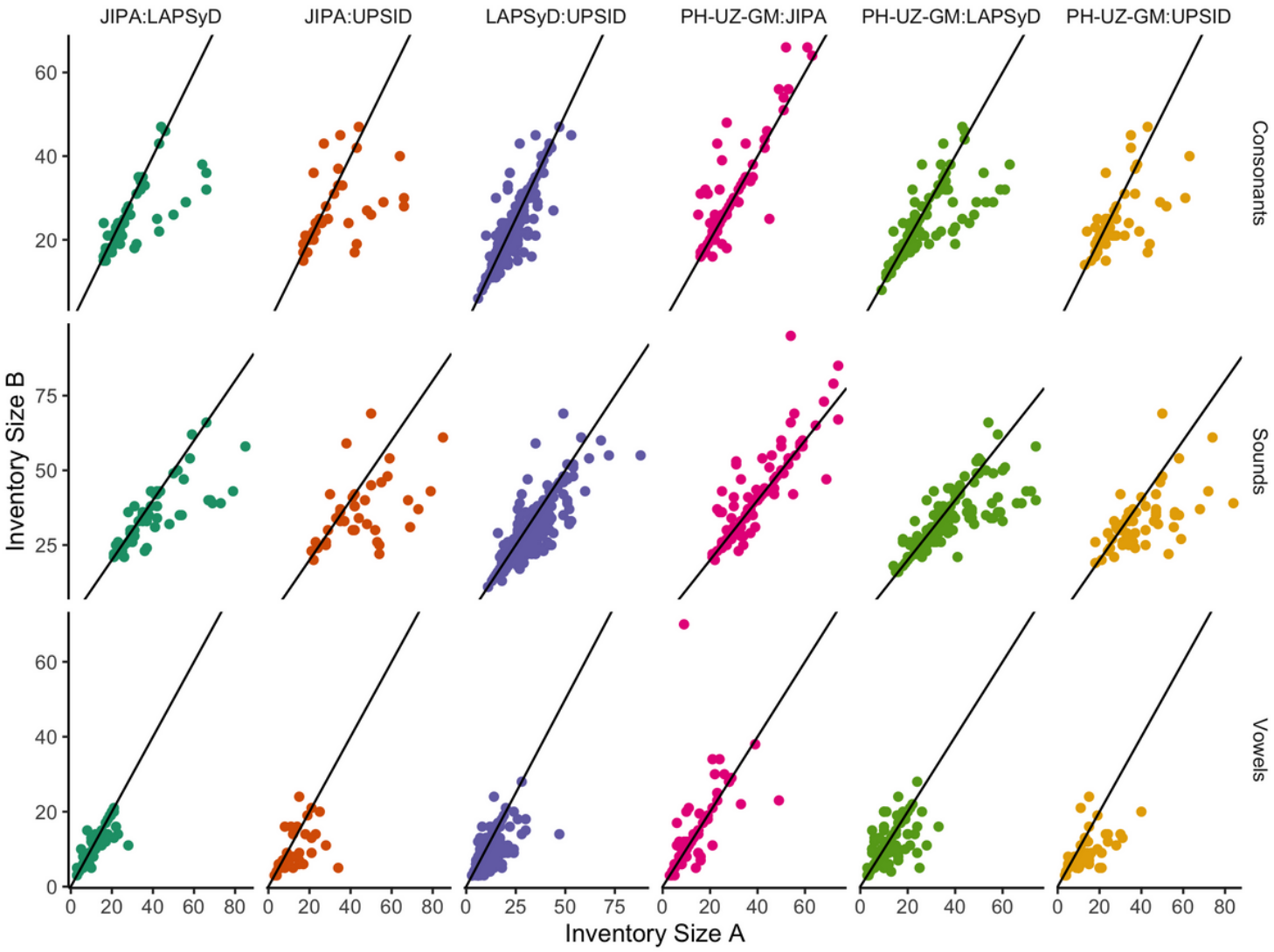


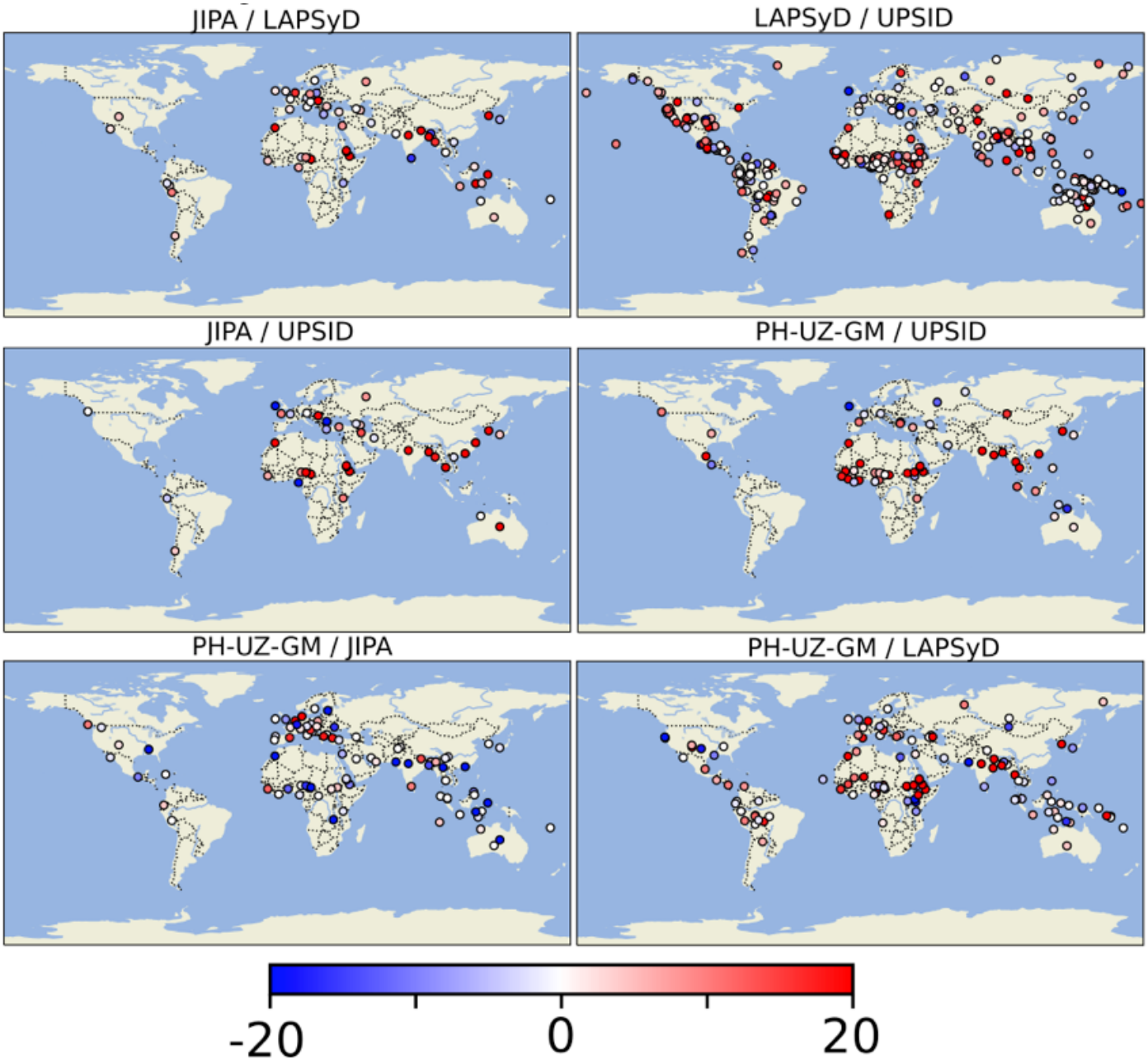

**Figure 1**

Web-based interactive application for inspecting phoneme inventories across sources and datasets, showing here the LAPSyD and UZ-PH-GM inventories for Kharia




## Figure 2

Comparing phoneme inventory sizes, consonant inventory sizes, and vowel inventory sizes for all six combinations of datasets




## Figure 3

Comparing individual differences for sound inventory sizes for the six combinations of datasets. Dots represent language varieties. In the colour bar, red indicates that the first inventory is larger than the second, blue if the second is larger than the first, with inventories of equal or very similar size surfacing as white dots.