BDPROTO: A Database of Phonological Inventories from Ancient and Reconstructed Languages

Egidio Marsico[1], Sebastien Flavier[1], Annemarie Verkerk[2], Steven Moran[3]
[1]Dynamique Du Langage, CNRS, Université de Lyon, Lyon, France
[2]Max Planck Institute for the Science of Human History, Jena, Germany
[3]Department of Comparative Linguistics, University of Zurich
{egidio.marsico, sebastien.flavier}@cnrs.fr, verkerk@shh.mpg.de, steven.moran@uzh.ch

Abstract

Here we present BDPROTO, a database comprised of phonological inventory data from 137 ancient and reconstructed languages. These data were extracted from historical linguistic reconstructions and brought together into a single unified, normalized, accessible, and Unicode-compliant language resource. This dataset is publicly available and we aim to engage language scientists doing research on language change and language evolution. We provide a short case study to highlight BDPROTO's research viability; using phylogenetic comparative methods and high-resolution language family trees, we investigate whether consonantal and vocalic systems differ in their rates of change over the last 10,000 years.

Keywords: historical linguistics, phonological inventories, rates of change, language evolution

1. Overview

First we provide some background on ancient language reconstruction and why it is interesting in light of studying the evolution of human language. Then we describe the BDPROTO language sample and our data extraction and aggregation pipelines. Afterwards we present a short case study using ancient language data from BDPROTO together with phonological inventory data from currently spoken languages to investigate evolutionary trends in consonant and vowel systems. Finally, we discuss how consonant and vowel inventories have changed over the last 10,000 years and we detail avenues for further research in historical and evolutionary linguistics.

2. Background

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

Comparative reconstruction is accomplished through systematic comparison of word forms in descendant languages. By identifying groups of potentially related words, cognates between related languages are identified (words that share form and meaning and are genealogically related). Consider the modern word for 'tooth' in four Romance languages: Spanish 'diente', Italian 'dente', French 'dent', and Portuguese 'dente'. The similarities in these words could be due to chance correspondence, word borrowing, or linguistic universals, cf. Blasi et al. (2016). However, these three factors are highly unlikely in this example because each descendent form shares the same meaning and a similar phonetic form. That is, each word begins with a 'd' sound and it contains a consonant cluster 'nt'. Thus it is highly probable that the ancestral word contained a form resembling 'd ... nt ...' (Fortson, 2004, 3).

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

Recently the comparative reconstruction approach outlined above has been implemented programmatically (Steiner et al., 2011), so that many of the time-consuming and redundant tasks of the historical linguist are automated, for example inferring regular sound change (Bouchard-Côté et al., 2013; Hruschka et al., 2015). The resulting score of similarity from pairwise sets of words across all languages in a sample can help to identify cognates. Expert judgment is still needed, but tools (List and Moran, 2013) and interfaces (List, 2017) allow even the non-tech-savvy linguist to quickly identify cognates from masses of raw data, such as word lists from thousands of languages (Wichmann et al., 2017). Word lists that are coded for cognacy and phonetic similarity scores can be used as input for one of many al-

gorithms that generate language family phylogenies. These language family trees can then be used as input to phylogenetic comparative methods developed by biologists for investigating the tree of life, but adopted and adapted by linguists and evolutionary anthropologists to address research questions about ancient language structures, cultures and population movements, e.g. Dunn et al. (2011), Gray et al. (2009), Bouckaert et al. (2012).

3. Data extraction and aggregation

The phonological inventories in BDPROTO were extracted manually from source texts,[1] interpreted by experts, and then codified according to standardized Unicode conventions (Moran and Cysouw, In press) for the International Phonetic Alphabet (International Phonetic Association, 1999). The resulting dataset was put into a Github repository, additional metadata were added, and an aggregation script was written to bring three independent and disparate input data sources together.[2]

The datasets include the original BDPROTO data from Marsico (1999) and reconstruction data collected more recently at the Department of Comparative Linguistics at the University of Zurich. The former was originally stored in SQL tables in ISO 8859-1 encoding and was for this work transformed into CSV files in UTF-8 NFC with LF and no BOM. Given the legacy character encoding, we standardized character representations using the PHOIBLE conventions.[3] Additional phonological inventories were entered by hand into Excel spreadsheets and exported as compliant Unicode Standard UTF-8 to complement and extend the sample in Marsico (1999). For each of the 478 unique speech sounds reported in BDPROTO, we integrated a distinctive feature vector from the 37 phonetic features described in PHOIBLE (Moran et al., 2014).

Supplemental metadata for each inventory was collected and is stored in the BDPROTO repository, including for each language: estimates for its age and the homeland where it was spoken. Both the time depth and the homeland of language families are hotly debated issues; see for example the discussion of the age and heartland of Indo-European (Bouckaert et al., 2012; Chang et al., 2015). Each language data point in BDPROTO is also associated with a Glottolog language identifier, so that it is positioned within a language family phylogeny (Hammarström et al., 2017).[4] Each inventory has one or more bibliographic citations, which are stored in a text-based BibTeX file, where the BDPROTO ID is mapped to the BibTeX key for easy perusal of original data sources.

The aggregation of the phonological inventory data and metadata is accomplished with a script written in the R programming language (R Core Team, 2013). This script combines the inventory data from CSV files, joins in the additional linguistic and non-linguistic metadata described above, and outputs the combined data sources as an R data object and CSV files.

[1] https://github.com/bdproto/raw-data/metadata/bdproto-references.bib

[2] https://github.com/bdproto

[3] http://phoible.github.io/conventions/

[4] http://glottolog.org/

4. The language sample

There are 137 phonological inventories in the current BDPROTO sample, which represent 126 distinct reconstructed and ancient languages from 67 different language families. The original BDPROTO sample was devised without duplicates by considering the coherence of the proposed reconstructions and their relations to their modern daughter languages. The aggregation of the original BDPROTO sample with our more recent work of collecting inventories results in duplicate data points. We consider multiple entries a feature of our database, thus allowing the user to explore and compare different reconstructions by different experts.

Figure 1 lists 25 of the oldest languages in the sample and approximately when they were spoken. Some data points in BDPROTO represent root-level language family nodes, such as Indo-European. Other data points in BDPROTO are intermediate nodes in existing proposed phylogenies. For example, there are expert reconstructions of ancient Germanic, Nordic, and Anatolian, each of which represents an intermediate node within the branches of Indo-European tree, i.e. daughter languages of Indo-European but also parent languages of currently spoken languages. Note that it is generally agreed-upon that 10,000 years is the maximum time depth of reconstruction for the comparative method (Nichols, 1992). Past this time depth, languages have simply had too much time to mutate in vocabulary through regular processes of sound change and it has not yet been discovered how to peer further back in time (although this is an active area of research, e.g. Pagel et al. (2013)). The fact that most language families have resided in geographically disparate areas and have been influenced by many other factors, including linguistic and cultural, is not beneficial for deep reconstruction.

5. Case study: consonant vs vowel rates

In a study of whether phonological inventories have become more or less complex over time, Marsico (1999) shows that languages dating back as far as 10,000 years are equally-complex in terms of their number of segments, consonant/vowel ratio, average number of consonants and vowels, and frequency hierarchy of the segments. However, Marsico (1999) also notes that modern languages tend to have slightly more consonants today than their ancestors did in the past. The same does not apply to vowels. On average the number of consonants and vowels across proto-languages in BDPROTO are 18 and 8, respectively. In comparison, modern spoken languages have on average 22 consonants and 8 vowels (Maddieson, 1984).[5]

Why is it that we observe more consonants in phonological inventories today than we see in reconstructed ancient languages of the past? We decided to test whether six language families show greater rates of change in consonant inventory size as compared to vowel inventory size using phylogenetic comparative methods. Specifically, we use BayesTraits V2 (Meade and Pagel, 2014), which implements a generalized least squares approach to modeling the evolution of continuously varying traits (Pagel,

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

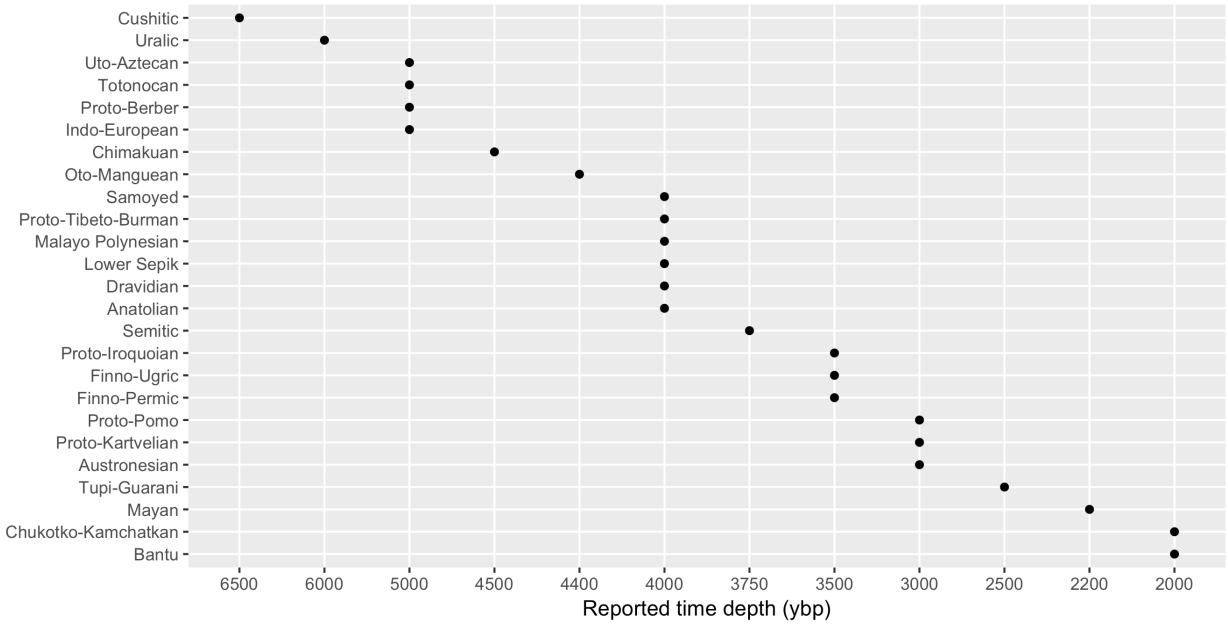

Figure 1: Approximate age of 25 language families in the BDPROTO sample

1997; Pagel, 1999). We chose these six language families because they have high-resolution expert-created phylogenies: Arawakan (language sample n=38; Walker and Ribeiro (2011)), Austronesian (83; Gray et al. (2009)), Bantu (114; Grollemund et al. (2015)), Indo-European (58; Bouckaert et al. (2012)), Pama-Nyungan (134; Bowern and Atkinson (2012)), and Tupi-Guarani (30; Michael et al. (2015)) and because they are in the BDPROTO sample. Figure 2 shows box plots of the ranges of vowel and consonant inventory size in the language samples used for phylogenetic ancestral state estimations. P gives proto-language reconstruction from BDPROTO. R gives the ancestral state estimation. An asterisk * indicates whether the rate of change of vowel or consonant inventory size is faster.[6]

Language family	Consonants	Vowels
Tupi-Guarani	5.02 ± 0.60	4.07 ± 0.52
Pama-Nyungan	3.28 ± 0.18	0.74 ± 0.04
Arawakan	17.57 ± 1.51	31.41 ± 2.76
Austronesian	129.07 ± 3.85	43.67 ± 1.22
Bantu	63.98 ± 2.94	8.52 ± 0.23
Indo-European	29.21 ± 2.42	47.00 ± 3.88

Table 1: Rates of change in 1000s of years

Table 1 gives the mean rates of change of consonant and vowels on the branches of the listed phylogenetic tree sets by 1000s of years. Our results suggest a mixed picture for the acquisition of new consonants vs vowels over the last 10,000 years. In two-thirds of the language families sampled, the rate of change in consonants systems is greater. But in Arawakan and Indo-European vowel inventory size changes faster than consonant inventory size. These two language families have in common a wider range of vowel inventory sizes as compared to the other families. However, if we take into account the mean and standard deviations of the rates given in Table 1, a high variance does not always entail a high rate and vice versa. Austronesian, for instance, has the highest rate of change for consonants, but the Austronesian languages are less variable in their inventory size than Bantu and Indo-European.

Thus our results suggest differential rates of change in consonants and vowels by language family. This finding is surprising to us because the synchronic data suggest that there is a diachronic pressure on languages to expand their consonant inventories at a greater rate than vowels; in line with the finding by Marsico (1999). For example, on average languages have more consonants than vowels, so we might expect phonological inventories to universally acquire consonants at a faster rate. Consonants are more likely to be borrowed than vowels (Moran et al., 2014). The synchronic data also show more phonetic diversity in consonant inventories, suggesting a greater number of lexical contrasts available by consonants. For example, there are three times as many contrastive consonants than vowels in the world's languages. Consonant inventories also range more in size from 6–90 (Rotokas in Papua New Guinea vs the click language !Xu, spoken in Botswana and Namibia) and vowel qualities from 2–14 (Maddieson, 2013a; Maddieson, 2013b).

Our finding warrants further research, but we might already speculate on where to look next. Inventories of both vow-

[6]We also observe that ancestral state estimates of vowel and consonant inventory sizes are generally closer to the mean of the range than expert reconstructions of proto-languages. This means there is a difference between the well-worked historical comparative method used by linguists to reconstruct proto-languages and the automated ancestral rates generated through pyhylogenetic analysis. This observation warrants a closer evaluation using directional models of feature change.

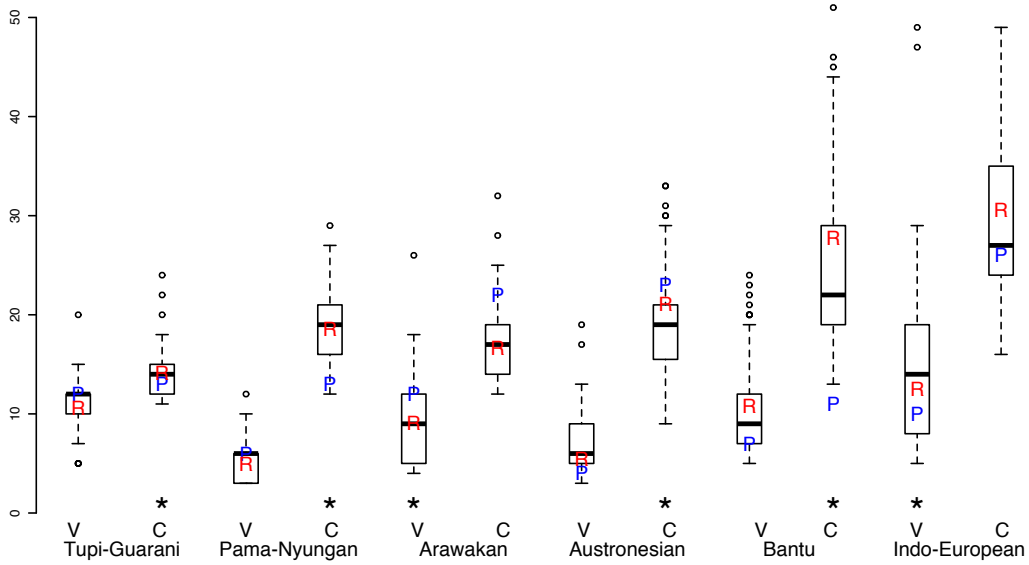

Figure 2: Ranges, reconstructions, and ancestral state estimations of vowel and consonant inventory size

els and consonants can be extended through the use of secondary articulatory features. For example, a vowel space can be expanded straightforwardly by the contrastive features length and nasalization. On the other hand, length and labialization, palatalization, and velarization, can expand consonant inventories.[7] It may be the case that mathematically there is a greater number of dimensions for consonant inventories to expand, but that there are other constraints on how consonant or vowel inventories increase in size. Hence to create more and more vocabulary, increasing the number of contrastive sounds in the phonological inventory while keeping the number of distinctive phonetic features at a minimum is said to encompass the principle of feature economy (Clements, 2009). An example is given in Moran (2012), who shows that vowel systems tend to expand from the cardinal vowels through the highly economic features length and nasalization before filling in the vowel space with peripheral vowels that require finer articulatory features and distinctions. Furthermore, Coupé et al. (2011) show that there is an asymmetry between feature economy in which vowel inventories tend to be more economical than consonant inventories. Thus the articulatory and perceptual constraints that may govern the changes in phonological inventories over time must be incorporated into models of the evolution of spoken languages.

6. Summary

Here we present BDPROTO, an open-access database of phonological inventories from a sample of 137 ancient and reconstructed languages. BDPROTO provides a rich resource for investigating historically reconstructed languages and whether they show any significant changes with languages spoken today. After an initial brief overview of the historical comparative method, we describe the data extraction and aggregation pipelines that we used to create the BDPROTO database. Finally, in a short case study we use phylogenetic methods to show that the evolution of consonant and vowel systems have differential rates of change – an unexpected observation given what we know about the ancient and reconstructed languages in the BDPROTO sample and their modern descendants.

[7] Consider for example palatalization in Russian, which increases the number of possible lexical contrasts in Russian, while being as perceptually salient a feature as primary features like voicing (Kavitskaya, 2006).

7. Acknowledgments

Special thanks to Ian Maddieson for his help with the data analysis and standardization. We also thank Balthasar Bickel, Joël Brogniart, Paul Widmer, and three anonymous reviewers.

8. Author contributions

EM, SF, SM designed the research. EM, SM collected the data. EM, SF, SM designed the database. AV, SM designed and implemented the case study. SM, AV wrote the paper.

9. Bibliographical References

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., and Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, page 201605782.

Bouchard-Côté, A., Hall, D., Griffiths, T. L., and Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Bowern, C. and Atkinson, Q. D. (2012). Computational phylogenetics and the internal structure of Pama-Nyungan. *Language*, 88:817–845.

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244.

Clements, G. N. (2009). The Role of Features in Phonological Inventories. In Eric Raimy et al., editors, *Contemporary Views on Architecture and Representations in Phonology*, pages 19–68. MIT Press.

Coupé, C., Marsico, E., and Philippson, G. (2011). How economical are phonological inventories? In *Proceedings of the 17th International Congress of Phonetic Sciences*.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Fortson, B. W. (2004). *Indo-European Language and Culture: An Introduction*. Number 13 in Blackwell Textbooks in Linguistics. Blackwell, Oxford.

Gray, R. D., Drummond, A. J., and Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, 323(5913):479–483.

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., and Pagel, M. (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, 112(43):13296–13301.

Hammarström, H., Forkel, R., and Haspelmath, M. (2017). Glottolog 3.0. Online: `http://glottolog.org`.

Hruschka, D. J., Branford, S., Smith, E. D., Wilkins, J., Meade, A., Pagel, M., and Bhattacharya, T. (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, 25(1):1–9.

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

Kavitskaya, D. (2006). Perceptual salience and palatalization in russian. *Laboratory Phonology*, 8:589–610.

List, J.-M. and Moran, S. (2013). An open source toolkit for quantitative historical linguistics. In *ACL (Conference System Demonstrations)*, pages 13–18.

List, J.-M. (2017). A Web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pages 9–12, Valencia. Association for Computational Linguistics.

Maddieson, I. (1984). *Pattern of Sounds*. Cambridge University Press, Cambridge, UK.

Maddieson, I. (2013a). Consonant inventories. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Maddieson, I. (2013b). Vowel quality inventories. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Marsico, E. (1999). What can a database of proto-languages tell us about the last 10,000 years of sound changes? In *Proceedings of the XIVth International Congress of Phonetic Sciences*, pages 353–356, San Francisco.

Meade, A. and Pagel, M. (2014). Bayestraits v2. Online: `http://www.evolution.rdg.ac.uk/BayesTraitsV3/BayesTraitsV3.html`.

Michael, L., Chousou-Polydouri, N., Keith, B., Donnelly, E., Meira, S., Wauters, V., and O'Hagan, Z. (2015). A bayesian phylogenetic classification of Tupí-Guaraní. *LIAMES - Línguas Indígenas Americanas*, 15:193–221.

Moran, S. and Cysouw, M. (In press). *The Unicode Cookbook for Linguists*. Language Science Press.

Moran, S., McCloy, D., and Wright, R. (2014). *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Moran, S. (2012). *Phonetics Information Base and Lexicon*. Ph.D. thesis, University of Washington.

Nichols, J. (1992). *Linguistic diversity in space and time*. University of Chicago Press.

Pagel, M., Atkinson, Q. D., Calude, A. S., and Meade, A. (2013). Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*, 110(21):8471–8476.

Pagel, M. (1997). Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, 26(4):331–348.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877.

R Core Team, (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Steiner, L., Stadler, P. F., and Cysouw, M. (2011). A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1):89–127.

Walker, R. S. and Ribeiro, L. A. (2011). Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1718):2562–2567.

Wichmann, S., Holman, E. W., and Brown, C. H. (2017). The ASJP Database (version 17). Online: `http://asjp.clld.org/`.