JAMBU: A historical linguistic database for South Asian languages

Aryaman Arora
Georgetown University
aa2190@georgetown.edu

Adam Farris
Stanford University
adfarris@stanford.edu

Samopriya Basu
Simon Fraser University
samopriya_basu@sfu.ca

Suresh Kolichala
Microsoft
suresh.kolichala@gmail.com

Abstract

We introduce JAMBU, a cognate database of South Asian languages which unifies dozens of previous sources in a structured and accessible format. The database includes 287k lemmata from 602 lects, grouped together in 23k sets of cognates. We outline the data wrangling necessary to compile the dataset and train neural models for reflex prediction on the Indo-Aryan subset of the data. We hope that JAMBU is an invaluable resource for all historical linguists and Indologists, and look towards further improvement and expansion of the database.[1]

1 Introduction

A particular concern of historical linguists is studying relatedness and contact between languages. Two languages are related if they share words the arose from a common source, having undergone (potentially different) regular sound changes.[2] For example, the German words *schlafen* and *Schiff* are cognate to the English words *sleep* and *ship* respectively, with the German words having undergone the sound change /p/ → /f/. Using evidence like this from all of the Germanic languages, historical linguists have reconstructed the historical words that gave rise to these terms: **slāpan* and **skipą* (Kroonen, 2013).

Computational historical linguistics is a growing field that seeks to apply modern computational methods to studying this kind of change (Jäger, 2019; List, 2023). Massive datasets of multilingual cognates are necessary for much of the current research in this area, e.g. on multilingual cognate detection and phoneme-level alignment (List et al., 2018) and automatic comparative reconstruction of historical ancestors of languages (Ciobanu and Dinu, 2018).

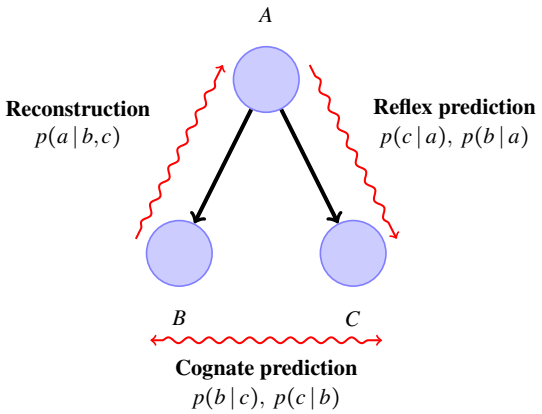

Figure 1: Three tasks of interest in computational historical linguistics. In this diagram, *A* is the ancestor language of *B* and *C*.

South Asia[3] as a region is home to a complex historical mesh of language contact and change, especially between the Indo-Aryan and Dravidian language families (Masica, 1976). Yet, South Asia is relatively understudied by linguists compared to the linguistic diversity of the region (Arora et al., 2022). There is no cross-family lexical dataset to facilitate computational study on South Asian historical and contact linguistics. In order to improve this unfortunate state of affairs, we introduce the **JAMBU** cognate database for South Asian languages. JAMBU includes all cognacy information from the major printed etymological dictionaries for the Indo-Aryan (Turner, 1962–1966) and Dravidian (Burrow and Emeneau, 1984) languages, as well as data from several more recent sources. In this paper, we introduce and analyse our database and train neural models on the reflex prediction task. We hope that this resource brings us closer to the ultimate goal of understanding how the lan-

[1] The entire dataset is available at https://github.com/moli-mandala/data, and a web interface for browsing it is at https://neojambu.herokuapp.com/.

[2] Per the Neogrammarian hypothesis, sound changes are regular and *exceptionless* (Osthoff and Brugmann, 1878; Paul, 1880). The reality of sound change is sometimes less ideal.

[3] When using the term *South Asia* we refer to the Indian Subcontinent.

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 68–77

guages of South Asia have evolved and interacted over time.

2 Related work

CLDF format. CLDF was proposed by Forkel et al. (2018) as a standard, yet highly flexible, format for linguistic data (including cognate databases, etymological dictionaries with reconstructions, and even dictionaries). We use this format for the JAMBU database. Many etymological databases use CLDF to effectively encode complex relations (e.g. loaning) and metadata (e.g. references, phonetic forms, alignments). Some which informed our database design were Rankin et al. (2015); Greenhill et al. (2008).

Cognates. Batsuren et al. (2019) compiled a *cognate database* covering 338 languages from Wiktionary. They noted that the meaning of *cognate* varies between research communities—for our purposes as historical linguists, we prefer grouping terms with shared direct etymological sources, while much computational work (e.g. Kondrak et al., 2003) takes a broader definition which includes loanwords or even all semantic equivalents as cognates.

As shown in figure 1, computational historical linguistics has taken on tasks involving cognates such as automatic *cognate identification* from wordlists (Rama et al., 2018; List et al., 2018; Rama, 2016), *cognate/reflex prediction*, i.e. predicting the form of a cognate in another language based on concurrent or historical data (List et al., 2022; Bodt and List, 2022; Fourrier et al., 2021; Marr and Mortensen, 2020), and *reconstruction* of the ancestor form of a cognate set (Durham and Rogers, 1969; Bouchard et al., 2007; Ciobanu and Dinu, 2018; Meloni et al., 2021; He et al., 2022, *inter alia*).

Other South Asian cognate databases. Cathcart (2019a,b, 2020) and Cathcart and Rama (2020) also previously made use of data from Turner (1962–1966) by scraping the version hosted online by *Digitial Dictionaries of South Asia*.

There was an effort to create a new digital South Asian etymological dictionary in the early 2000s, termed the **SARVA** (South Asian Residual Vocabulary Assemblage) project (Southworth, 2005a). This was unsuccessful however, and only a small portion of the possible cross-family entries were complete. Our database does not incorporate it.

	Languages	**Cognate sets**	**Lemmata**
Indo-Aryan	433	16,464	194,834
Dravidian	78	5,649	78,502
Nuristani	22	3,645	12,088
Other	52	163	311
Munda	15	129	1,352
Burushaski	2	41	48
Total	602	23,024	287,135

Table 1: Statistics about the JAMBU database, factored by language family. **Cognate sets** counts the number of such sets that include at least one cognate from that family (and so does not sum to the total).

3 Database

The JAMBU database incorporates data from three major language families of South Asia: Indo-European (the Indo-Aryan and Nuristani subbranches), Dravidian, and Austroasiatic (the Munda subbranch). This comes out to 287k lemmata from 602 lects across 23k cognate sets (table 1).

The data is stored in the CLDF structured data format. The overall database structure is described in the file `Wordlist-metadata.json`, which includes information about the type of data recorded in each column of each file. The file `forms.csv` includes all lemmata (word form) and associated etymological and linguistic information. The files `parameters.csv` and `cognates.csv` include all cognateset headwords and etymological notes for each. The file `languages.csv` lists all languages in the database and their geographical location. Finally, `sources.bib` lists all data references in BibTeX format.

For each lemma in `forms.csv`, we store the following information: a unique *form ID*; the *language ID*; the *cognate set ID*, linking it to other cognate lemmata; a *normalised* representation of the lemma itself, using our transcription scheme; a *gloss* in English; the spelling of the lemma in the *native script*; the phonemic *IPA* representation of the lemma; the *unnormalised* form of the lemma taken from the original source; a finer-grained *cognate set ID*; *notes*; and *references*.

For each cognate set, we store a **headword**, which is usually a common ancestor of the words in that cognateset or a reconstruction of that ancestor if possible. We also store desiderata such as definitions and etymological notes.

Finally, we take an expansive view of what constitutes a "language" in our database. If a word is

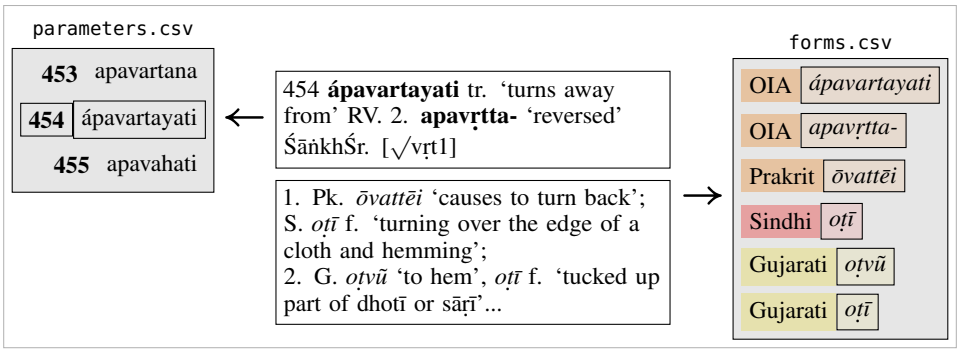

Figure 2: Diagram of some of the data in JAMBU parsed from CDIAL entry 454 (*ápavartayati*, 'turns away from').

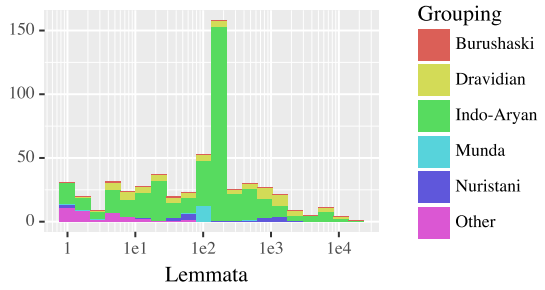

Figure 3: Distribution of languages by number of lemmata entered in JAMBU.

known to only be attested in a particular dialect, we list that dialect separately. For example, for the Shina language (northwestern Indo-Aryan), we list 32 geographical dialects. The distribution of languages by number of lemmata is depicted in figure 3.

3.1 Data sources and scraping

The two major data sources are CDIAL (Turner, 1962–1966) and DEDR (Burrow and Emeneau, 1984), which have been scraped in their entirety from web versions hosted by the University of Chicago's Digital Dictionaries of South Asia project.[4] Since the raw data is in HTML with limited structured markup, extracting CLDF-suitable data is a significant hurdle, including matching lemmata to the appropriate language and grouping associated metadata like grammatical gender and etymological notes under the correct form (figure 2). Further cleanup of data from these two sources will have to be done manually.

Since CDIAL and DEDR have not been updated in decades, we are also incorporating more recent sources that refer to them into our database, as well as etymologising newer fieldwork data manually. The additional sources we added (some partially) are listed in appendix B.

3.2 Transcriptions

One serious issue has been reconciling differing transcription systems from different sources; transcription schemes vary across sources even for the same language, since there is no standard transcription for South Asian linguistics. An illustrative example of this issue is the variable transcription of the labiodental fricative as *v* or *w*.

Turner (1962–1966) normalises entries from various sources into a relatively mundane Indological transcription, i.e., IAST[5] with many extensions for the varying phonologies of South Asian languages, but not always consistently. For example, the phoneme /eː/ is notated ⟨ē⟩ for Sanskrit entries, but ⟨e⟩ for Hindi (and in Burrow and Emeneau (1984), as ⟨é⟩ for Malto entries). Elsewhere, e.g., in Bengali and Punjabi, transcriptions adhere to the written form, which do not always adhere to any phonemic analysis of the languages in question. In the case of Kashmiri, Shina, and many other languages, there are now better analyses to base romanisation on than existed at the time of compilation of the sources of Turner (1962–1966). Meanwhile, (Burrow and Emeneau, 1984) does not attempt to conventionalise transcription at all, instead strictly copying the transcription from the original source; e.g. all Bengali entries strictly reflect spelling and do not indicate the differing surface realisations of the orthographic schwa (Johny and Jansche, 2018).

We created a new, more rigidly standardised transcription system based on Indological conventions to unify all our data. We did not want to use pure IPA because it obscures useful cross-lingual pat-

[4] https://dsal.uchicago.edu/dictionaries/

[5] https://en.wikipedia.org/wiki/International_Alphabet_of_Sanskrit_Transliteration

Language	Original	Normalised
Old Indo-Aryan	*anugr̥bhāyati	*anugr̥b^{h}āyati
European Romani	učhar	uc^{h}ar
Shumashti	ášin	ǽśin
Palula	beedhríi	bēd^{h}rî
Pashai: Degano	dew'âz	devǻz

Table 2: Examples showing how our orthographic normalisation process affected forms from various sources.

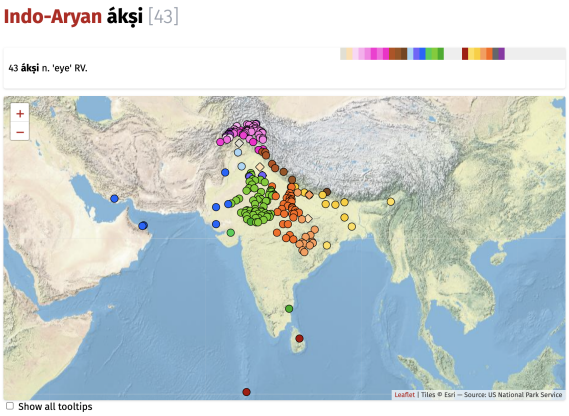

Figure 4: Web interface for Jambu, displaying reflexes of CDIAL entry 43 (*ákṣi*, 'eye'). See https://neojambu.herokuapp.com/entries/43.

terns[6] and is not conventional in the Indological research community (especially considering that the database may be of use to non-linguist Indologists as well). For that reason, we use a modified IAST (for instance, using a superscript ⟨^{h}⟩ to notate aspiration and breathy voice distinguishing these from genuine h-clusters) to suit cross-linguistic needs. Some contrasts are made more explicit while notational consistency is maintained across the board.

We used the `segments` Python library to create orthography normalisation profiles for each source's transcription scheme (Moran and Cysouw, 2018); some examples of the changes are shown in table 2. So far, forms from all source have not yet been orthographically standardised to our system. However, we developed standardisation scripts covering 204k lemmata, of which 99.7% were automatically converted without errors.

3.3 Web interface

Finally, we developed a web interface for the JAMBU database; see figure 4. Originally, we used the pre-existing `clld` webapp toolkit for the publication of Cross-Linguistic Linked Data,[7] but we later switched to a custom Flask web app designed from scratch in order to have finer control over the database structure and to execute searches on the backend more efficiently. This web interface supports search, filtering, and geographical visualisation. We hope this supersedes the unstructured search interfaces currently available for browsing older etymological dictionaries for these languages (Turner, 1962–1966; Burrow and Emeneau, 1984).

Model	Perplexity	BLEU	TER
GRU	2.57	55.91	**34.40**
Transformer	**2.53**	**56.03**	35.15

Table 3: Performance of the two models on reflex prediction on the Indo-Aryan segment of JAMBU.

4 Experiment

As a demonstration of the usability of the dataset for computational historical linguistics, we replicate the reflex prediction task of Cathcart and Rama (2020). We train neural models on the task of reflex prediction in Indo-Aryan languages, i.e. predicting the descendant of an Old Indo-Aryan word in a given Indo-Aryan language. Rather than being restricted to data from Turner (1962–1966), we can draw on all the sources present in JAMBU.

We train on 80% of the data and test on the remaining 20%. We compare two models: a bidirectional GRU encoder-decoder with Bahdanau attention and a transformer encoder-decoder with learned positional embeddings. The optimised hyperparameters for the GRU are a learning rate of $2 \cdot 10^{-3}$, 4 layers, and embedding and hidden size of 64. The transformer had a learning rate of 1 (using the parameter-based adjustment and warmup/decay schedule from Huang et al., 2022), 3 layers, 4 attention heads per layer, embedding size of 64, and FFN size of 128. Both models were trained for 50 epochs without early stopping with a batch size of 1024 on a single Quadro RTX 6000, with a run completing in about 15 minutes.

We evaluate BLEU and TER on the held-out set using the SacreBLEU implementation (Post, 2018), treating a single phoneme as a 'word'. Even after comprehensive hyperparameter tuning we find that both models achieve similar performances, per the results in table 3. We leave analysis of these models for future work.

[6] E.g. the Indological *a* (called a schwa) varies in pronunciation across South Asia, from [a̲] (Telugu) to [ɜ] (Hindi) to [ɔ~o] (Bengali) to [ʌ] (Nepali).

[7] https://github.com/clld/clld

5 Conclusion

In this paper, we introduced JAMBU, the largest and most up-to-date cognate database for South Asian languages. We are continuing to expand the database, incorporating all lexical data that has so far been unused in comparative linguistic work on the region. We believe that the open questions of South Asian historical linguistics cannot be resolved without examining all the information (both synchronic and diachronic) that linguists have collected about language of the region. The old etymological dictionaries are in desparate need of an update. However, much work remains. We briefly discuss some avenues of future work.

Many sources are yet to be incorporated, especially those recording loanwords from external languages (especially Persian, Arabic, English, and Portuguese) and from local literary languages (particularly Sanskrit). We have yet to distentangle cross-lectal interactions and mark lexical isoglosses, which seem necessary to understand the history of language interactions in the region; Kalyan et al. (2018)'s wave model of linguistic change has been thought by many scholars to be suited for South Asian languages, but it has not been operationalised yet due to a lack of comprehensive data (Toulmin, 2006; Kogan, 2017).

Another significant task ahead is extending our database structure to support indicating and analysing more complex cross-lingual interactions. For example, the database as it stands does not distinguish between inheritance from the parent language and loaning mediated by a sibling language.

We have also been working on a consistent orthography for tonemes in the languages where tones are contrastive, such as the northwestern Indo-Aryan languages (Baart, 2014). Older data from these languages either does not notate tone at all (for tonality was not yet recognized, as in Gawri and Torwali), or represents it indirectly through diachronically correct, but synchronically confusing, spelling systems (as in Punjabi and Kishtwari). So, our work will also involve analyzing and incorporating new data from tonal languages, both from existing sources and our own fieldwork.

Finally, we hope to manually improve data quality once the parsing of old sources is stable. This includes fixing known mistakes, reorganising entries to better indicate indirect derivations and cross-lectal loans, and etymological notes that summarise the extant literature.

References

Binny Abraham, Binoy Koshy, and Vimal Raj R. 2012. Sociolinguistic survey of selected Rajasthani speech varieties of Rajasthan, India, volume 2: Mewari. *Journal of Language Survey Reports*.

Said Al Jahdhami. 2017. Zadjali: The dying language. *International Journal of Language and Linguistics*, 4.

Said Humaid Al Jahdhami. 2022. Maimani language and Lawati language: Two sides of the same coin? *Journal of Modern Languages*, 32(1):37–57.

Aryaman Arora and Ahmed Etebari. 2020–2021. Kholosi dictionary.

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2022. Computational historical linguistics and language diversity in South Asia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1396–1409, Dublin, Ireland. Association for Computational Linguistics.

Joan Baart. 2014. Tone and stress in north-west Indo-Aryan. *Above and beyond the segments: Experimental linguistics and Phonetics, Amsterdam: John Benjamins*, pages 1–13.

Joan L. G. Baart. 1997. *The sounds and tones of Kalam Kohistani: with wordlist and texts*. National Institute of Pakistan Studies, Quaid-i-Azam University and Summer Institute of Linguistics, Islamabad.

Peter C. Backstrom and Carla F. Radloff. 1992. *Sociolinguistic Survey of Northern Pakistan, Volume 2. Languages of Northern Areas*. National Institute of Pakistan Studies, Islamabad.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2019. CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy. Association for Computational Linguistics.

Theodor Gipson Benjamin and Liahey Ngwazah. 2012. Sociolinguistic survey of selected Rajasthani speech varieties of Rajasthan, India, volume 5: Dhundari and Shekhawati. *Journal of Language Survey Reports*.

Hermann Berger. 1998. *Die Burushaski-Sprache von Hunza und Nager*. Harrassowitz.

Murli D. Bhawnani. 1979. *Descriptive analysis of Thari: A dialect of Sindhi language*. Ph.D. thesis, Deccan College Post Graduate and Research Institute Pune, Pune.

Timotheus A. Bodt and Johann-Mattis List. 2022. Reflex prediction: A case study of Western Kho-Bwa. *Diachronica*, 39(1):1–38.

Ed Boehm. 2017. *A Sociolinguistic Profile of Bundeli*. Journal of Language Survey Reports. SIL International, Dallas, Texas.

Edward Daniel Boehm. 1998. A phonological reconstruction of Proto-Tharu. Master's thesis, The University of Texas at Arlington.

Kelly Kilgo Boehm. 2002. *A Preliminary Sociolinguistic Survey of the Chhattisgarhi-Speaking Peoples of India*. SIL International, Dallas, Texas.

Alexandre Bouchard, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896, Prague, Czech Republic. Association for Computational Linguistics.

Thomas Burrow and Murray Barnson Emeneau. 1984. *A Dravidian Etymological Dictionary*, 2 edition. Clarendon Press, Oxford.

Chundra Cathcart. 2019a. Gaussian process models of sound change in Indo-Aryan dialectology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 254–264, Florence, Italy. Association for Computational Linguistics.

Chundra Cathcart. 2019b. Toward a deep dialectological representation of Indo-Aryan. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 110–119, Ann Arbor, Michigan. Association for Computational Linguistics.

Chundra Cathcart. 2020. A probabilistic assessment of the Indo-Aryan Inner–Outer Hypothesis. *Journal of Historical Linguistics*, 10(1):42–86.

Chundra Cathcart and Taraka Rama. 2020. Disentangling dialects: a neural approach to Indo-Aryan historical phonology and subgrouping. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 620–630, Online. Association for Computational Linguistics.

Sajayan Chacko and Liahey Ngwazah. 2012. Sociolinguistic survey of selected Rajasthani speech varieties of Rajasthan, India, volume 6: Marwari, Merwari, and Godwari. *Journal of Language Survey Reports*.

Alina Maria Ciobanu and Liviu P. Dinu. 2018. Ab initio: Automatic Latin proto-word reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1604–1614, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stanton P. Durham and David Ellis Rogers. 1969. An application of computer programming to the reconstruction of a proto-language. In *International Conference on Computational Linguistics COLING 1969: Preprint No. 5*, Sånga Säby, Sweden.

Josef Elfenbein. 1994. Notes on Khetrāni phonology. *Studien zur Indologie und Iranistik*, 19:71–82.

M. B. Emeneau and T. Burrow. 1962. *Dravidian Borrowings from Indo-Aryan*. Number 26 in University of California Publications in Linguistics. University of California Press, Berkeley.

Robert Forkel, 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(1):1–10.

Clémentine Fourrier, Rachel Bawden, and Benoît Sagot. 2021. Can cognate prediction be modelled as a low-resource machine translation task? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 847–861, Online. Association for Computational Linguistics.

Sonja Fritz. 2002. *The Dhivehi language: a descriptive and historical grammar of Maldivian and its dialects*. Ergon-Verlag.

Harjeet Singh Gill. 1973. *Linguistic atlas of the Punjab*. Munshiram Manoharlal Publishers, Delhi.

Simon J. Greenhill, Robert Blust, and Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:EBO.S893. PMID: 19204825.

Andre He, Nicholas Tomlin, and Dan Klein. 2022. Neural unsupervised reconstruction of protolanguage word forms. *arXiv*, abs/2211.08684.

Austin Huang, Suraj Subramanian, Jonathan Sum, Khalid Almubarak, and Stella Biderman. 2022. The annotated transformer.

Mathews John and Bezily P. Varghese. 2021. The Kannauji-speaking people of Uttar Pradesh: A sociolinguistic profile. *Journal of Language Survey Reports*.

Cibu C Johny and Martin Jansche. 2018. Brahmic schwa-deletion with neural classifiers: Experiments with bengali. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 259–263.

Thomas Jouanne. 2014. *A Preliminary Analysis of the Phonological System of the Western Pahāṛī Language of Kvār*. Ph.D. thesis, University of Oslo, Oslo.

Gerhard Jäger. 2019. Computational historical linguistics. *Theoretical Linguistics*, 45(3–4):151–182.

Siva Kalyan, Alexandre François, et al. 2018. Freeing the comparative method from the tree model: A framework for historical glottometry. *Senri Ethnological Studies*, 98:59–89.

Masato Kobayashi. 2022. Proto-Dravidian origins of the Kuṛux-Malto past stems. *Bhasha. Journal of South Asian Linguistics, Philology and Grammatical Traditions*, 1(2):263–282.

Anton I Kogan. 2017. Genealogical classification of New Indo-Aryan languages and lexicostatistics. *Journal of Language Relationship*, 14(3–4):227–258.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 46–48.

Binoy Koshy. 2012. Sociolinguistic survey of selected Rajasthani speech varieties of Rajasthan, India, volume 3: Hadothi. *Journal of Language Survey Reports*.

Guus Kroonen. 2013. *Etymological Dictionary of Proto-Germanic*. Leiden Indo-European Etymological Dictionary Series; 11. Brill, Leiden, Boston.

Henrik Liljegren. 2013. Notes on Kalkoti: A Shina language with strong Kohistani influences. *Linguistic Discovery*, 11(1):129–160.

Henrik Liljegren. 2019. Palula dictionary. *Dictionaria*, (3):1–2700.

Johann-Mattis List. 2023. Computational historical linguistics.

Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan Hill, and Ryan Cotterell. 2022. The SIGTYP 2022 shared task on the prediction of cognate reflexes. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 52–62, Seattle, Washington. Association for Computational Linguistics.

Johann-Mattis List, Mary Walworth, Simon J Greenhill, Tiago Tresoldi, and Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, 3(2):130–144.

Clayton Marr and David R. Mortensen. 2020. Computerized forward reconstruction for analysis in diachronic phonology, and Latin to French reflex prediction. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 28–36, Marseille, France. European Language Resources Association (ELRA).

Colin P. Masica. 1976. *Defining a Linguistic Area: South Asia*. University of Chicago Press.

Eldose K. Mathai. 2011. Bagri of Rajasthan, Punjab, and Haryana: A sociolinguistic survey. *Journal of Language Survey Reports*.

Eldose K. Mathai. 2012. Sociolinguistic survey of selected Rajasthani speech varieties of Rajasthan, India, volume 4: Mewati. *Journal of Language Survey Reports*.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, Online. Association for Computational Linguistics.

Steven Moran and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press.

Ram Dayal Munda. 1968. Proto-Kherwarian phonology. Master's thesis, University of Chicago.

Hermann Osthoff and Karl Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Hirzel, Leipzig, Germany.

Hukam Chand Patyal. 1982. Etymological notes on some Maṇḍyāḷī words (Indo-Aryan Studies II). *Indo-Iranian Journal*, 24:289–294.

Hukam Chand Patyal. 1983. Etymological notes on some Maṇḍyāḷī words (Indo-Aryan Studies IV). *Indo-Iranian Journal*, 25:41–49.

Hukam Chand Patyal. 1984. Etymological notes on some Maṇḍyāḷī words (Indo-Aryan Studies V). *Indo-Iranian Journal*, 27:121–132.

Hukam Chand Patyal. 1991. Etymological notes on some Ḍogri words (Indo-Aryan Studies III). *Indo-Iranian Journal*, 34:123–124.

Hermann Paul. 1880. *Prinzipien der Sprachgeschichte*. Max Niemeyer, Halle, Germany.

Martin Pfeiffer. 2018. *Kurux Historical Phonology Reconsidered: With a Reconstruction of Pre-Kurux-Malto Phonology*. PubliQation.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027, Osaka, Japan. The COLING 2016 Organizing Committee.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400, New Orleans, Louisiana. Association for Computational Linguistics.

Robert L. Rankin, Richard T. Carter, A. Wesley Jones, John E. Koontz, David S. Rood, and Iren Hartmann, editors. 2015. *Comparative Siouan Dictionary*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Felix Rau. 2019. Munda cognate set with proto-munda reconstructions.

Khawaja A. Rehman and Joan L. G. Baart. 2005. A first look at the language of Kundal Shahi in Azad Kashmir. *SIL Electronic Working Papers*.

Ruth Laila Schmidt and Vijay Kumar Kaul. 2008. A comparative analysis of Shina and Kashmiri vocabularies. *Acta Orientalia*, 69:231–302.

Christopher Shackle. 1995. *A Guru Nanak Glossary*. Routledge.

F. C. Southworth. 2005a. The SARVA (South Asia Residual Vocabulary Assemblage) project.

Franklin C. Southworth. 2005b. Prehistoric implications of the Dravidian element in the NIA lexicon with special reference to Marathi. *International Journal of Dravidian Linguistics*, 34(1):17–28.

Franklin C Southworth. 2006. Proto-Dravidian agriculture. In *Proceedings of the Pre-symposium of Rihn and 7th ESCA Harvard-Kyoto Roundtable. Research Institute for Humanity and Nature, Kyoto*, pages 121–150.

Richard F. Strand. 1997–2021. Nuristân: Hidden land of the Hindu-Kush.

Matthew William Stirling Toulmin. 2006. *Reconstructing linguistic history in a dialect continuum: The Kamta, Rajbanshi, and Northern Deshi Bangla subgroup of Indo-Aryan*. Ph.D. thesis, The Australian National University.

Ralph Lilley Turner. 1962–1966. *A comparative dictionary of the Indo-Aryan languages*. Oxford University Press, London.

Govindaswamy Srinivasa Varma. 1970. *Vaagri boli, an Indo-Aryan language*. Ph.D. thesis, Annamalai University.

Stephen Watters. 2013. A sociolinguistic profile of the Bhils of northern Dhule district. *Journal of Language Survey Reports*.

Claus Peter Zoller. 2005. *A grammar and dictionary of Indus Kohistani: Dictionary*, volume 1. Walter de Gruyter.

Saeed Zubair. 2016. A phonological description of Wadiyari, a language spoken in Pakistan. Master's thesis, Payap University, Chiang Mai, Thailand.

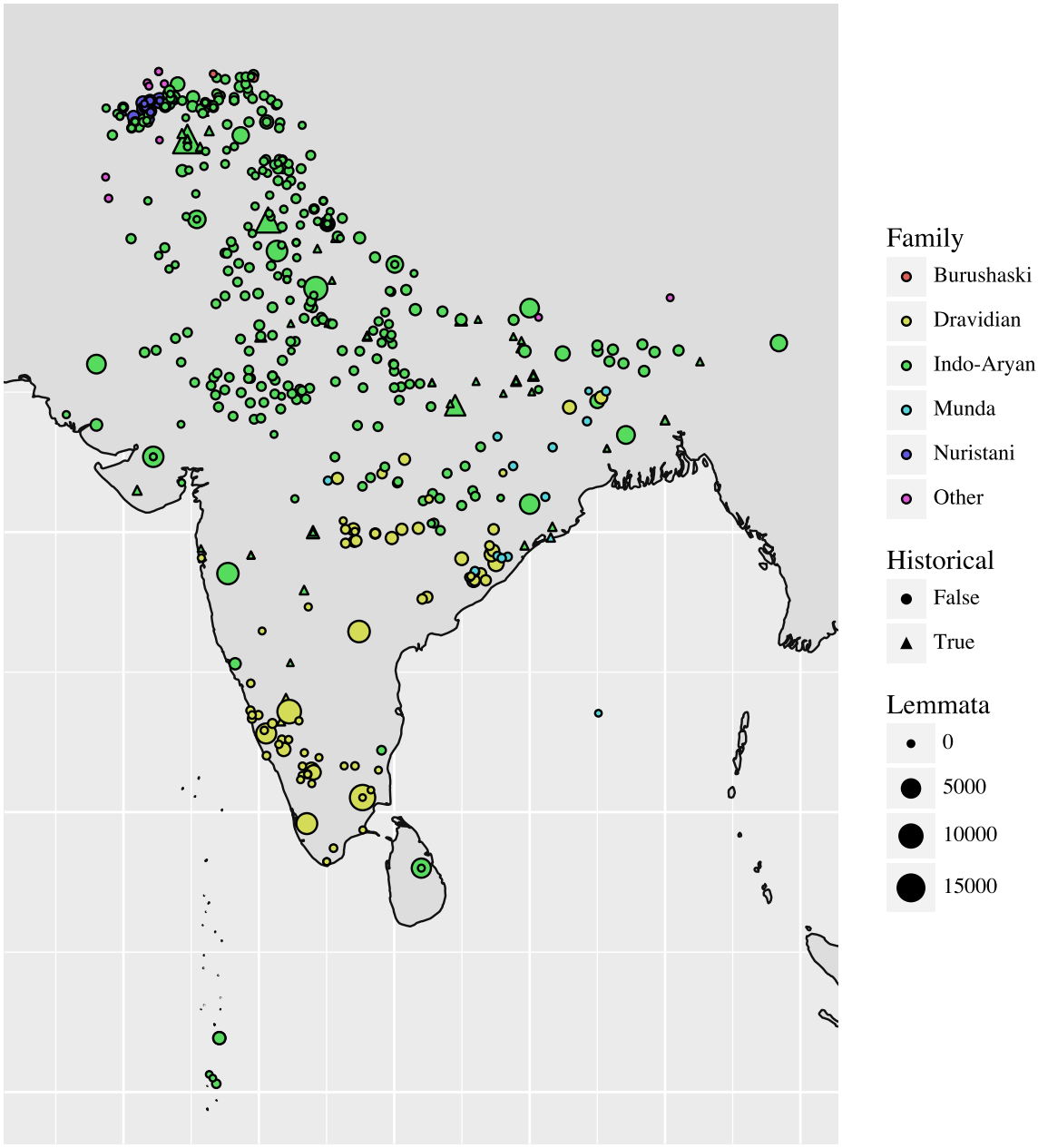

Figure 5: Map of South Asian languages present in JAMBU, coloured by phylogenetic grouping and sized by number of lemmata included in the database. 74 lects (mostly varieties of Romani, an Indo-Aryan language, spoken in Europe and the Middle East) are not visible within the bounds of this map.

A Licensing

Data from Burrow and Emeneau (1984) and Turner (1962–1966) has been scraped using the approval of the SARVA project (of which one of the authors was previously involved in) for strictly academic purposes. Additional data added to the dataset has either been manually etymologised (and therefore is an original academic contribution) or obtained with permission of the respective authors.

B Other data sources

Language(s)	Reference	Etymologised?	In JAMBU?
Burushaski	Berger (1998)	✓	†
Dravidian	Burrow and Emeneau (DEDR; 1984)	✓	✓
	Emeneau and Burrow (DBIA; 1962)	✓	
	Southworth (2006)	✓	✓
	Southworth (2005b)	✓	✓
Kurux, Malto	Kobayashi (2022)	✓	†
	Pfeiffer (2018)	✓	†
Indo-Aryan	Turner (CDIAL; 1962–1966)	✓	✓
Bagri	Mathai (2011)		✓
Bhil	Watters (2013)		
Bundeli	Boehm (2017)		✓
Chhattisgarhi	Boehm (2002)		✓
Dhivehi	Fritz (2002)	✓	✓
Dogri	Patyal (1991)	✓	✓
Gawri	Baart (1997)		✓
Indus Kohistani	Zoller (2005)	✓	
Kalkoti	Liljegren (2013)		✓
Kamtapuri, etc.	Toulmin (2006)	✓	✓
Kannauji	John and Varghese (2021)		†
Khetrani	Elfenbein (1994)		✓
Kholosi	Arora and Etebari (2020–2021)	✓	✓
Kundal Shahi	Rehman and Baart (2005)		✓
Kvari	Jouanne (2014)		✓
Maimani, Luwati	Al Jahdhami (2022)		†
Mandeali	Patyal (1982, 1983, 1984)	✓	✓
Palula	Liljegren (2019)	✓	✓
Punjabi, etc.	Gill (1973)		†
	Shackle (1995)	✓	†
Rajasthani	Abraham et al. (2012)		✓
	Benjamin and Ngwazah (2012)		✓
	Chacko and Ngwazah (2012)		✓
	Koshy (2012)		✓
	Mathai (2012)		✓
Shina, Domaaki	Backstrom and Radloff (1992)		✓
Shina, Kashmiri	Schmidt and Kaul (2008)		†
Thari	Bhawnani (1979)		†
Tharu	Boehm (1998)		✓
Vaagri Boli	Varma (1970)	✓	†
Wadiyara Koli	Zubair (2016)		†
Zadjali	Al Jahdhami (2017)		✓
Munda	Rau (2019)	✓	✓
	Munda (1968)	✓	
Nuristani	Strand (1997–2021)	✓	✓

Table 4: All sources included in JAMBU, grouped together by language and family. **Etymologised?** indicates whether the original sources provided etymologies for the terms it listed; if not, we manually proposed etymologies. **In JAMBU?** indicates what portion of the work has been incorporated into the current version of the database; ✓ means entirely while † means partially.