Bhāṣācitra: Visualising the dialect geography of South Asia

Aryaman Arora
Georgetown University
aa2190@georgetown.edu

Adam Farris
San Mateo High School
adamfarris@gmail.com

Gopalakrishnan R
EFL University, Hyderabad
gopalkrishnan11251@gmail.com

Samopriya Basu
University of North Carolina – Chapel Hill
sampr0b@live.unc.edu

Abstract

We present Bhāṣācitra,[1] a dialect mapping system for South Asia built on a database of linguistic studies of languages of the region annotated for topic and location data. We analyse language coverage and look towards applications to typology by visualising example datasets. The application is not only meant to be useful for feature mapping, but also serves as a new kind of interactive bibliography for linguists of South Asian languages.

1 Introduction

South Asia is extremely linguistically diverse. There is a common saying illustrating this diversity, present in several languages of the region; it is given in Hindi below.

> *kos kos par pānī badle, cār kos par bānī.*
> 'The taste of water changes every mile,
> and the language every four.'

One issue with this vast scale of diversity is the difficulty it poses for linguists in collecting and cataloguing linguistic data, which further impedes comprehensive typological analysis. India alone contains known living speakers of 461 languages (Eberhard et al., 2021).[2] It is also difficult to assess the availability of linguistic literature for all of these languages, leading to gaps in the typological databases we end up compiling; print linguistic bibliographies for the region become outdated as new work is published and do not encode useful metadata, such as the specific dialect studied in each work or the linguistic features studied.

In this paper we present **Bhāṣācitra**, a database of linguistic sources for South Asian languages that we have compiled and annotated, as well as a dialect mapping and visualising system built from the location data extracted from those sources. Currently it includes 1104 labelled sources covering 311 lects. The site is online at http://aryamanarora.github.io/bhasacitra.

2 Background and related work

Dialects[3] are defined by *isoglosses*, geographical boundaries separating linguistic features. The mapping of dialect geography is a well-established problem in linguistics, and has been done for many languages; two illustrative examples are English (Orton et al., 1998; Kretzschmar, 2001) and Japanese (Kumagai, 2016). Dialect mapping is instrumentally important for the study of historical–comparative linguistics, since the present-day geography of isoglosses is a result of past *language change* and *language contact*. The distribution of synchronic features is data for theories of diachronic language change.

Computational approaches to dialect geography have worked on many parts of the issue, including the compilation of broad databases of linguistic features (Dryer and Haspelmath, 2013; Carling et al., 2018), dialect identification and clustering on modern social media corpora (Abdul-Mageed et al., 2018; Jones, 2015), and statistical modelling of dialect groups (e.g. Murawaki, 2020).

South Asia is a *linguistic area* (Masica, 1993; Bashir, 2016), a region of typological convergence due to historical contact between speakers of languages of different families. Families represented

[1] From Sanskrit *bhāṣā* 'language' + *citra* 'ornament, appearence'; lit. 'language map'.

[2] But note Asher (2008): "It is impossible to be at all precise about either the number of languages spoken in the region or the number of speakers of each."

[3] *Dialect* for the purposes of this paper refers to any speech variety. South Asia as a region is prone, due to geographical and historical factors, to fuzzy boundaries between speech varieties. The situation is best explained by Deo (2018) in describing the distribution of Indo-Aryan as "sociolinguistically rich and complex, characterized by plurilinguality and dialect continua spread over large regions spanning multiple languages".

Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021, pages 51–57

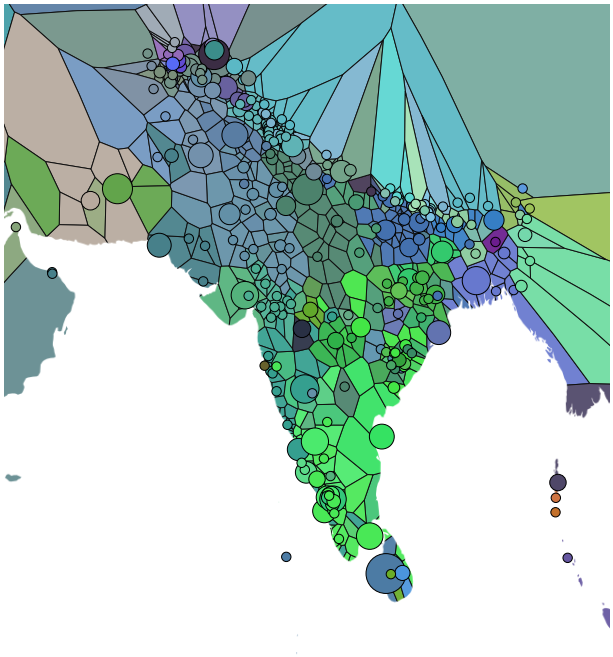


Figure 1: The primary interface map for Bhāṣācitra generated with D3.js and Voronoi partitioning.

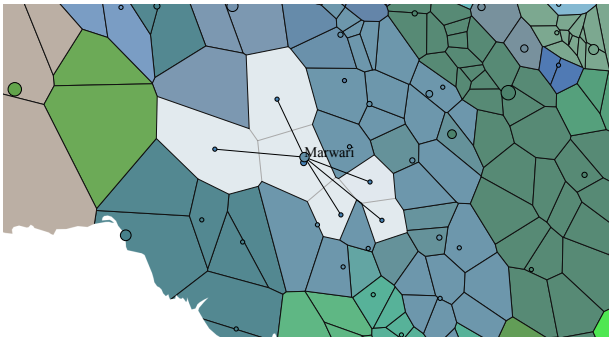


Figure 2: Hovering on the circle for Marwari (a language of Rajasthan, India) highlights the regions from which linguistic sources for it draw data.

in South Asia are Indo-European, Dravidian, Austroasiatic, Sino-Tibetan, and some unclassified isolates (Nihali, Kusunda, and Burushaski).

Visualisation of data for linguistic typology has a long history, beginning with the first lexical isogloss maps created by aggregating data from dialect surveys and with more recent work specifically for visualising historical change, such as Kalouli et al. (2019). As linguists adopt computational methods that deal with vast amounts of data, it becomes a challenge for humans to interpret datasets. Modern approaches to visualisation like Visual Analytics (VA) try to address this issue (Keim et al., 2008; MacEachren, 2017).

The use of point-based mapping in linguistic data visualisation is well-known, in e.g. WALS (Dryer and Haspelmath, 2013). This format has been used to map data in South Asian languages (Arsenault, 2017; Liljegren et al., 2021) as well as the languages of Iran (Anonby et al., 2019, 2018). We develop this paradigm further to map areal language extents based on the location data in published linguistic fieldwork.

3 Data model

We built Bhāṣācitra to be an easy-to-use system for researchers with no computational background. We implemented the application in JavaScript on a statically-hosted webpage. There are three data files in JSON format, for reference metadata (in BibTEX-compatible format with additional fields for location and topic information; see appendix A), language metadata (traditional genetic classification and coordinates for reference locations), and the typological database (containing per-language per-location data).

The primary interface is an interactive map displaying geographical points corresponding to locations from which language data has been collected. The map is generated and manipulated using the D3.js library which has a complete pipeline for web cartography (Bostock et al., 2011). Dialect zones are partitioned using the Voronoi algorithm; for a point P_k in the set of points P, its Voronoi region R_k is defined as all points closer to P_k than to any other point.

$$R_k = \left\{ x \in X \;\middle|\; \arg\min_i \left(\mathrm{dist}(P_i, x)\right) = k \right\} \quad (1)$$

In the primary interface (see figure 3), zones are colour-coded by consensus genetic classification of the languages covering the zone, with circles (with size proportional to the number of sources) centered at the weighted average of the coordinates of descriptions of the languages. In the case where multiple languages share a zone, the RGB components of the colouring are averaged.

3.1 Interface

The primary interface map is fully interactive (draggable and zoomable). Hovering over a language circle shows all the geographical points and Voronoi polygons associated with the sources compiled for that language (see figure 2). Like the language circles, each geographical point's size is weighted by the number of sources corresponding to it. Clicking on a language circle brings up the scrollable bibliography for that language, with each entry in

Topic	Count
overview (descriptive grammars)	494
syntax	141
phonetics/phonology	125
historical	111
morphology	100
sociolinguistics	91
lexicography	83
corpora	51
dialectology	48
comparative	44
Total	1104

Table 1: Count of sources labelled under the top 10 topics. A single source can be labelled with multiple topics.

human-readable format with the corresponding location and topic annotations appended.

3.2 Limitations

In South Asia (as elsewhere), geography is hardly the only variable encoding language use. As noted by Deo (2018) and shown in sociolinguistic studies (Gumperz, 1958) factors such as caste, social status, political affiliation, and religion play a large role in language use and adoption. Migrant speaker communities have also developed distinct dialects even in regions where they are a minority language group (e.g. Marathi speakers in Thanjavur and Burushaski speakers in Srinagar).

To deal with geographical overlap (different language sources for the same location), we allowed the areal zones of multiple languages to encompass the same location. A complete solution to the limitations of the geographical model would require collection of demographic data indexed to language use, which has not yet been collected on a large scale in South Asia.

4 Compiling the database

There are some existing bibliographies of language references for South Asia. In compiling data for Bhāṣācitra, we prioritised the incorporation of sources that provided the greatest coverage of language information, such as grammars and grammatical sketches, analysed corpora, and sociolinguistic surveys.

We began with data from Glottolog for broad coverage (Hammarström et al., 2020); South Asia-specific sources we drew from are Peterson (2018); Baart and Baart-Bremer (2001); Perera (2021). We then searched for literature not included in existing bibliographies. Many new sources were obtained

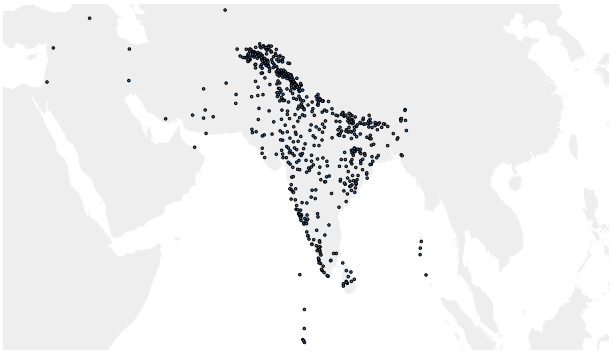


Figure 3: Map of all locations extracted from the sources in the Bhāṣācitra database.

from Shodhganga,[4] a platform for open-access digitised theses completed at Indian universities. These theses were difficult to access before the past decade, so from this resource we were able to incorporate many new references.

We annotated information on topic coverage for every source (see table 1) and location data (see §4.1) when possible. We also preferred to link to open-access versions of sources. In total, we compiled **1104 sources** describing **311 lects** with data collected from **763 locations**. This number is continually increasing as we actively improve our coverage of the linguistic literature and new work is published.

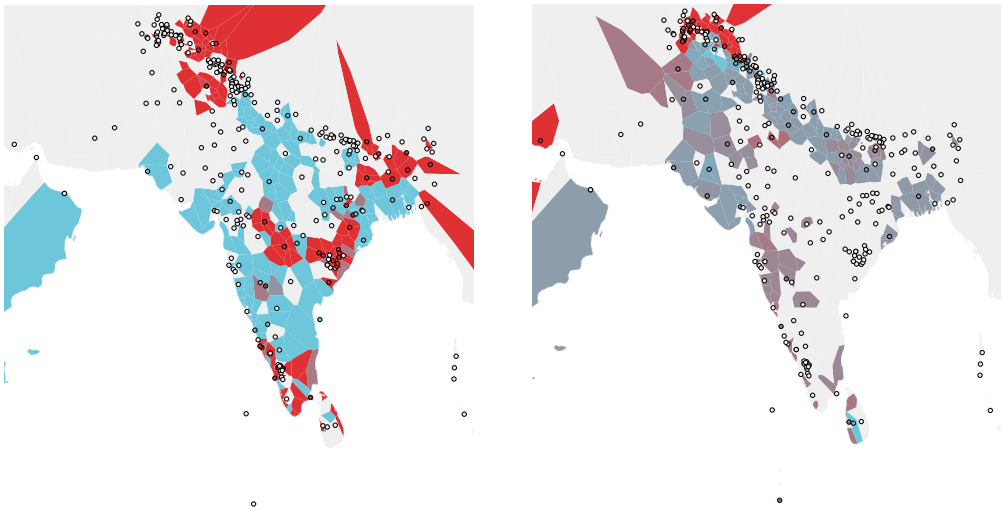
4.1 Locations

The primary new contribution of the Bhāṣācitra database is location data manually collected from the included references, shown in figure 3. The geocoding of the locations was done through the Google Maps API and manually verified.

While databases such as Glottolog and WALS do include location data for languages, their representation reduces the language's geographical distribution to a single point. We instead represent multiple points per language based on data from the sources we catalogued.

For example, in Glottolog, Hindi is placed at a single point in central India, whereas in Bhāṣācitra there are 21 locations associated with Hindi–Urdu, with most sources describing the standard dialect in Delhi, but also work dealing with varieties in Varanasi, Lahore, and the rural regions surrounding Delhi. Areal mapping of linguistic references allows for better assessment of the coverage of dialects in our sources, and for explicit coverage of dialect variation when mapping features.

[4] https://shodhganga.inflibnet.ac.in/



(a) Distribution of the breathy-voiced retroflex stop (/ɖʱ/) in South Asian languages.

(b) Percent of sound changes of Sanskrit /kṣ/ that result in /k(ː)ʰ/ in various Indo-Aryan languages.

Figure 4: Example datasets mapped in Bhāṣācitra. Scale: Yes/100%, No/0%

5 Mapping datasets

To illustrate the value of areal visualisation of language features, we mapped two datasets: the phoneme inventories of a large number of Indian languages from Ramaswami (1999), and the outcomes of selected sound changes from Sanskrit to the modern Indo-Aryan languages based on the Jambu database (Arora and Farris, 2021) parsed from Turner (1962–1966).

Note that we only visually analyse the map in these examples; these observations would need to be corroborated with statistic analysis and modelling to result in any verifiable claims.

5.1 Phoneme inventories

From the data in Ramaswami (1999) collected in the PHOIBLE database (Moran and McCloy, 2019) we were able to map the phoneme inventories of 62 major South Asian languages. Several works have studied the phonetic typology of the South Asian linguistic area, e.g. Ramanujan and Masica (2016); Arsenault (2017), but have not used areal mapping visualisations.

Some interesting phonological features for mapping are retroflexion (which is prevalent throughout the region, but weakly distinguished or not distinguished at all in the eastern periphery) and breathy-voiced stops (which are less common in much of the Dravidian and Munda families and in the northwestern languages). Figure 4a shows the distribution of the breathy-voiced retroflex stop /ɖʱ/ (in IAST: *ḍh*) using the Bhāṣācitra system.

While Arsenault (2017) did use mapping, the feature-separating lines were calculated based on point coordinates for each language, not areal zones. Bhāṣācitra produces more accurate visualisations; it is immediately clear that the northwest Indo-Aryan and Nuristani, Dravidian, and Munda languages lack the phoneme, and this information can be used to inform locations for future fieldwork at the isogloss boundaries to refine our data.

5.2 Indo-Aryan sound changes

As another demonstration, we use an underdevelopment etymological database of Indo-Aryan languages (Arora and Farris, 2021) that builds on Turner (1962–1966) to map the outcomes of some key Indo-Aryan sound changes.[5]

The Indo-Aryan (IA) languages show complex overlapping phonological isoglosses as a symptom of intense cross-dialectal contact over a long period of time, whose complexity makes it difficult to make sense of the family's linguistic history. For example, the Sanskrit cluster /kṣ/ generally develops to /kʰ/ in the core region of modern Indo-Aryan and /t͡ʃʰ/ in the periphery, but some doublets are evidence of dialect contact, e.g. Sanskrit /kṣaːrə/ > Hindi /t͡ʃʰaːr/ 'ashes' as well as /kʰaːr/ 'alkali' (Masica, 1993). The variability of these sound changes has recently been used

[5]The compilation of the Jambu database is not in the scope of this work, but, briefly, it has been compiled by parsing data from the digitised version of Turner (1962–1966) and augmenting it with several more recent diachronic dictionaries for Indo-Aryan languages.

to statistically model dialect components in IA languages (Cathcart, 2019a,b, 2020; Cathcart and Rama, 2020).

Thus, a visualisation of the probability of certain IA sound changes based on a lexical database would be useful for finding isoglosses and the geographical extent of historical dialect contact. We aligned the cognate forms given in Arora and Farris (2021) using the LingPy library's multiple alignment function (List et al., 2019). Based on the alignments, the likelihood of /kṣ/ > /k(ː)^{h}/ is mapped in figure 4b. A rough core–periphery distinction indeed emerges, with languages in the northwest, south, and east having fewer outcomes of /k(ː)^{h}/. It is also apparent that the language coverage in Turner (1962–1966) is limited, with a great deal of core IA languages lacking data.

6 Future work

We intend to maximise coverage of South Asian languages in Bhāṣācitra. In the interest of achieving this goal we welcome contributions to our open-source database on GitHub: `https://github.com/aryamanarora/bhasacitra`. Ultimately, this sort of database would be useful for all languages of the world, but we lack the domain knowledge for non-South Asian languages, so we welcome any collaborators who feel this system would be beneficial.

As for directions for technical work, Bhāṣācitra would benefit from a SQL database for faster querying and precomputation of some data (e.g. language circle sizes and coordinates) to improve performance in the browser. In the interface, we will explore continuous alternatives to discretised Voronoi polygons, which force rigid transitions between lects[6] and do not show where location coverage is sparse. This will also help us with the issue of large polygons at the edges of our research area. Also, a basemap with administrative boundaries and other contextual geographical information would be useful. All of these will require substantial changes to the code beyond the capabilities of visualisation with pure D3.js.

Bhāṣācitra is one step of our larger goal of improving the study of South Asian languages with computational methods. Our future work on historical/comparative linguistics (Arora and Farris, 2021) and corpus linguistics for under-studied languages of the region will benefit from Bhāṣācitra's visualisation capabilities.

[6] We thank both reviewers for pointing out this limitation.

7 Conclusion

We developed and presented Bhāṣācitra, a database of linguistic resources for South Asia and a language visualisation system based on location data from those resources. We analysed the coverage of our database and used the areal mapping system to visualise phoneme inventories and Indo-Aryan sound change outcomes. We hope that researchers find the tool useful especially as we move forward with studying the typology of South Asian languages.

Acknowledgments

We thank Kaushalya Perera for providing her personal linguistic bibliography for Sinhala, Erik Anonby for showing us the *Atlas of the Languages of Iran* (ALI) project, and Henrik Liljegren for pointing us to his work on Hindu Kush typology.

We also thank Nathan Schneider for his helpful comments on the paper and Ananya Chakravarti for the useful discussion when devising this project.

References

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Erik Anonby, Mortaza Taheri-Ardali, and Amos Hayes. 2019. The Atlas of the Languages of Iran (ALI): A research overview. *Iranian studies*, 52(1–2):199–230.

Erik Anonby, Mortaza Taheri-Ardali, Fraser Taylor, and Amos Hayes. 2018. Atlas of the Languages of Iran.

Aryaman Arora and Adam Farris. 2021. *Jambu*. Georgetown University, Washington.

Paul Arsenault. 2017. Retroflexion in South Asia: Typological, genetic, and areal patterns. *Journal of South Asian Languages and Linguistics*, 4(1):1–53.

Ronald E Asher. 2008. Language in historical context. *Language in South Asia*, pages 31–48.

Joan L. G. Baart and Esther L. Baart-Bremer. 2001. *Bibliography of languages of northern Pakistan*. NIPS–SIL Working Paper Series. National Institute of Pakistan Studies, Quaid-i-Azam University and Summer Institute of Linguistics.

Elena Bashir. 2016. Contact and convergence. In Hans Henrich Hock and Elena Bashir, editors, *The Languages and Linguistics of South Asia: A Comprehensive Guide*. De Gruyter Mouton.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

Gerd Carling, Filip Larsson, Chundra A. Cathcart, Niklas Johansson, Arthur Holmer, Erich Round, and Rob Verhoeven. 2018. Diachronic Atlas of Comparative Linguistics (DiACL)—a database for ancient language typology. *PLOS ONE*, 13(10):1–20.

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.

Ashwini Deo. 2018. Dialects in the Indo-Aryan landscape. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pages 535–546. Wiley Online Library.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2021. *Ethnologue: Languages of the World*, 24th edition. SIL International.

John J. Gumperz. 1958. Dialect differences and social stratification in a North Indian village. *American Anthropologist*, 60(4):668–682.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. *Glottolog 4.3*. Max Planck Institute for the Science of Human History.

Taylor Jones. 2015. Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, 90(4):403–440.

Aikaterini-Lida Kalouli, Rebecca Kehlbeck, Rita Sevastjanova, Katharina Kaiser, Georg A. Kaiser, and Miriam Butt. 2019. ParHistVis: Visualization of parallel multilingual historical data. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 109–114, Florence, Italy. Association for Computational Linguistics.

Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer.

William A. Kretzschmar. 2001. Linguistic databases of the American Linguistic Atlas Project (ALAP).

Yasuo Kumagai. 2016. Developing the Linguistic Atlas of Japan Database and advancing analysis of geographical distributions of dialects. In Marie-Hélène Côté, Remco Knooihuizen, and John Nerbonne, editors, *The future of dialects: Selected papers from Methods in Dialectology XV*. Language Science Press.

Henrk Liljegren, Robert Forkel, Nina Knobloch, and Noa Lange. 2021. Hindu Kush areal typology (version v1.0).

Johann-Mattis List, Simon J Greenhill, Tiago Tresoldi, and Robert Forkel. 2019. LingPy. A Python library for quantitative tasks in historical linguistics.

Alan M MacEachren. 2017. Leveraging big (geo) data with (geo) visual analytics: Place as the next frontier. In *Spatial data handling in big data era*, pages 139–155. Springer.

Colin P. Masica. 1993. *The Indo-Aryan languages*. Cambridge Lamguage Surveys. Cambridge University Press, Cambridge.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Yugo Murawaki. 2020. Latent geographical factors for analyzing the evolution of dialects in contact. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 959–976, Online. Association for Computational Linguistics.

Harold Orton, Stewart Sanderson, and John Widdowson, editors. 1998. *The linguistic atlas of England*. Psychology Press.

Kaushalya Perera. 2021. Personal communication.

John Peterson. 2018. Bibliography for seldom studied and endangered South Asian languages.

A. K. Ramanujan and Colin Masica. 2016. *Toward a Phonological Typology of the Indian Linguistic Area*, pages 543–577. De Gruyter Mouton.

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

Christopher Shackle. 1980. Hindko in Kohat and Peshawar. *Bulletin of the School of Oriental and African Studies, University of London*, 43(3):482–510.

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

A Source format

Below is reference metadata for Shackle (1980) in JSON format; note the location annotations and the topic data.

```
{
    "type": "article",
    "title": "Hindko in Kohat and Peshawar",
    "author": ["Christopher Shackle"],
    "journal": "Bulletin of the School...",
    "year": 1980,
    "volume": 43,
    "number": 3,
    "pages": "482--510",
    "url": "https://www.jstor.org/stable/615737",
    "languages": {
        "Hindko": ["Kohat", "Peshawar"]
    },
    "topics": ["overview"]
}
```