Computational historical linguistics and language diversity in South Asia

Anonymous ACL submission

Abstract

South Asia is home to a plethora of languages, most of which are severely lacking access to language technologies that have been developed with the maturity of NLP/CL. This linguistic diversity, however, also results in a research environment conducive to the study of comparative, contact, and historical linguistics—fields which necessitate the gathering of extensive data from many languages. We claim that data scatteredness (rather than scarcity) is the primary obstacle in the development of South Asian language technology, and suggest that the study of language history is uniquely aligned with surmounting this obstacle. We review recent developments in, and the intersection of, South Asian NLP and historical–comparative linguistics, explaining our current efforts in this area while also offering new paths towards breaking the data barrier.

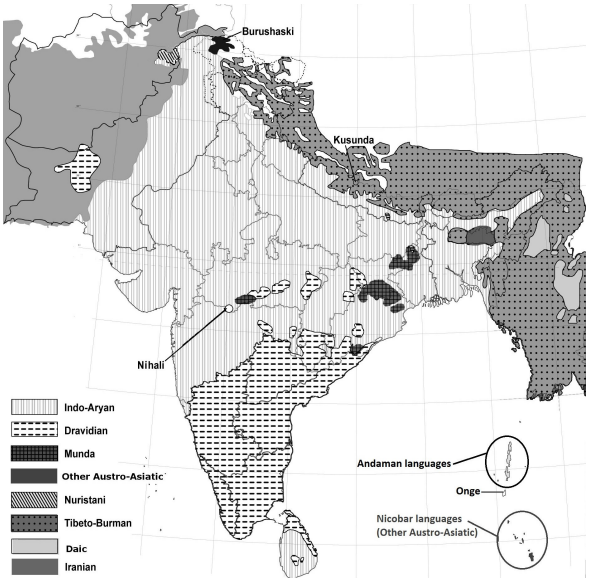

Figure 1: The major language families of South Asia (Kolichala, 2016).

1 Introduction

South Asia[1] is home to one-quarter of the world's population and boasts immense linguistic diversity (Saxena and Borin, 2008; Bashir, 2016). With members of at least five major linguistic families[2] and several putative linguistic isolates, this region is a fascinating arena for linguistic research. The languages of South Asia, moreover, have a long recorded history, and have undergone complex change through genetic descent, sociolinguistic interactions, and contact influence.

Nevertheless, South Asian languages for the most part remain severely underdocumented (van Driem, 2008), and several languages with even official administrative status (e.g. Sindhi) are low-resourced for the purposes of all natural language processing tasks (Joshi et al., 2020). This data scarcity persists despite long native traditions of linguistic description, continued language vitality with active use on the internet, and vast numbers of speakers (Rahman, 2008; Groff, 2017).

We argue that the most basic problem in NLP/CL work on South Asian languages is not data scarcity, but **data scatteredness**. There is much data to be extracted for even the most endangered languages (e.g. Burushaski, a language isolate of the northwest), from annotated corpora and grammatical descriptions compiled by linguists, *if only* one is willing to wrangle idiosyncratic data formats and digitise existing texts. Thus far, commercial interests and scientific agencies have only intermittently supported the development of language technology for the region—taking a new approach, we propose a research programme from the perspective of computational historical linguistics, outlining current data gathering initiatives in this discipline and potential benefits to other work across NLP.

[1] Roughly the Indian Subcontinent, or the geographic and cultural region enclosed by the Himalayas, the Indian Ocean, and the Hindu Kush.

[2] Indo-European (Indo-Aryan, Iranian, Nuristani), Dravidian, Austroasiatic (Munda, Khasian, Nicobarese), Tibeto-Burman

Level	Languages
4: Underdogs	Hindi
3: Rising Stars	Urdu, Bengali, Tamil
2: Hopefuls	Konkani, Marathi, Sanskrit, Punjabi
1: Scraping-Bys	Malayalam, Bhojpuri, Nepali, Doteli, Gujarati, Newar, Dzongkha, Maithili, Tulu, Kannada, Odia, Kashmiri, Romani, Pashto, Bishnupriya Manipuri, Divehi, Sindhi, Tibetan, Pali, Sinhala, Santali, Assamese, Telugu
0: Left-Behinds	(several hundred languages)

Table 1: A brief overview of NLP/CL research progress on South Asian languages grouped by Joshi et al. (2020)'s categories.

2 Related work

The state of NLP in South Asia. So far, initiatives for improving language technology in South Asia have largely focused on languages with official status and some degree of standardisation. These include cross-lingual projects such as IndicNLPSuite (Kakwani et al., 2020), the EMILLE corpus (McEnery et al., 2000), and iNLTK (Arora, 2020), and workshops like DravidianLangTech (Chakravarthi et al., 2021) and WILDRE (Jha et al., 2020). As table 1 shows, only a select few languages benefit from NLP research—even fewer benefit from (commericialised) products like Google Translate or OCR tools.

NLP/CL has proven to be an expansive field as of late. Computational historical linguistics is inextricably linked with computational approaches to fundamental linguistic tasks: corpus building, POS tagging and dependency parsing, morphological analysers, and lexical databases. Work on these has progressed fast for the big languages. For example, Hindi, the highest-resourced South Asian language, has massive hand-annotated dependency treebanks (Bhatt et al., 2009), state-of-the-art neural distributional semantic transformer models (Jain et al., 2020; Khanuja et al., 2021), and machine translation models to and from English (Saini and Sahula, 2018).

This is not to say that there are no resources at all for the languages (Joshi et al., 2020) terms "the Left-Behinds". Linguists, for example, have compiled rudimentary treebanks for many languages, simply waiting to be digitised and converted to a multilingual format like Universal Dependencies; these include Palula (Liljegren and Haider, 2015) and Toda (Emeneau, 1984), which are yet to be the subject of any NLP research work. There are also new treebanks in Universal Dependencies for Kangri, Mandeali, Bhojpuri (Ojha and Zeman, 2020), and Magahi.

Comparative linguistics in South Asia. The study of historical and comparative linguistics has a long history in South Asia, beginning well before similar threads of inquiry in the Western linguistic tradition, with grammarians like Pāṇini (c. 5th century BCE) and Hēmacandra (1088–1173) analysing historical and dialectal language from a comparative perspective.

Following the recognition by Western philologists of an Indo-European language family that includes Sanskrit, comparative study of the languages of South Asia began in earnest. As a result, several comprehensive comparative grammars featuring the Dravidian (Caldwell, 1856; Andronov, 2003; Krishnamurti, 2003) and Indo-Aryan families (Beames, 1872; Hoernlé, 1880; Bloch, 1934; Masica, 1993) have appeared in the years since. Emeneau (1956) was the first to posit a South Asian zone of language contact and convergence spanning multiple families. Subsequent work on micro-areal zones has yielded many insights into the nature of linguistic interactions in the region (Peterson, 2017; Liljegren et al., 2021; Toulmin, 2006).

The sole South Asia-wide linguistic data collection effort to ever be undertaken was the **Linguistic Survey of India**, completed about a century ago (Grierson, 1903–1928). To date, there has been no comparable centralised data resource on South Asian languages of its magnitude–covering typological features, the lexicon, and sociolinguistic phenomena.

Data in the earliest comparative works was frequently sourced from high-prestige standard varieties like Delhi Hindi, with progress on studying and collecting data from more localised lects largely proceeding in isolation. Compilation of comparative data continued sporadically throughout the 20th century, resulting in works such as the *Comparative Dictionary of the Indo-Aryan Languages* (Turner, 1962–1966) and the *Dravidian Etymology Dictionary* (Burrow and Emeneau, 1984) which attempt at a more diverse spectrum of language data. Meanwhile, progress on documentation and comparative analysis of the Austroasiatic (Anderson, 2008), Sino-Tibetan, and isolate languages (e.g. Burushaski, Nihali, Kusunda) of South Asia is still in its infancy. As a consequence, stud-

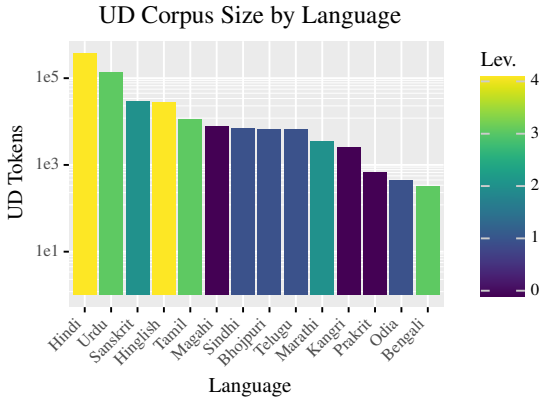

Figure 2: Universal Dependencies corpus sizes, in tokens, for all South Asian languages available thus far. Colors correspond to Joshi et al. (2020)'s level categorization.

ies drawing upon their data for purposes such as substrate analysis often lack nuance and family-internal consistency.

3 Ongoing work

Having established the issue of data scarcity, the mutual benefit inherent to data collection (for historical/comparative linguistic work and other NLP tasks), as well as possible interesting avenues for future research, we present a compilation of our ongoing projects in this direction, most involving languages that have not been studied in NLP before.

3.1 Dependency treebanks

Structured, syntactically-parsed corpora are not only essential for (1) **downstream NLP tasks** such as information extraction (Gamallo et al., 2012) and semantic role labelling (Li et al., 2019), but also have the potential to (2) aid quantitative **comparative and historical linguistic study**. Parsing according to several formalisms is possible, though dependency formalisms in particular are better equipped to handle the flexible word-order characteristic of many South Asian languages (assuming the parsing algorithm used adequately handles non-projective dependency trees[3]) (Palmer et al., 2009).

Multilingual dependency formalisms such as Universal Dependencies (UD) (Nivre et al., 2016) have established consistent guidelines for the annotation of binary dependency relations, morphology, and other linguistic features, resulting in the recent appearance of treebanks for several low-resourced languages of the region (Bhojpuri, Kangri, etc.) as well as their older diachronic stages (Vedic and Classical Sanskrit).

Towards the second goal listed above, Farris and Arora (2022) compiled a UD treebank for the Ashokan Prakrit dialect continuum–a parallel corpus of 14 pillar/rock inscriptions in six Middle Indo-Aryan (MIA) dialects dating back to the 3rd c. BCE. As the first study of MIA from a computational perspective, their work calls for an analysis of Indo-Aryan regional fragmentation through dialectometry, approaching contentious linguistic issues with statistical arguments curated using treebank data.

In a similar vein, we are currently working towards filling other chronological gaps in corpora (e.g. the Old Sinhala Sīgiri Graffiti of the Early New Indo-Aryan stage) through treebanking *in parallel* with their low-resourced, modern stages (e.g. Sinhala). To the best of our knowledge, we are unaware of any studies involving such **diachronic transfer** frameworks, where knowledge transfer between two historically-separated stages of the same language can be used to dependency-parse a given stage using resources from the other. Other historically-attested langauges we plan to include in this pipeline include Old Kashmiri, Old Maldivian, and Old Tamil.

Multilingual dependency parsing. More broadly, we are interested in cross-lingual transfer models (Duong et al., 2015; Guo et al., 2015; Schuster et al., 2019) as a means of expediting dependency parsing for low-resourced, South-Asian languages. A similar approach for Uralic languages is (Lim et al., 2018). They propose a dependency-parsing model for North Saami and Komi using annotated corpora and bilingual word-embeddings from high-resourced genetically related (Finnish) and typologically similar

[3]For vertex set V, weighted edge set $E \subseteq \{i \xrightarrow{w} j \mid i,j \in \mathbb{R}, w \in \mathbb{R}\}$, and root $\rho \in V$, let $G = (\rho, V, E)$ be a rooted weighted directed graph. A **dependency tree** is a spanning subgraph $D = (\rho, V, E')$, $E' \subseteq E$ subject to the following well-formedness constraints (Zmigrod et al., 2020):
(C1) Each non-root vertex of D has one incoming edge
(C2) D is acyclic
(C3) Root ρ of D has exactly one outgoing edge
In other words, dependency trees are *arborescences* (directed, rooted trees) equipped with the root constraint (C3). Graph-based parsing algorithms find the optimal dependency tree D^*, that is, the dependency tree D with maximum total edge weight in the set of all possible dependency trees $D(G)$, for a given sentence (maximum weight spanning arborescence). A **treebank** is a corpus of such dependency trees.

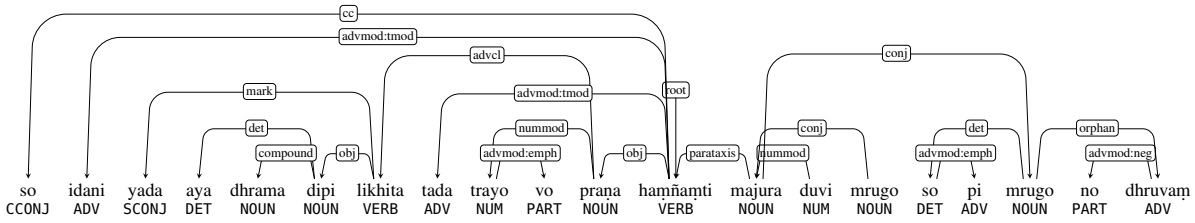

Figure 3: A sample dependency-parse from the Ashokan Prakrit UD treebank (Shahbazgarhi dialect)

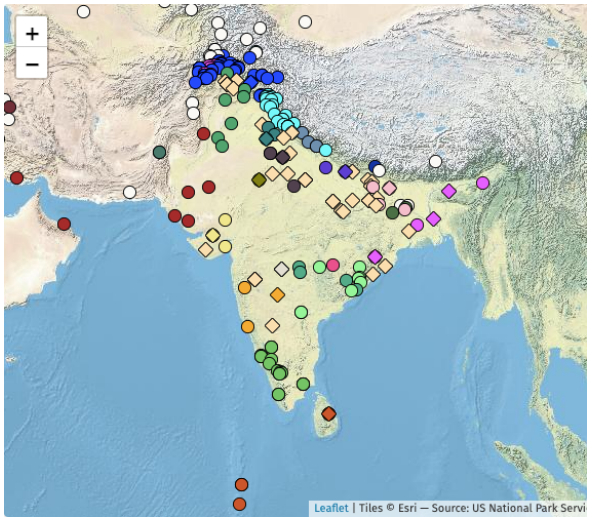

Figure 4: A map of languages included in Jambu, colour-coded by subfamily designation with point-geometry variation by diachronic stage.

(Russian) languages, without the requirement of extensive parallel texts for training. They conclude that while genetically related pairs (Komi–Finnish, North Saami–Finnish) allow for highly efficient parsing, pairs of unrelated languages in contact (Komi–Russian) also provide valuable input for further correction. Given the languages of South Asia exhibit common typological features by virtue of sharing a linguistic area, treebanking efforts will undoubtedly benefit from a multilingual dependency parsing approach. Languages like Sindhi, Punjabi, and Sinhala, which have genetic relatives and contact languages that are comparatively more resourced, are our immediate targets for such efforts.

3.2 *Jambu* etymological database

One of our major efforts in data-collection for the region has been the **Jambu** project. Jambu is a compiled cognate lexicon of all South Asian languages, cutting across phylogenetic groupings and historical language stages. It has a web interface online at `https://neojambu.glitch.me/`. It includes data parsed and compiled from the University of Chicago's Digital Dictionaries of South Asia project (Turner, 1962–1966; Burrow and Emeneau, 1984), existing web databases (Liljegren et al., 2021; Strand, 1997–2021), and individual articles and theses (Toulmin, 2006; Jouanne, 2014), totalling 294 lects and 202,653 lemmas. Some of these sources have been used in previous work on South Asian historical linguistics, e.g. Cathcart and Rama (2020); Cathcart (2019b,a, 2020)—this is the first attempt to consolidate them. Note some previous work in this direction: while the SARVA project (Southworth, 2005) did not reach fruition, a searchable database of the DEDR was developed by Suresh Kolichala under its auspices.[4]

Past etymological research in South Asian languages was primarily focused on internal comparisons within linguistic families. Unknown etyma was often blindly attributed to Dravidian or Munda without comprehensive cross-linguistic analyses.[5] In fact, we find a large number of common words in languages of several families with uncertain origin, possibly substrate loans from undocumented languages.[6] In order to provide reliable data for the robust reconstruction of the history of the ancient linguistic contact, a comprehensive South Asia-wide linguistic data is desideratum.

Consolidating Indo-Aryan data. While Turner (1962–1966) and its supplements remain the undisputed gold standard for Indo-Aryan comparative etymologies, many later works on individual languages have considerably expanded our knowl-

[4] `http://kolichala.com/DEDR/`

[5] Recent comparative work on Munda and Indo-Aryan contact such as Ivani et al. (2020) in general find very limited influence of Munda, restricted primarily to the (eastern) Indo-Aryan languages in close proximity with them. Prior work had a tendency to exaggerate the impact of Munda to explain unusual features of other Indo-Aryan languages; notably, Witzel (1999), who advocated for a historical 'Para-Munda' family that influenced Indo-Aryan as far as in the northwest, the historical location of Rigvedic Sanskrit.

[6] Dr. Felix Rau (p.c.) terms these unattested substrate(s) 'the **big X** of South Asian linguistic history', and other (possible same) substrate(s) responsible for words reconstructable to Proto-Munda without secure cognates in other Austroasiatic branches 'the **big Y**'.

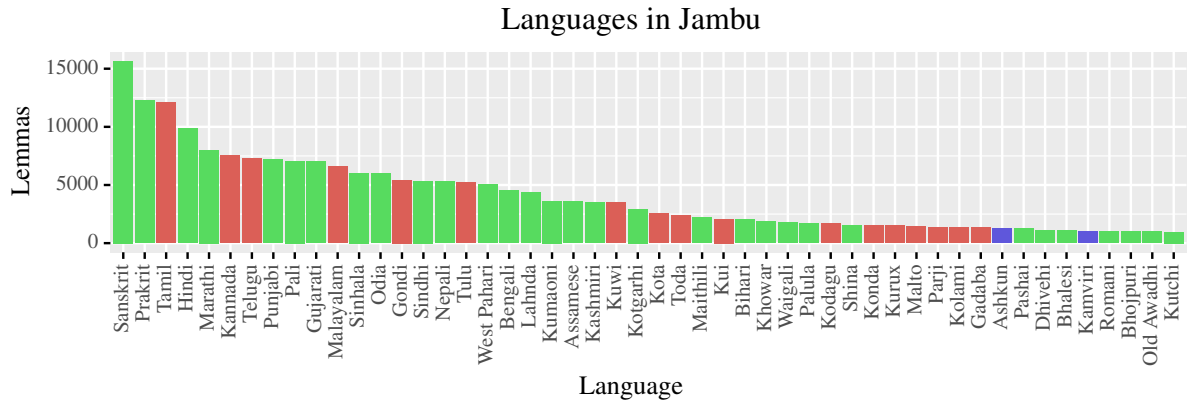

Figure 5: Top 50 languages by number of lemmas included in the Jambu database, colour-coded by language family.

edge of cognate relations in underdocumented languages; e.g. Liljegren et al. (2021); Toulmin (2006); Zoller (2005). Inclusion of data from these newer works is ongoing. We also expanded coverage of the isolated and linguistically archaic Nuristani lects (Strand, 1997–2021), which are contended not to be Indo-Aryan—comparative lexical data will help cement their exact phylogenetic status.

Updates to Dravidian data. A *Dravidian Etymological Dictionary* published by Burrow and Emeneau (1984) (2nd edition; abbreviated DEDR) remains the latest effort to gather etymological data on Dravidian. Although Krishnamurti (2003) provides reconstructions for about 500 entries, systematic historical reconstruction for all known cognates of Dravidian is still pending. Subrahmanyam (2013) published an update to the DEDR utilises new data on serveral non-literary languages that became available after 1984.

Recent fieldwork on several non-literary languages have produced grammars with new vocabulary lists, providing rich data to be updated in DEDR. In addition, several dictionaries with attempted etymologies for many literary languages have appeared since 1984, and can become a source for the realignment of cognates as well as new additions.

Cognate databases in NLP. The obvious benefit of cognate databases for upstream NLP tasks is for low-resourced languages that lack adequate corpora on the web. Similar work in this area is the pan-lingual CogNet (Batsuren et al., 2019), and also earlier WordNets (Miller, 1995). Cognate data can be used for transfer learning, where a low-resourced language can map onto existing models for high-resource languages, such as a distributional semantic model which generally requires massive corpora to train (Sharoff, 2017). Typological data in general offers modest improvements in performance on a variety of NLP tasks (Ponti et al., 2019).

Unified transcription. Since many languages of South Asia are unwritten or are lacking standardised orthographies (even in their respective linguistics works), we developed a preliminary system for phonemic transcription of all South Asianlanguages, which all our cognate data will be converted to. For cognate identification and reconstruction work (both by humans and using NLP tools), a unified phonemic representation is important. This system combines features of the International Alphabet of Sanskrit Transliteration (IAST)[7] with IPA and Americanist phonetic transcription systems. Future work will outline it in depth, along with examples of its focus on cross-family diacritical consistency.

3.3 Historical linguistic analyses

One of our main objectives for building extensive comparative lexical and grammatical databases is to ensure credible data from up-to-date, modern sources are available to researchers working on comparative and diachronic linguistics in the South Asian linguistic area. Below, we highlight two such projects we are currently engaged in involving three low-resourced languages of northern Pakistan: Burushaski, Gawri, and Torwali (Torwali, 2018).

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

3.3.1 Gawri tonogenesis and UniMorph

The languages of northern Pakistan have synchronically been analyzed to have phonemic tonal contrasts. Baart (2003) has classified such tonal languages into three broad groups based on the type of tonal contrast displayed:

- *Shina-type*: Shina varieties, Palula, Indus Kohistani (all Indo-Aryan), Burushaski (isolate) etc.
- *Punjabi-type*: Punjabi, Hindko, some Gujari varieties, extending into the Himachali languages of northern India, as well as Kishtwari,[8] which is usually classified as a divergent dialect of Kashmiri (all Indo-Aryan).
- *Kalami-type*: Gawri (Kalami), Kalkoti and Torwali (all Indo-Aryan) and possibly other undiscovered varieties of the area.

To these, one may also add the simpler accentual systems of Kalasha-mon (Heegård-Petersen, 2015) and Khowar (Liljegren and Khan, 2017), which we term *Chitrali-type*.

The tonal system and the historical mechanism of tonogenesis is broadly understood for Punjabi proper and some Hindko varieties (Shackle, 1980; Bashir and Conners, 2019; Bhatia, 2013), but specifics for individual varieties further east (Kishtwari and Himachali) remain underdescribed (Hendriksen, 1986; Jouanne, 2014). This system arises primarily from the disappearance of phonemic breathy voice, but the phonetic specifics differ from language to language. The Shina-type tonal system is both the best described and the best understood diachronically. It continues the Vedic (hence Indo-European) pitch-accent system subject to later changes necessitated by regular apocope (Liljegren, 2008, 2016; Kümmel, 2015). Vedic pitch-accent is also partly continued by the Chitrali-type accentual system (Heegård-Petersen, 2012), though less conservatively.

The tonal diachrony of the Kalami-type system, on the other hand, has not yet been fully understood. Part of the reason is that this system is considerably more complex than the other three accentual systems, contrasting as many as five distinct tonemes (Baart, 1997; Lunsford, 2001; Liljegren, 2013). In ongoing work, based on the Gawri data compiled from Baart (1997, 1999); Baart and Sagar (2004); Baart et al. (2004), we are investigating the origin of the system, and will be appended in the future by Torwali data we are now collecting.

[8] Not mentioned in (Baart, 2003), but independently identified by one of the present authors.

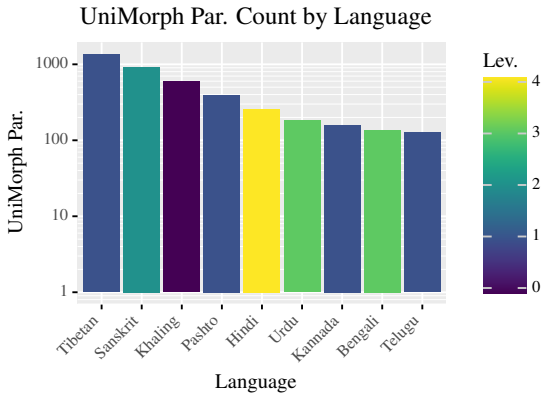

Figure 6: UniMorph paradigm counts for all South Asian languages available thus far. Colors correspond to Joshi et al. (2020)'s level categorization.

Morphology in NLP. In addition to working out the history of the Kalami-type tonal system, we intend to incoporate our annotated lexical dataset into the **UniMorph** database (Kirov et al., 2018). The morphology of Gawri and Torwali marks gender, number and case for nouns and adjectives primarily by tonal changes and vowel alterations (historical umlaut) unlike other Indo-Aryan languages which use suffixation, though they still encode much the same categories and do not behave any different syntactically either. This makes them prime targets for testing out computational methods for morphological analysis, especially to compare performance vis-à-vis a related language like Hindi that has a similar grammar but different morphological profile.

UniMorph has only a few South Asian languages thus far, as shown in figure 6—this is part of a broader project to expand coverage in the region, using existing morphological data stored in analysers (e.g. for Sindhi, Motlani et al., 2016) and grammars (e.g. for Palula, Liljegren et al., 2021). In this vein, we also mention that UniMorph has only a handful of languages that signal morphological alterations tonally. So, our contribution will also improve typological diversity in the database to a considerable extent.

3.3.2 Proto-Burushaski reconstruction

Our understanding of the linguistic pre-history of South Asia is heavily reliant on disciplined studies of the histories of the non-Indo-European languages of the subcontinent. This is primarily because while we do have reliable estimates on the time-frame of Indo-European migration into the

subcontinent, for the families endemic to the region (including isolates) analogous dating is not possible.

Burushaski, spoken in a few mountain valleys of the Karakoram, is among these endemic languages of South Asia. It has attracted quite a bit of scholarly attention since its academic discovery as it stands out both typologically and genealogically in its current neighborhood (cf. the latest descriptive grammars Berger (1974, 1998); Munshi (2018); Yoshioka (2012)). The history of the language and its speakers is virtually unknown until the first linguistic documentation in the mid-nineteenth century. The first secure pre-modern attestation of Burushaski speakers is in Tibetan chronicles dating from the ninth century where a people *bru-ża* or *bru-ša* to the west of Tibet find mention (Jäschke, 1881).[9,10]

As of now, both major varieties of Burushaski are well-documented, but there has been precious little comparative work done. The dictionaries in Berger (1974, 1998) lay the foundation of comparative studies by identifying several layers of potential loans in the language, cf. also Rybatzki (2010). Conversely, potential Burushaski interaction with and influence on the older stages of Indo-Iranian have been explored in Tikkanen (1988); Kümmel (2018), the former mainly dealing with how Burushaski broadly fits into the South Asian linguistic zone. A handful of Burushaski loans in Purik Tibetan are identified in Zemp (2018), not all of them convincing, and Steblin-Kamenskij (1999) contains shared lexemes with Wakhi. More speculative are the claimed Burushaski loans in (Proto-)Romani collected in Berger (1959), believed to be borrowed before the Roma migrated westward toward Europe (presumably) through Burushaski territory.

However, all these studies share a common drawback in that we do not yet have a principled way of identifying Burushaski lexemes or grammatical features. A first step toward this goal is Holst (2014), where the author attempts an internal reconstruction of Burushaski through a comparative lexical and morphological study of the two main dialect groups of Yasin and Hunza–Nager. Holst's work, though, is still just a preliminary investigation and there is much to be added and improved on. In particular, the book does not undertake a systematic study of loanwords to and from neighboring languages as previous areal studies involving Burushaski have, nor does it exhaustively utilize the descriptive literature available resulting in a few avoidable but significant errors of interpretation (Munshi, 2015). This is a major shortcoming because external comparisons are a vital component to reconstructing the histories of language isolates and smaller families, cf. Trask (2013) for Basque and Nikolaeva (2011) for Yukaghir, among others.

Computational reconstruction. We have already started a principled reconstruction of Proto-Burushaski building on Holst's work, but utilizing more sources and laying a greater emphasis on loanword etymologizing and chronologizing. Our databases, compiled from available lexical and descriptive sources, are intended to aid this goal of comparative analysis, as well as to make data from Burushaski and neighboring languages available to other researchers.

Proto-language reconstruction is an interesting task in computational historical linguistics, and so far work has been under way in a supervised setting on known, high-quality cognate data across related languages, e.g. on Romance languages (Ciobanu and Dinu, 2018; Meloni et al., 2021).

Low-resource dependency corpora. In addition, starting with annotated texts from descriptive grammars, we plan to build a dependency treebank for Burushaski as described in §3.1. Burushaski is a low-resourced language in the sense that its domain of use is very restricted and there is no readily available internet corpus one can subject to sophisticated (computational) linguistic analyses automatically.

However, as mentioned before, there has been a steady stream of quality descriptive work on it and all published grammars come with a wealth of oral texts one can build a functional corpus with.

4 Future Work

The data resources we are in the process of compiling for South Asian languages will enable a variety of research to be conducted into language history. We lay out some of the immediate potential pathways for this further research in hopes of stimulating work in this area.

[9] We are grateful to Dr. Diego Loukota (p.c.) for informing us that a short text in the *bru-ša* language is also attested in Tibetan records with translation in Sanskrit. We are, however, not aware of any scholarly attempt to interpret said text through modern Burushaski.

[10] It is also possible that an older ethnonym recorded as Sanskrit *mūja-*, *maujavata-* and Avestan *muža-* refer to the same people but that is harder to establish.

4.1 Substrate studies and language history

A perennial question in South Asian language history for at least a century has been the Indus Valley Civilisation inscriptionary corpus, and the problem of deciphering it (if it even encodes a language) and whether it belongs to a known language family of South Asia or something else entirely (Farmer et al., 2004; Fairservis, 1983). Notably, in the mid-20th century a team of Finnish and Soviet linguists and computer scientists claimed evidence that the Indus inscriptions represent a Dravidian language (Parpola, 1986).

Recent computational information-theoretic work also suggests language-like properties in the text, a subject of subsequent vociferous debate (Rao et al., 2009, 2010). A serious issue is that we do not have sufficiently diverse data from modern languages of the region against which to compare any purported decipherments of the Indus script (e.g. Proto-Dravidian reconstruction is as of now still in a preliminary stage), and thus even if the Indus language provided any substrate loans into modern families, we would be unable to comprehensively list out possible candidates. The *Jambu* database can help inform research on substrate contact in the languages of the region.

4.2 Text digitisation and OCR

One of the major bottlenecks in compiling existing linguistic data on South Asian languages is that it remains machine-unreadable. For example, many linguistics theses completed at Indian universities have recently been digitised and uploaded to Shodhganga,[11] but most are scanned images in PDF format. Optical character recognition (OCR) of such texts also requires difficult parsing of diacritics and low-resource scripts.

A recent initiative to digitise old linguistic data is the digitisation of the Linguistic Survey of India (Grierson, 1903–1928) under the project *South Asia as a linguistic area? Exploring big-data methods in areal and genetic linguistics* (Borin et al., 2020, 2018, 2014). Using OCR and subsequent information extraction from the text, Borin et al. have shown that "old" data still has much to tell for the computational study of typology and comparative linguistics.

Future work on extracting data from non-digitised South Asian language sources will have to use OCR, possibly a neural model finetuned for the purposed of our domain on a platform like Transkribus (Kahle et al., 2017).

4.3 Fieldwork initiatives

Hämäläinen (2021), calling for the NLP community to make a consistent distinction between "endangered" and "low-resourced" languages, implores researchers to 'stop complaining about how low-resourced [a language] is, [and] get up and gather the data.'

In response to this call, we announce several currently-underway (online) fieldwork/data elicitation efforts for Indo-Aryan languages that are both endangered and low-resourced (from the perspective of NLP). These include Kholosi, Poguli, Kishtwari, Bhaderwahi, Torwali, and certain divergent dialects of Maldivian (e.g. Huvadhoo). By virtue of their geographical spread (Northern India/-Pakistan, Iran, Maldives), linguistic data collected from these languages will further enable the consturction of typologically viable datasets for both NLP and computational historical linguistic tasks.

5 Conclusion

In this paper, we gave an overview of the state of NLP in South Asia with a special focus on historical–comparative linguistics, a research program of which we believe will help address the issue of data scatteredness. South Asian languages are not obliged to remain low-resource (in the NLP sense), and have plenty of speakers who would like access to and would benefit from language technologies, along with a multitude of raw linguistic resources that can be used to cultivate them. Incentives have not been in place to support those demands, however, so we suggest an alternative route founded in linguistic research to gather data.

Collective efforts have had great success recently in NLP—besides institutional efforts like the Stanford Center for Research on Foundation Models (Bommasani et al., 2021) and HuggingFace's BigScience Workshop,[12] there are grassroots organisations like MaskhaneNLP for African languages (Nekoto et al., 2020) and AI4Bharat (Kakwani et al., 2020) that are working towards improving resource availability. Our proposals in this paper are the first seeds of a programme similar in spirit, motivated by a dual interest in understanding South Asian language history and remedying inequalities in technological availability.

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

[12] https://bigscience.huggingface.co/

References

Gregory D.S. Anderson. 2008. *The Munda Languages*. Routledge, London.

Mikhail Sergeevich Andronov. 2003. *A Comparative Grammar of the Dravidian Languages*. Beitrage zur Kenntnis sudasiatischer Sprachen und Literaturen. Harrassowitz, Wiesbaden.

Gaurav Arora. 2020. iNLTK: Natural language toolkit for indic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 66–71, Online. Association for Computational Linguistics.

Joan LG Baart. 1997. The sounds and tones of Kalam Kohistani. *National institute of Pakistan studies Quaid-i-Azam University and Summer Institute of Linguistics. P*, pages 1–9.

Joan LG Baart. 1999. *A sketch of Kalam Kohistani grammar*, volume 5. National Institute of Pakistan Studies, Quaid-i-Azam University.

Joan LG Baart. 2003. Tonal features in languages of northern pakistan. *Pakistani languages and society: problems and prospects*, pages 132–144.

Joan LG Baart, Esther L Baart-Bremer, and Muhammad Zaman Sagar. 2004. Names of plants in Kalam Kohistani (Pakistan). *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session*, 48(1):1.

Joan LG Baart and Muhammad Zaman Sagar. 2004. *Kalam Kohistani texts*, volume 9. National Institute of Pakistan Studies.

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.

Elena Bashir and Thomas J Conners. 2019. *A descriptive grammar of Hindko, Panjabi, and Saraiki*. De Gruyter Mouton.

Khuyagbaatar Batsuren, Gabor 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.

John Beames. 1872. *A comparative grammar of the modern Aryan languages of India: to wit, Hindi, Panjabi, Sindhi, Gujarati, Marathi, Oriya, and Bangali*, volume I: Sounds. Trübner & Co., London.

Hermann Berger. 1959. Die Burušaski-Lehnwörter in der Zigeunersprache. *Indo-Iranian Journal*, 3(1):17–43.

Hermann Berger. 1974. *Das Yasin-Burushaski (Werchikwar): Grammatik, Texte, Wörterbuch*, volume 3. Otto Harrassowitz Verlag.

Hermann Berger. 1998. *Die Burushaski-Sprache von Hunza und Nager*, volume 13. Otto Harrassowitz Verlag.

Tej Bhatia. 2013. *Punjabi*. Routledge.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for Hindi/Urdu. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 186–189, Suntec, Singapore. Association for Computational Linguistics.

Jules Bloch. 1934. *L'Indo-Aryen du Véda aux temps modernes*. Adrien-Maisonneuve, Paris.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models.

Lars Borin, Anju Saxena, Bernard Comrie, and Shafqat Mumtaz Virk. 2020. A bird's-eye view on south asian languages through LSI: Areal or genetic relationships? *Journal of South Asian Languages and Linguistics*, 7(2):151–185.

Lars Borin, Anju Saxena, Taraka Rama, and Bernard Comrie. 2014. Linguistic landscaping of South Asia using digital language resources: Genetic vs. areal linguistics. In *Proceedings of the Ninth International*

Conference on Language Resources and Evaluation (LREC'14), pages 3137–3144, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lars Borin, Shafqat Mumtaz Virk, and Anju Saxena. 2018. Language technology for digital linguistics: Turning the Linguistic Survey of India into a rich source of linguistic information. In *Computational Linguistics and Intelligent Text Processing*, pages 550–563, Cham. Springer International Publishing.

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

Robert Caldwell. 1856. *A comparative grammar of the Dravidian or South-Indian family of languages*. Trübner & Co., London.

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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Parameswari Krishnamurthy, and Elizabeth Sherly, editors. 2021. *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Kyiv.

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.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 113–122, Beijing, China. Association for Computational Linguistics.

M. B. Emeneau. 1956. India as a linguistic area. *Language*, 31(1).

Murray Barnson Emeneau. 1984. *Toda grammar and texts*, volume 155. American Philosophical Society.

Walter A Fairservis. 1983. The script of the Indus Valley civilization. *Scientific American*, 248(3):58–67.

Steve Farmer, Richard Sproat, and Michael Witzel. 2004. The collapse of the Indus-script thesis: The myth of a literate Harappan civilization. *Electronic journal of Vedic studies*, 11(2):19–57.

Adam Farris and Aryaman Arora. 2022. For the Purpose of Curry: A UD Treebank for Ashokan Prakrit. In press.

Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 10–18, Avignon, France. Association for Computational Linguistics.

George A. Grierson. 1903–1928. *Linguistic Survey of India*. Office of the Superintendent of Government Printing, India.

Cynthia Groff. 2017. Language and language-in-education planning in multilingual india: A minoritized language perspective. *Language Policy*, 16(2):135–164.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, Beijing, China. Association for Computational Linguistics.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! *arXiv preprint arXiv:2103.09567*.

Jan Heegård-Petersen. 2012. Animacy, Vedic accent and Kalasha case allomorphy. *Muenchener Studien zur Sprachwissenschaft*, 66(1):55–64.

Jan Heegård-Petersen. 2015. Kalasha texts–with introductory grammar. *Acta Linguistica Hafniensia*, 47(sup1):1–275.

Hans Hendriksen. 1986. *Himachali Studies: Grammar*. Munksgaard.

Rudolf Hoernlé. 1880. *A Comparative Grammar of the Gaudian Languages*. Trübner Company.

Jan Henrik Holst. 2014. *Advances in Burushaski linguistics*, volume 547. BoD–Books on Demand.

Hēmacandra. 1088–1173. *Siddʰa-Hēma-Śabdanuśāśana*.

Jessica K. Ivani, Netra Paudyal, and John Peterson. 2020. Indo-Aryan — a house divided? Evidence for the east–west Indo-Aryan divide and its significance for the study of northern South Asia. *Journal of South Asian Languages and Linguistics*, 7(2):235–274.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-Transformers: An analysis of transformer language models for Indian languages.

Heinrich August Jäschke. 1881. *A Tibetan-English Dictionary: with special reference to the prevailing dialects, to which is added an English-Tibetan vocabulary*. Routledge.

Girish Nath Jha, Kalika Bali, Sobha L., S. S. Agrawal, and Atul Kr. Ojha, editors. 2020. *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*. European Language Resources Association (ELRA), Marseille, France.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Thomas Jouanne. 2014. *A preliminary analysis of the phonological system of the Western Pahārī language of Kvār*. Ph.D. thesis, University of Oslo.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus—a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24. IEEE.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual representations for Indian languages.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Suresh Kolichala. 2016. Dravidian languages. In Hans Henrich Hock and Elena Bashir, editors, *The Languages and Linguistics of South Asia: A Comprehensive Guide*, pages 73–107. De Gruyter Mouton.

Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge Language Surveys. Cambridge University Press, Cambridge.

Martin Kümmel. 2015. Developments in the dissolution of the Indo-Iranian accentual system. In *the Workshop on Diachronic Morphophonology: Lexical Accent Systems at the 22nd International Conference on Historical Linguistics*. Naples, July, pages 27–31.

Martin Kümmel. 2018. Substrata in Indo-Iranic and related questions. In a handout of a presentation held at the conference *Loanwords and Substrata in the Indo-European Languages*. Limoges, June.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dependency or span, end-to-end uniform semantic role labeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6730–6737.

Henrik Liljegren. 2008. *Towards a grammatical description of Palula: An Indo-Aryan language of the Hindu Kush*. Ph.D. thesis, Institutionen för lingvistik.

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

Henrik Liljegren. 2016. *A grammar of Palula*. Language Science Press.

Henrik Liljegren and Naseem Haider. 2015. Palula texts. Forum for Language Initiatives.

Henrik Liljegren and Afsar Ali Khan. 2017. Khowar. *Journal of the International Phonetic Association*, 47(2):219–229.

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

KyungTae Lim, Niko Partanen, and Thierry Poibeau. 2018. Multilingual dependency parsing for low-resource languages: Case studies on north saami and Komi-Zyrian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wayne A Lunsford. 2001. *An overview of linguistic structures in Torwali, a language of Northern Pakistan*. The University of Texas at Arlington.

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

Anthony McEnery, Paul Baker, Rob Gaizauskas, and Hamish Cunningham. 2000. EMILLE: building a corpus of South Asian languages. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the new Millennium: MT 2000*, University of Exeter, UK.

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.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Raveesh Motlani, Francis Tyers, and Dipti Sharma. 2016. A finite-state morphological analyser for Sindhi. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2572–2577, Portorož, Slovenia. European Language Resources Association (ELRA).

Sadaf Munshi. 2015. Review of Holst: *Advances in Burushaski linguistics*. *Journal of South Asian Languages and Linguistics*, 2(2):251–261.

Sadaf Munshi. 2018. *Srinagar Burushaski: A Descriptive and Comparative Account with Analyzed Texts*. Brill.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Irina Nikolaeva. 2011. *A historical dictionary of Yukaghir*. De Gruyter Mouton.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Atul Kr. Ojha and Daniel Zeman. 2020. Universal Dependency treebanks for low-resource Indian languages: The case of Bhojpuri. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 33–38, Marseille, France. European Language Resources Association (ELRA).

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Asko Parpola. 1986. The Indus script: A challenging puzzle. *World Archaeology*, 17(3):399–419.

John Peterson. 2017. Towards a linguistic prehistory of eastern-central South Asia (and beyond). *Journal of South Asian Languages and Linguistics*, 4(2):211–257.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Pāṇini. c. 5th century BCE. *Aṣṭād^{h}yāyī*.

Tariq Rahman. 2008. Language policy, multilingualism and language vitality in pakistan. In Anju Saxena and Lars Borin, editors, *Lesser-known languages of South Asia*, pages 73–106. De Gruyter Mouton.

Rajesh P. N. Rao, Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, Ronojoy Adhikari, and Iravatham Mahadevan. 2010. Commentary and discussion: Entropy, the indus script, and language: A reply to R. sproat. *Computational Linguistics*, 36(4):795–805.

Rajesh PN Rao, Nisha Yadav, Mayank N Vahia, Hrishikesh Joglekar, R Adhikari, and Iravatham Mahadevan. 2009. Entropic evidence for linguistic structure in the indus script. *Science*, 324(5931):1165–1165.

Volker Rybatzki. 2010. Türkische Lehnwörter im Burushaski. *Studia Orientalia Electronica*, 108:149–180.

Sandeep Saini and Vineet Sahula. 2018. Neural machine translation for English to Hindi. In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 1–6. IEEE.

Anju Saxena and Lars Borin. 2008. *Lesser-known Languages of South Asia: Status and Policies, Case Studies and Applications of Information Technology*, volume 175. Walter de Gruyter.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Serge Sharoff. 2017. Toward pan-Slavic NLP: Some experiments with language adaptation. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 1–2, Valencia, Spain. Association for Computational Linguistics.

F. C. Southworth. 2005. The sarva (south asia residual vocabulary assemblage) project.

I. M. Steblin-Kamenskij. 1999. *Ètimologičeskij slovarʲ vaxanskogo jazyka*. Peterburgskoe Vostokovedenie.

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

P S Subrahmanyam. 2013. *A Supplement to Dravidian Etymological Dictionary*.

Bertil Tikkanen. 1988. On Burushaski and other ancient substrata in northwestern South Asia. *Studia Orientalia Electronica*, 64:303–326.

Zubair Torwali. 2018. Language revitalization—a case study of Torwali. *Criterion Quarterly,(12)*, 4.

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.

Robert Lawrence Trask. 2013. *The history of Basque*. Routledge.

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

George van Driem. 2008. Endangered languages of south asia. In Matthias Brenzinger, editor, *Language Diversity Endangered*. De Gruyter Mouton.

Michael Witzel. 1999. Substrate languages in Old Indo-Aryan (Ṛgvedic, Middle and Late Vedic). *Electronic Journal of Vedic Studies*, 5(1):1–67.

Noboru Yoshioka. 2012. A reference grammar of Eastern Burushaski. *Tokyo: Tokyo University of Foreign Studies PhD Dissertation*.

Marius Zemp. 2018. *A grammar of Purik Tibetan*. Brill.

Ran Zmigrod, Tim Vieira, and Ryan Cotterell. 2020. Please mind the root: Decoding arborescences for dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4809–4819, Online. Association for Computational Linguistics.

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