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.

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

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

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

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

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; Toullmin, 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-
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 his-
torical/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 be-
fore.

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 com-
parative and historical linguistic study. Pars-
ing 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 (as-
suming the parsing algorithm used adequately han-
dles non-projective dependency trees)

Multilingual dependency formalisms such as
Universal Dependencies (UD) (Nivre et al., 2016)
have established consistent guidelines for the anno-
tation 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 cor-
pus 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 computa-
tional perspective, their work calls for an analy-
sis of Indo-Aryan regional fragmentation through
dialectometry, approaching contentious linguistic
issues with statistical arguments curated using tree-
bank data.

In a similar vein, we are currently working to-
wards filling other chronological gaps in corpora
e.g. the Old Sinhala Sigiri Graffiti of the Early
New Indo-Aryan stage through treebanking in par-
allel 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 languages we plan to include
in this pipeline include Old Kashmiri, Old Maldi-
vian, 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 geneti-
cally related (Finnish) and typologically similar
rected, 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 to-
tal edge weight in the set of all possible dependency trees
$D(G)$, for a given sentence (maximum weight spanning ar-
borescence). A treebank is a corpus of such dependency
trees.
Asia exhibit common typological features by (Komi–Russian) also provide valuable input for (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/](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 Emenau, 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.  

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. In fact, we find a large number of common words in languages of several families with uncertain origin, possibly substrate loans from undocumented languages. 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-
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.


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 linguistic works), we developed a preliminary system for phonemic transcription of all South Asian languages, 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 Kishthwari,\(^8\) 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 underscribed (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.

**Morphology in NLP.** In addition to working out the history of the Kalami-type tonal system, we intend to incorporate 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

---

\(^8\)Not mentioned in (Baart, 2003), but independently identified by one of the present authors.
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-za or bru-sa 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 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 Stehlin-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 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.

---

9We are grateful to Dr. Diego Loukota (p.c.) for informing us that a short text in the bru-sa 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.

10It is also possible that an older ethnonym recorded as Sanskrit mājā, mājavatta- and Avestan mača- 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, 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 purpose 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 construction 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, 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.

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

12https://bigscience.huggingface.co/
References


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


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


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


