What NLP Gets Wrong About Contact: Implications for Field Linguistic Evidence

Manodnya K H
CLASIC,
University of Colorado Boulder
Boulder, CO, USA
manodynak@gmail.com

Abstract

Field linguistics increasingly relies on computational tools to organize, analyze, and preserve linguistic data, yet the classificatory assumptions embedded in these tools are rarely examined. A pervasive assumption is that languages can be treated as discrete, genealogically defined units, with relatedness modeled as tree-structured descent. We argue that this assumption misrepresents linguistic evidence in contact-heavy regions and risks distorting the computational mediation of field linguistic data. Focusing on South Asia, we show that widely assumed boundaries—such as the Indo-Aryan–Dravidian divide—collapse in long-standing contact zones characterized by convergence, dialect continua, and institutional multilingualism. Through historically grounded case studies including Kannada–Telugu and Tamil–Malayalam, we demonstrate how convergence, script-mediated distance, and post-hoc standardization reshape how field data is segmented, compared, and interpreted when organized through genealogical labels. We argue that contact-aware, relational models of linguistic relatedness are necessary if NLP tools are to support, rather than distort, the documentation and analysis of linguistic diversity.

1 Introduction

Genealogical classification has long framed linguistic relatedness in terms of divergence, modeling languages as splitting and branching along tree-structured lineages. This view has been foundational for historical reconstruction and typology, but it is increasingly embedded—often implicitly—into computational tools used to organize, analyze, and preserve linguistic data. In contact-heavy regions such as South Asia, this abstraction captures only part of linguistic reality. Alongside divergence, the region exhibits persistent patterns of convergence across families and branches, produced through centuries of contact, cohabitation, and institutional multilingualism (Emeneau, 1956; Masica, 1976; Southworth, 2005). When computational systems inherit genealogical labels as organizing primitives, these dynamics are systematically obscured.

The Indo-Aryan–Dravidian divide illustrates this problem clearly. Although typological and official classifications present these families as distinct, extensive linguistic work shows that the boundary collapses in transitional regions shaped by sustained contact (Gumperz and Wilson, 1971; Thomason and Kaufman, 1988; Bashir, 2016). In such zones, cross-boundary alignment frequently outweighs internal divergence within standardized languages, giving rise to dialect continua and areal nexuses that resist discrete classification (Masica, 1991; Nichols, 1992). Treating these regions through rigid taxonomic lenses risks misrepresenting the linguistic systems documented through fieldwork.

These issues are not peripheral to field linguistics. Courtly bilingualism, shared literary registers, and long-standing diglossia enabled grammatical constructions, lexemes, and pragmatic conventions to circulate across what are now treated as firm genealogical divides (Ramanujan and Masica, 1969; Krishnamurti, 2003). Scriptal divergence, often taken as evidence of linguistic separation, frequently lags behind structural convergence and instead reflects post-hoc identity formation driven by political and administrative standardization (Southworth, 2005; Trautmann, 2006). When field linguistic data is computationally organized through genealogical categories, these historical processes directly shape how linguistic evidence is segmented, compared, and interpreted.

This paper argues that, in contact-heavy settings, genealogical classification functions poorly as a computational organizing principle for field linguistic data. Using South Asia as a critical case, we show that many distortions introduced by NLP-assisted analysis arise not from data sparsity or

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 8–15

modeling limitations, but from the imposition of tree-based abstractions onto contact-driven linguistic systems. Recognizing this mismatch is essential if computational tools are to support, rather than distort, the documentation and analysis of linguistic diversity.

2 Genealogical Classification as a Computational Prior

In computational workflows applied to linguistic documentation, genealogical classification rarely appears as an explicit analytical choice. Instead, it enters indirectly through the labels, resources, and organizational frameworks inherited from linguistic surveys, census practices, and standard-language corpora. ISO language codes, census categories, script conventions, and standardized orthographies together function as a de facto ontology of linguistic relatedness. When NLP tools are used to organize, compare, or archive linguistic materials, this ontology quietly structures what counts as a language, what is comparable, and what variation is treated as noise. In contact-heavy regions such as South Asia, these inherited categories encode assumptions of discreteness and divergence that are poorly aligned with the linguistic record (Grierson, 1903; Masica, 1991; Southworth, 2005).

Genealogical trees, originally developed as heuristic tools for historical reconstruction, thus come to function as computational ground truth in the processing of field and documentary data. Linguistic materials are partitioned into discrete labels; similarity is inferred across pre-defined language units; and cross-corpus comparison is constrained by presumed family membership (Campbell, 2003; Garrett, 2006). In this way, genealogical classification does not merely describe linguistic structure—it actively shapes how linguistic evidence is grouped, stored, and interpreted by computational systems.

This framing becomes especially problematic in regions characterized by dialect continua and areal convergence. As Emeneau's formulation of a linguistic area made clear, South Asia exhibits widespread structural diffusion across family boundaries that cannot be reduced to inheritance alone (Emeneau, 1956). Subsequent scholarship has documented how features such as retroflexion, case marking, clause-finality, and evidential strategies circulate across Indo-Aryan and Dravidian languages through sustained contact rather than descent (Masica, 1976; Thomason and Kaufman, 1988; Bashir, 2016). When computational tools continue to treat these languages as maximally distinct once genealogical boundaries are crossed, they impose a classificatory logic that obscures precisely the forms of continuity most salient in field linguistic evidence.

Scriptal differentiation further reinforces this abstraction. In South Asia, script is frequently conflated with language identity, despite extensive historical evidence that scriptal divergence often lags behind linguistic convergence and functions as an identity marker rather than a structural delimiter (Southworth, 2005; Trautmann, 2006). When scripts are treated as proxies for language boundaries in data organization and analysis, surface distance is amplified while deeper grammatical and lexical alignment is suppressed. This script-mediated distance is then reified in tokenization practices, representational spaces, and similarity measures, shaping how linguistic relatedness is inferred from field data.

The consequences of these assumptions are visible in computationally mediated linguistic inquiry. Transitional varieties are routinely misattributed to dominant standards or excluded altogether (Nichols, 1992). Measures of similarity overestimate distance across administrative or scriptal boundaries while underestimating divergence within standardized languages that exhibit substantial internal variation (Nichols, 1997). Cross-corpus comparison and reuse of field data consequently privilege the wrong affinities while overlooking structurally aligned contact zones (Kunchukuttan and Bhattacharyya, 2020). In such cases, distortion arises not from the linguistic data itself, but from the classificatory scaffolding through which that data is processed.

Importantly, these effects cannot be attributed solely to limitations of data or model capacity. They reflect a prior commitment to genealogical abstraction as the organizing principle of computational representation. Recent computational studies have begun to expose this mismatch. Dialect-level embeddings and geospatial clustering reveal similarity structures that align more closely with areal proximity and contact history than with genealogical family membership (Arora et al., 2021, 2022, 2023). Large-scale reanalyses of survey data likewise question whether genetic classification adequately captures South Asian language relationships (Borin et al., 2021). These findings do not

challenge historical linguistics; they highlight how selectively genealogical structure has been operationalized in computational treatments of linguistic evidence.

In this paper, we therefore treat genealogical classification not as neutral background context, but as an active computational prior. In contact-heavy settings, this prior systematically flattens networked, historically entangled linguistic ecologies into administratively convenient abstractions, shaping what computational tools can recognize as structure or variation in field linguistic data.

3 Contact Zones and Areal Nexuses in South Asia

South Asian linguistic structure cannot be adequately captured through discrete language units alone. Instead, it is organized around zones of sustained contact—areal nexuses in which linguistic features circulate across genealogical boundaries through prolonged multilingualism, shared institutions, and overlapping communicative domains. The concept of South Asia as a linguistic area, first articulated by Emeneau (Emeneau, 1956), foregrounded convergence as a co-equal force alongside divergence. Subsequent work has confirmed that areality is not a peripheral phenomenon, but a defining characteristic of the region's linguistic ecology (Masica, 1976, 1991; Southworth, 2005; Hook, 1987).

Areal convergence in South Asia operates at multiple linguistic levels. Phonological features such as retroflexion and vowel harmony, morphosyntactic patterns including postpositional case marking and clause-finality, and pragmatic strategies such as honorific alignment and evidential marking diffuse across Indo-Aryan and Dravidian languages through contact rather than inheritance (Ramanujan and Masica, 1969; Hook, 1976; Thomason and Kaufman, 1988). These shared structures do not eliminate genealogical distinctions, but they routinely outweigh them in actual language use, particularly in regions characterized by dense multilingual interaction (Nichols, 1992, 1997).

Dialect continua are most visible in transitional zones that lie between major phylogenetic divisions. Rather than exhibiting sharp boundaries, varieties in these regions form gradients of mutual intelligibility, lexical overlap, and structural alignment. Classical dialectological work has long documented such continua in South Asia, particularly along the Indo-Aryan–Dravidian interface and within Eastern Indo-Aryan (Grierson, 1903; Southworth, 2005). More recent regional studies of Kannada, Marathi, Konkani, Odia, and related varieties reinforce the view that linguistic distance increases gradually rather than categorically across space (Sridhar, 1990; Rane, 2010; Behera, 2006; Patnaik, 2015).

Institutional multilingualism has played a central role in sustaining these contact zones. Precolonial courts, religious institutions, and administrative systems routinely operated across multiple languages and registers, enabling grammatical constructions, lexemes, and stylistic conventions to circulate widely (Talbot, 2001; Zvelebil, 1973). Courtly bilingualism in particular fostered stable patterns of registeral alignment, where literary and bureaucratic norms were shared across languages without being perceived as foreign (Gumperz and Wilson, 1971). These practices produced layered linguistic ecologies in which speakers navigated multiple codes without rigid boundaries.

Scriptal differentiation, often treated as a proxy for linguistic separation in computational pipelines, must be understood within this institutional context. Historical evidence shows that scriptal divergence frequently follows linguistic convergence, crystallizing only when political, religious, or educational regimes seek to formalize identity (Southworth, 2005; Trautmann, 2006). In South Asia, the consolidation of distinct scripts for languages such as Kannada, Telugu, Tamil, and Malayalam reflects processes of standardization rather than deep structural rupture. As King and Pollock have shown in different contexts, scripts often function as symbolic markers of authority and identity rather than transparent reflections of linguistic distance (King, 1994; Pollock, 2006).

Colonial and postcolonial language administration further intensified this process. Large-scale surveys and gazetteers, while invaluable as documentary resources, imposed classificatory grids that privileged discrete languages over continua and standardized forms over local practice (Hunter, 1881, 1885). Educational policy and state formation in the twentieth century hardened these boundaries, aligning language, script, and territory in ways that obscured long-standing zones of overlap (Annamalai, 2001; Mohanty, 2019). Linguistic state reorganization after 1956 represents a particularly consequential moment, transforming fluid contact zones into administratively policed borders

(Trautmann, 2006; Patnaik, 2015).

For field linguistics and computational documentation, these historical processes are not merely background context. They determine how corpora are labeled, how scripts are segmented, and how linguistic relatedness is encoded in machine-readable form. When areal nexuses and dialect continua are flattened into discrete categories, computational representations mischaracterize similarity and erase contact-driven structure. The case studies that follow examine this dynamic in detail, showing how specific South Asian contact zones expose the limits of tree-based assumptions and motivate contact-aware, network-oriented approaches to computationally mediated linguistic analysis.

4 Case Studies: Convergence Against Classification

The following case studies illustrate how genealogical classification collapses in South Asian contact zones. Rather than presenting exhaustive historical surveys, each case foregrounds a specific mode of entanglement—morphological, scriptal, or registeral—that exposes the limits of tree-based models and motivates an areal, network-oriented account of relatedness.

4.1 Kannada–Telugu: Courtly Bilingualism and Morphological Interflow

While modern classifications assign Kannada and Telugu to distinct Dravidian subgroups—South Dravidian and South-Central Dravidian respectively (Krishnamurti, 2003)—their historical record reveals prolonged co-evolution rather than divergence. From the ninth to the fourteenth centuries, both languages functioned as courtly and inscriptional media under the Western Chalukyas, Hoysalas, and the Vijayanagara Empire, producing overlapping literary registers and shared administrative conventions (Talbot, 2001; Gumperz and Wilson, 1971).

Epigraphic evidence from regions such as Hampi, Bagali, Molkalmuru, and Anantapur demonstrates scriptal hybridity prior to the formal divergence of Kannada and Telugu scripts in the thirteenth century (Southworth, 2005). Contemporary spoken varieties along the Molkalmuru–Anantapur belt retain interoperable lemma inventories, case marking, verb-final constructions, and honorific systems, with variation largely restricted to phonological realization rather than grammatical function (Ramanujan and Masica, 1969; Krishnamurti, 2003). These patterns reflect a shared grammatical substrate differentiated only later through scriptal standardization and administrative boundary formation.

4.2 Tamil–Malayalam: Selective Divergence and Script-Mediated Distance

Tamil and Malayalam are often cited as a canonical example of genealogical divergence within Dravidian. Yet this divergence is highly stratified. Early Malayalam inscriptions and literary texts remain closely aligned with contemporaneous Tamil in core morphosyntax, with differentiation concentrated in lexicon, script, and register (Menon, 1933; Krishnamurti, 2003). Border varieties in regions such as Palakkad and Kanyakumari preserve high degrees of mutual intelligibility and shared grammatical structure (Ramanujan and Masica, 1969; Hook, 1976).

Much of the perceived distance between modern standard Tamil and Malayalam reflects post-medieval processes of Sanskritization, orthographic consolidation, and literary standardization. Scriptal divergence in particular amplifies surface distance while masking deeper structural continuity, functioning as an identity marker rather than a reliable indicator of grammatical separation (Southworth, 2005; Trautmann, 2006).

4.3 Marathi–Konkani–Tulu: Littoral Continua and Registeral Layering

Along the western coast, Marathi, Konkani, and Tulu participate in a littoral contact zone characterized by intense multilingualism and registeral layering. Konkani varieties in particular exhibit extensive lexical and morphosyntactic overlap with both Marathi and Kannada/Tulu, reflecting sustained contact rather than clear genealogical alignment (Rane, 2010; Sridhar, 1990). Script choice—Devanagari, Roman, or Kannada—further fragments representation without corresponding structural divergence.

Here, standard-language gravity pulls varieties toward administratively dominant centers, while everyday usage preserves hybrid systems that resist categorical classification. Computationally, this produces unstable language identification and distorted similarity judgments when script or standard form is treated as primary signal.

4.4 Bangla–Odia: Eastern Indo-Aryan and Administrative Separation

Bangla and Odia, both Eastern Indo-Aryan languages, share extensive phonological and morphosyntactic structure, particularly in transitional regions such as Medinipur and Ganjam (Masica, 1991; Behera, 2006). Their contemporary separation is reinforced by scriptal differentiation and colonial-era administrative boundaries rather than deep structural rupture.

As with Dravidian contact zones, dialect continua across the Bangla–Odia interface exhibit gradual rather than categorical change. When treated as maximally distinct units, these varieties are artificially distanced in computational representations despite substantial grammatical alignment.

4.5 Hindi/Hindustani Continua: Internal Diversity Under a Single Label

The Hindustani continuum illustrates a complementary failure mode: internal diversity collapsed under a single standardized label. Varieties such as Bhojpuri, Awadhi, Maithili, and Dakhani differ systematically in morphosyntax and lexicon, yet are routinely subsumed under "Hindi" in corpora and NLP pipelines (Grierson, 1903; Southworth, 2005). This flattening erases meaningful internal structure and produces predictable errors in language identification, similarity modeling, and transfer.

5 What This Distorts in Computationally Mediated Field Linguistics

The case studies above show that genealogical classification fails descriptively in South Asian contact zones. When computational tools are used to document, annotate, organize, and compare linguistic data, however, this failure acquires broader consequences. Treating genealogical labels as ground truth does not merely simplify linguistic reality—it reshapes how linguistic evidence is partitioned, aligned, and rendered legible within computational workflows supporting field linguistics. What follows is not a catalog of technical errors, but an account of how a classificatory prior propagates through computational mediation, with direct consequences for how linguistic structure is recorded and interpreted.

5.1 Language Identification as Documentary Misattribution

Computational tools used in field linguistics often presuppose that linguistic material can be unambiguously mapped to discrete language categories. In South Asia, this assumption collapses in contact zones and dialect continua. Varieties spoken along interfaces such as Kannada–Telugu, Bangla–Odia, or within the Hindustani continuum are not marginal or noisy instantiations of a single language, but stable hybrid systems shaped by long-term contact.

When such material is forced into genealogical labels during annotation or corpus organization, linguistic proximity is treated as error and administrative categories as signal. The result is not merely misclassification, but misattribution: texts, utterances, and speakers are reassigned to categories that obscure the linguistic systems they instantiate. Documentation practices that enforce categorical assignment further entrench this distortion, penalizing representations that capture genuine overlap while rewarding conformity to inherited taxonomies. Apparent inconsistency in field data thus reflects classificatory mismatch rather than deficiencies in the data itself.

5.2 Similarity Modeling and the Production of Artificial Distance

Computational analysis of field linguistic data frequently relies on similarity and distance measures to organize corpora, align varieties, or infer relatedness across datasets. When genealogical boundaries are assumed to define similarity space, these measures inherit a distorted geometry. Distance is exaggerated across scriptal or administrative boundaries and suppressed within standardized languages, even where internal variation is substantial.

In South Asia, where script-mediated distance often masks deep grammatical continuity, surface divergence becomes over-weighted. Representational spaces trained on standardized corpora encode this bias, producing similarity structures that mirror census categories more closely than contact history. As a result, varieties that share morphosyntactic and pragmatic structure are rendered artificially distant, while internally heterogeneous standards are treated as coherent units. The distances inferred through computational analysis are thus not discovered in the data, but produced by classificatory assumptions.

5.3 Cross-Corpus Comparison and Misplaced Affinities

Computational workflows increasingly support the reuse, aggregation, and comparison of field linguistic data across projects and corpora. In practice, genealogical classification often serves as a proxy for determining which materials are comparable or transferable. In contact-heavy settings, this proxy misaligns linguistic evidence.

Structurally aligned varieties that cross genealogical boundaries are excluded from comparison, while aggregation within standardized languages suppresses meaningful internal diversity. In South Asian contact zones, convergence-driven affinities are systematically overlooked, while administratively consolidated categories dominate corpus structure. Apparent incompatibilities across datasets therefore reflect misplaced assumptions about where similarity resides, rather than intrinsic differences in linguistic structure.

5.4 Annotation and Organization Against Administrative Abstractions

The most consequential effects of genealogical abstraction emerge at the level of annotation and data organization. When computational tools treat standardized language labels as ground truth, they structure field data according to administrative artifacts rather than linguistic evidence. In such regimes, capturing gradient similarity, overlap, or hybridity becomes difficult or impossible within available annotation schemes.

This has direct implications for field linguistics. Datasets that inherit rigid labels from census categories or ISO standards conflate linguistic adequacy with taxonomic conformity. Over time, tools, annotation practices, and corpora co-evolve around the same abstractions, reinforcing a closed loop in which classificatory schemes are reproduced rather than questioned. Linguistic variation documented in the field is thereby flattened, misaligned, or rendered invisible.

5.5 From Technical Limitation to Epistemic Misalignment

Taken together, these distortions point to a deeper issue. Persistent challenges in the computational handling of field linguistic data are not primarily technical. They reflect an epistemic misalignment between how linguistic relatedness is conceptualized and how linguistic evidence is organized for analysis. In South Asia, where convergence, areality, and institutional multilingualism are foundational, tree-based classification is not a neutral simplification but an interpretive distortion.

Recognizing this misalignment reframes the role of computational tools in field linguistics. The goal is not merely to scale existing annotation and organization practices, but to interrogate the classificatory assumptions that structure them. Without this shift, increasingly sophisticated tools risk producing ever more precise representations of increasingly impoverished abstractions, undermining the very diversity field linguistics seeks to preserve.

6 Discussion

The preceding analysis reframes a problem often treated as technical or task-specific in computational linguistics. Rather than viewing South Asian linguistic diversity as an unusually difficult setting for computational analysis, we argue that many persistent failures arise from a deeper misalignment between linguistic reality and classificatory abstraction. Genealogical labels, inherited from historical linguistics and institutionalized through census practice, script standardization, and corpus design, function as silent priors that shape how linguistic evidence is organized, annotated, and interpreted in computationally mediated fieldwork.

What distinguishes South Asia is not merely the presence of diversity, but the centrality of convergence. Linguistic structure in the region is shaped by sustained multilingualism, registeral layering, and institutional bilingualism, producing contact zones in which relatedness is relational rather than categorical. When computational tools assume divergence as the default and treat convergence as noise, they invert this reality. Hybrid and transitional varieties are mischaracterized precisely because they reflect the linguistic ecologies in which they emerge.

This perspective clarifies why incremental technical improvements often fail to resolve longstanding challenges in computational support for field linguistics. Improved models, better tokenization, or expanded datasets cannot compensate for annotation schemes and data structures that encode administrative abstractions as ground truth. As long as computational workflows privilege rigid labels over contact-driven structure, they will continue to obscure the very forms of variation that field linguistics seeks to document.

More broadly, the South Asian case exposes a general vulnerability in computationally mediated linguistic inquiry. Whenever genealogical classification is treated as exhaustive rather than partial—when trees are mistaken for terrain—computational tools risk organizing classification systems rather than linguistic evidence. For field linguistics, this distinction is not incidental: it determines what kinds of structure become visible, preservable, and interpretable.

7 Conclusion

This paper has argued that genealogical classification, while indispensable for historical reconstruction, becomes a liability when operationalized as computational ground truth in contact-driven linguistic ecologies. In South Asia, where convergence across families and branches is foundational rather than exceptional, tree-based assumptions systematically misrepresent linguistic relatedness, continuity, and variation.

Through historically grounded case studies, we showed how morphological, scriptal, and registeral entanglement undermines the assumptions embedded in computational treatments of language. These failures are not edge cases, nor are they artifacts of insufficient data or modeling capacity. They are the predictable consequences of imposing administrative and phylogenetic abstractions onto linguistic systems shaped by long-term contact.

We conclude that computational approaches supporting field linguistics must move beyond tree-based notions of linguistic organization. In contact-heavy regions, languages and varieties are better understood as nodes in overlapping, historically sedimented networks. Aligning computational tools with this reality is a prerequisite for documenting linguistic diversity without erasing it.

8 Future Work

Future work should translate this diagnostic account into computational practices that are explicitly contact-aware. Promising directions include annotation schemes that permit overlap and gradient membership, similarity measures grounded in areal proximity, and data models that represent continua rather than discrete endpoints.

Field-facing tools should also distinguish misannotation from faithful representation of hybridity, enabling documentation practices that preserve contact-driven structure rather than suppress it. Beyond South Asia, similar analyses should be extended to other regions characterized by long-term contact and dialect continua, including the Balkans, the Arabic-speaking world, and Romance dialect spaces.

Finally, sustained collaboration between NLP practitioners and field linguists is essential. Without close engagement with the epistemic commitments of fieldwork, computational tools risk reproducing inherited abstractions rather than supporting the preservation and analysis of linguistic diversity.

9 Limitations

This study is intentionally diagnostic and theory-driven. While it draws on a broad body of linguistic and computational scholarship, it does not introduce new datasets, tools, or empirical experiments. Its contribution lies in identifying structural failure modes in the computational mediation of linguistic evidence rather than in proposing immediate technical solutions.

Our focus on South Asia reflects the region's extreme linguistic density and long history of contact. Although analogous dynamics are likely present elsewhere, the extent to which these conclusions generalize beyond South Asia remains an empirical question.

Finally, this paper does not reject genealogical classification wholesale. Genealogy remains indispensable for many linguistic purposes. Our claim is narrower: in contact-heavy settings, genealogical labels are insufficient as organizing principles for computational documentation. Determining how genealogical and areal signals should be balanced in field-facing computational tools remains an open problem.

References

E. Annamalai. 2001. *Managing Multilingualism in India: Political and Linguistic Manifestations*. Sage, New Delhi.

A. Arora, A. F. Farris, S. Basu, and S. Kolichala. 2021. Bhasacitra: Visualizing the dialect geography of South Asia. arXiv.

A. Arora, A. F. Farris, S. Basu, and S. Kolichala. 2022. Computational historical linguistics and language diversity in South Asia. arXiv.

A. Arora, A. F. Farris, S. Basu, and S. Kolichala. 2023. JAMBU: A historical linguistic database for South Asian languages. arXiv.

Elena Bashir. 2016. Contact and convergence. In Hans Henrich Hock and Elena Bashir, editors, *The Languages and Linguistics of South Asia: A Comprehensive Guide*, pages 123–145. De Gruyter Mouton, Berlin and Boston.

D. Behera. 2006. The Odia language movement: A linguistic assertion. *Orissa Review*, 62(1):18–27.

Lars Borin, Anju Saxena, Bernard Comrie, and Shafqat Mumtaz Virk. 2021. A bird's-eye view on South Asian languages through LSI: Areal or genetic relationships? *Journal of South Asian Languages and Linguistics*, 7(2):203–237.

Lyle Campbell. 2003. How to show languages are related. In Roger D. Woodard, editor, *The Cambridge Encyclopedia of the World's Ancient Languages*, pages 108–120. Cambridge University Press, Cambridge.

Murray B. Emeneau. 1956. India as a linguistic area. *Language*, 32(1):3–16.

Andrew Garrett. 2006. Convergence in the formation of Indo-European subgroups: Phylogeny and the challenge of contact. In A. L. Sims, editor, *Historical Linguistics 2005*, pages 64–75. John Benjamins, Amsterdam and Philadelphia. Verify editor/booktitle details against your copy; chapter metadata varies by edition.

George A. Grierson. 1903. *Linguistic Survey of India*. Government of India Press. Published 1903–1928; Vols. 1–11.

John J. Gumperz and Robert Wilson. 1971. Convergence and creolization: A case from the Indo-Aryan/Dravidian border in India. In Dell H. Hymes, editor, *Pidginization and Creolization of Languages*, pages 151–167. Cambridge University Press, Cambridge.

Peter E. Hook. 1976. Case marking in South Asian languages: A survey. *Indian Linguistics*, 37:45–78.

Peter E. Hook. 1987. Linguistic areas: Getting at the grain of history. In George Cardona and Norman H. Zide, editors, *Festschrift for Henry Hoenigswald*, pages 155–168. Narr, Tübingen.

William Wilson Hunter. 1881. *The Imperial Gazetteer of India*, 1 edition. Trübner and Co., London. 9 vols.

William Wilson Hunter. 1885. *The Imperial Gazetteer of India*, 2 edition. Oxford University Press, Oxford. Published 1885–1887; Vols. 5–24.

Christopher R. King. 1994. *One Language, Two Scripts: The Hindi Movement in Nineteenth Century North India*. Oxford University Press, Oxford.

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

A. Kunchukuttan and P. Bhattacharyya. 2020. Leveraging language relatedness to improve low-resource machine translation. arXiv.

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

Colin P. Masica. 1991. *The Indo-Aryan Languages*. Cambridge University Press, Cambridge.

T. R. Menon. 1933. *A Primer of Malayalam Literature*. Asian Educational Services, New Delhi.

A. K. Mohanty. 2019. *Multilingualism, Education and Language Policy in India*. Springer, Singapore.

Johanna Nichols. 1992. *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago.

Johanna Nichols. 1997. Modeling ancient population structures in linguistics. *Annual Review of Anthropology*, 26:427–450.

D. Patnaik. 2015. Language identity and politics in eastern India. *Indian Journal of Linguistics*, 75(3):223–244.

Sheldon Pollock. 2006. *The Language of the Gods in the World of Men: Sanskrit, Culture, and Power in Premodern India*. University of California Press, Berkeley.

A. K. Ramanujan and Colin P. Masica. 1969. Toward a phonological typology of the Indian linguistic area. In Thomas A. Sebeok, editor, *Current Trends in Linguistics, Vol. 5: Linguistics in South Asia*, pages 543–577. Mouton, The Hague and Paris.

J. Rane. 2010. Konkani dialectology: Intracontinental variation. *Journal of Indo-Aryan Studies*, 24:57–74.

Franklin C. Southworth. 2005. *Linguistic Archaeology of South Asia*. Routledge, London and New York.

K. K. Sridhar. 1990. Kannada dialects and multilingualism in India. *International Journal of the Sociology of Language*, 86:99–119.

Cynthia Talbot. 2001. *Precolonial India in Practice: Society, Region, and Identity in Medieval Andhra*. Oxford University Press, Oxford.

Sarah G. Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley and Los Angeles.

Thomas R. Trautmann. 2006. *Languages and Nations: The Dravidian Proof in Colonial Madras*. University of California Press, Berkeley.

Kamil V. Zvelebil. 1973. *The Smile of Murugan: On Tamil Literature of South India*. Brill, Leiden.