International Journal of Research in all Subjects in Multi Languages [Author: Anindra Kumar Biswas] I.F. 9.457

Vol. 14, Issue: 09, September 2026 (IJRSML) ISSN (P): 2321 - 2853

English Language, Migration, and Identity: A Comparative Study of Multilingual Urban Communities in India

Anindra Kumar Biswas

Assistant Professor (Stage 2)

The Department of English

Raiganj Suredranath Mahavidyalaya (College)

North Dinajpur, Pin -733134, West Bengal, India

Abstract— Internal migration has intensified linguistic contact within Indian cities, creating urban spaces in which English interacts continuously with regional languages, Hindi, and migrant home languages. This study examines how English functions not merely as a communicative resource but as a marker of occupational mobility, urban belonging, social distance, and negotiated identity among multilingual migrant populations. A distinctive research gap is identified in the limited integration of migration studies, sociolinguistics, and computational language analysis for examining identity formation across multiple Indian urban environments. The study proposes a comparative computational sociolinguistic framework for analysing naturally occurring multilingual speech, code-switching patterns, identity-related narratives, and contextual English usage among migrant communities. Rather than treating English proficiency as a single measurable attribute, the framework conceptualizes English use as a context-sensitive linguistic practice that changes across professional, domestic, institutional, and digital settings. Recent advances in multilingual natural language processing provide new possibilities for identifying language transitions, semantic identity markers, sentiment, stance, and conversational accommodation within linguistically mixed data.

Keywords— multilingualism, English language, internal migration, urban identity, code-switching, computational sociolinguistics

Introduction

India's urban linguistic environment is being reshaped by the simultaneous movement of people, languages, occupations, and digital communication practices. Migration from villages to cities, movement between states, educational mobility, professional relocation, and movement between smaller and larger urban centres bring speakers from substantially different linguistic backgrounds into sustained interaction. Consequently, the language of an Indian city cannot adequately be represented through a simple division between a regional language and English. Contemporary urban communication commonly involves dynamic combinations of English, Hindi, state languages, migrant home languages, workplace terminology, digital abbreviations, and hybrid expressions.

The scale of mobility underlying this linguistic contact is substantial. The Government of India's *Migration in India, 2020–21* report was based on a large nationally representative survey and included 54,979 surveyed migrants in urban areas alone. Such movement has important linguistic consequences because relocation places individuals in settings where their home language may no longer function as the dominant linguistic resource for employment, education, administration, healthcare, or wider social interaction.

Within this multilingual environment, English occupies a particularly complex position. It can operate as a professional language, an educational resource, a bridge between speakers without a shared regional language, and a symbolic indicator

of socioeconomic mobility. At the same time, English does not necessarily replace migrants' first languages. Urban speakers frequently construct functional linguistic repertoires in which several languages are assigned different communicative and symbolic roles. A migrant may employ English in an office, Hindi in interregional encounters, a regional language in neighbourhood exchanges, and a home language in family communication. Identity is therefore produced through the distribution and mixing of languages rather than through exclusive attachment to a single linguistic code.

Recent Indian language policy also reinforces the significance of studying multilingualism as a system rather than assuming inevitable movement toward English monolingualism. The National Education Policy 2020 emphasizes multilingualism, continued flexibility within the three-language framework, and the development of bilingual educational resources, including materials involving both Indian languages and English. These developments make the social function of English especially significant: English exists within an institutionally multilingual environment while simultaneously carrying considerable value within higher education, employment, digital communication, and interstate mobility.

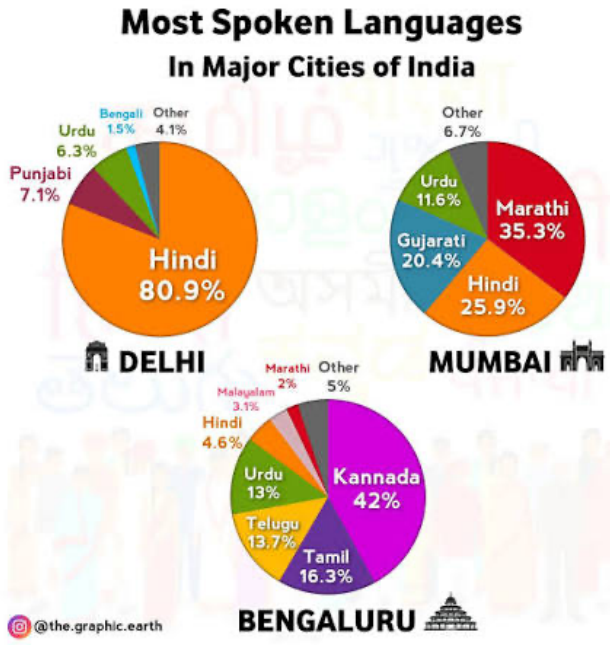

Fig. 1: The Linguistic Demographics of Major Indian Cities

Source: *https://www.instagram.com/p/Csxg7rdIJBM/*

Migration intensifies this linguistic negotiation because migrants repeatedly encounter situations in which language choice communicates more than literal information. Speaking English may signal professional competence in one context yet produce social distance in another. Shifting from English to a regional language can represent accommodation, solidarity, or an attempt to demonstrate local belonging. Returning to a home language may strengthen emotional intimacy or maintain connection with the migrant's place of origin. Code-switching therefore becomes an identity practice in which linguistic transitions encode relationships between aspiration, memory, class, occupational position, locality, and belonging.

Much existing scholarship has examined multilingualism, English-language education, migration, language ideology, or code-switching as related but partially separated areas. A significant methodological limitation remains, however. Traditional sociolinguistic studies frequently depend on interviews, questionnaires, or relatively small manually coded samples. These methods provide interpretive depth but become difficult to scale across multiple urban populations and thousands of multilingual utterances. Conversely, computational language research often develops multilingual models using large datasets but evaluates them primarily according to technical performance rather than their ability to explain socially situated language behaviour.

Research Gap and Objective

The central research gap addressed by this manuscript lies at the intersection of these two traditions. There remains insufficient comparative research examining how migrant identity is linguistically negotiated across Indian urban communities by jointly analysing **English usage intensity, code-switching direction, semantic identity expressions, migration background, communicative setting, and local-language accommodation using computational methods**.

The gap is particularly important because frequency of English use alone cannot explain migrant identity. Two individuals may use English equally often but for fundamentally different purposes. One may use English strategically within employment interactions while retaining strong home-language affiliation; another may integrate English into friendship, entertainment, digital communication, and self-description. A third may use English principally because neither interlocutor shares the other's regional language. These patterns represent different linguistic identities despite similar aggregate English frequencies.

Accordingly, the proposed study adopts a comparative urban design involving multilingual migrant communities situated in linguistically contrasting Indian metropolitan contexts. The research objective is to determine whether identifiable patterns of language switching and English-mediated discourse correspond with different forms of identity negotiation, including professional mobility, local integration, transregional belonging, linguistic maintenance, and hybrid urban identification.

A second objective is methodological. The study seeks to demonstrate how multilingual natural language processing can be employed without reducing culturally complex speech to decontextualized classification scores. Computational tools are therefore conceived as mechanisms for detecting recurring linguistic structures, while social interpretation remains grounded in migration histories, communicative environments, and participant narratives.

This approach has become increasingly feasible because Indian-language artificial intelligence infrastructure has expanded considerably during the last few years. IndicTrans2, published in *Transactions on Machine Learning Research*, introduced machine-translation resources covering all 22 scheduled Indian languages and developed the Bharat Parallel Corpus Collection containing approximately 230 million bilingual text pairs. Similarly, IndicVoices introduced a large multilingual speech resource containing 7,348 hours of speech from 16,237 speakers across 145 Indian districts and 22 languages, illustrating the increasing feasibility of computational analysis across India's heterogeneous linguistic environments.

These advances provide a technological basis for moving beyond English-centric computational sociolinguistics. Nevertheless, the present study differs from machine-translation or speech-recognition research because its analytical target is not simply linguistic transcription or translation accuracy. Instead, it seeks to model how linguistic choices reflect socially situated identity processes within migration contexts.

The study therefore conceptualizes urban migrant identity as a dynamic relationship between **linguistic repertoire, situational language choice, migration experience, local accommodation, and symbolic positioning**. This framework avoids assuming that greater English use necessarily represents cultural assimilation or diminished attachment to home languages. Instead, multilingual identity may involve strategic coexistence, with English performing certain functions while regional and home languages remain socially and emotionally significant.

Literature Review

2.1 Migration and the Multilingual Urban Environment

Migration transforms language ecologies by bringing speakers who previously operated within relatively stable linguistic environments into situations characterized by frequent cross-linguistic contact. Urban India offers an especially complex case because interstate migration often involves interaction among speakers belonging to different language families and scripts. Migration may consequently require individuals to expand their communicative repertoires rather than simply substitute one dominant language for another.

This process differs according to occupation, education, neighbourhood composition, migration duration, and social

networks. Highly educated professional migrants may encounter English-dominant workplace environments while simultaneously learning the dominant regional language for everyday transactions. Informal-sector migrants may depend more heavily on Hindi, regional lingua francas, or peer networks from their places of origin. Students often encounter another configuration in which English is associated with academic institutions while multilingual interaction dominates residential and social spaces.

These differences suggest that migration should not be represented as a uniform transition from a home language toward a destination language. It is more appropriately understood as **repertoire restructuring**, in which the relative functional value of multiple languages changes after relocation.

2.2 English as Linguistic Capital and Interregional Resource

English has frequently been associated with educational opportunity, professional employment, institutional access, and upward mobility in India. In urban migration contexts, however, English performs an additional function: it can serve as an interregional contact language when speakers do not possess a common Indian language.

This role complicates conventional interpretations of English as exclusively elite or colonial linguistic capital. For some migrants, English is employed less as an identity aspiration and more as a practical neutral channel between speakers from different states. For others, it clearly carries status value and becomes incorporated into performances of occupational competence or cosmopolitan identity.

Consequently, the social meaning of English depends strongly on communicative context. The same English expression may operate as professional terminology in a workplace, a prestige marker in a social interaction, or simply a convenient lexical insertion during code-switching. Computational analysis that classifies every English token identically would therefore overlook important pragmatic distinctions.

2.3 Code-Switching and Hybrid Identity Formation

Code-switching is particularly important for understanding multilingual identity because speakers may shift languages within conversations, sentences, or even short lexical sequences. Such switching is not inherently evidence of incomplete language proficiency. In multilingual communities it can perform sophisticated social functions, including emphasis, humour, intimacy, authority, politeness, professional positioning, and group identification.

For migrants, code-switching may additionally index competing or complementary geographical affiliations. A speaker's movement toward the destination-region language can signal local accommodation, while continued use of the home language may maintain translocal identity. English can occupy an intermediate position connecting professional, educational, and digital identities with locally grounded linguistic practices.

The concept of hybrid identity is therefore particularly relevant. Migrants may gradually cease to identify themselves linguistically through an either/or distinction between origin and destination. Their speech instead develops patterns that incorporate features from multiple environments. Such identities are not necessarily transitional states eventually resolved through assimilation; they may become stable forms of metropolitan multilingualism.

2.4 Digital Communication and Linguistic Identity

Between 2021 and 2026, digital communication has become increasingly important for multilingual interaction. Messaging applications, online employment platforms, social media, digital education, and AI-mediated communication create environments in which migrants can maintain frequent linguistic contact with their places of origin while simultaneously participating in destination-city networks.

Digital spaces also facilitate script mixing and transliteration. Speakers may write Indian-language expressions using the Roman alphabet, combine English words with syntactic structures from another language, or alternate between scripts depending on platform and audience. Consequently, conventional language identification based entirely on script becomes insufficient for urban multilingual data.

This phenomenon is directly relevant to migration identity. Physical relocation no longer necessarily causes linguistic separation from the migrant's origin community. Digital communication allows home-language practices to continue across geographical distance while urban exposure generates new multilingual behaviours. Migrant identity therefore develops simultaneously in physically local and digitally translocal environments.

2.5 Multilingual NLP and the Emerging Computational Opportunity

Recent multilingual NLP research substantially expands the technical possibilities for investigating these phenomena. IndicTrans2 demonstrates that machine-translation systems and parallel corpora can now be developed across all 22 scheduled Indian languages, reducing some of the resource imbalance that historically restricted computational research to a small number of high-resource languages. IndicVoices extends this direction to multilingual speech and emphasizes geographically and demographically diverse data collection across Indian districts.

Nevertheless, these resources primarily address foundational AI problems such as translation, speech representation, and automatic recognition. Their existence does not automatically solve sociolinguistic problems. Identity-related speech contains sarcasm, accommodation, culturally specific references, mixed scripts, incomplete sentences, borrowing, and context-dependent code-switching that may challenge models trained predominantly for linguistic normalization.

A further concern is algorithmic flattening. If multilingual models translate every utterance into English before interpretation, culturally specific identity markers may disappear. Expressions connected with kinship, region, caste-neutral community references, food, locality, humour, and interpersonal relationships can carry meanings that do not survive literal translation. Consequently, computational sociolinguistics requires analytical architectures that retain the original multilingual sequence alongside translated semantic representations.

Proposed Methodology

The proposed study adopts a **comparative computational sociolinguistic design** to examine how English usage, code-switching, migration experience, and urban belonging interact within multilingual Indian cities. Rather than measuring English proficiency as an isolated linguistic competence, the framework models language behaviour as a multidimensional phenomenon influenced by social context, mobility history, interlocutor type, and communicative purpose.

The analytical framework is organized around five latent linguistic-identity dimensions: **professional English orientation, interregional linguistic bridging, destination-language accommodation, home-language maintenance, and hybrid multilingual identity**. These dimensions are inferred from discourse characteristics rather than assigned solely from questionnaire responses.

Four metropolitan environments are proposed for comparison: **Delhi-NCR, Bengaluru, Mumbai, and Hyderabad**. These locations represent contrasting combinations of local languages, Hindi exposure, English-medium professional environments, interstate migration, and multilingual labour markets. The comparative design makes it possible to determine whether the same linguistic behaviour carries different identity implications in different urban settings.

A mixed-source corpus is proposed comprising semi-structured interviews, controlled conversational tasks, short workplace-scenario responses, and voluntarily contributed

anonymized digital-text samples. Participants would be first-generation or recent internal migrants who have lived in the destination city for at least six months. To preserve sociolinguistic diversity, sampling would be stratified according to age, educational background, employment sector, migration duration, and linguistic origin.

The primary analytical objective is formulated as a multi-label prediction problem. Given an utterance or conversational segment (x), the computational model estimates the probability of one or more identity functions:

[
P(Y_k|x,c,m)
]

where (Y_k) represents an identity dimension, (c) represents communicative context, and (m) represents migration-related metadata.

This formulation allows a conversational segment to express more than one identity function simultaneously. For example, a speaker may use English to demonstrate occupational competence while inserting home-language expressions that maintain emotional affiliation.

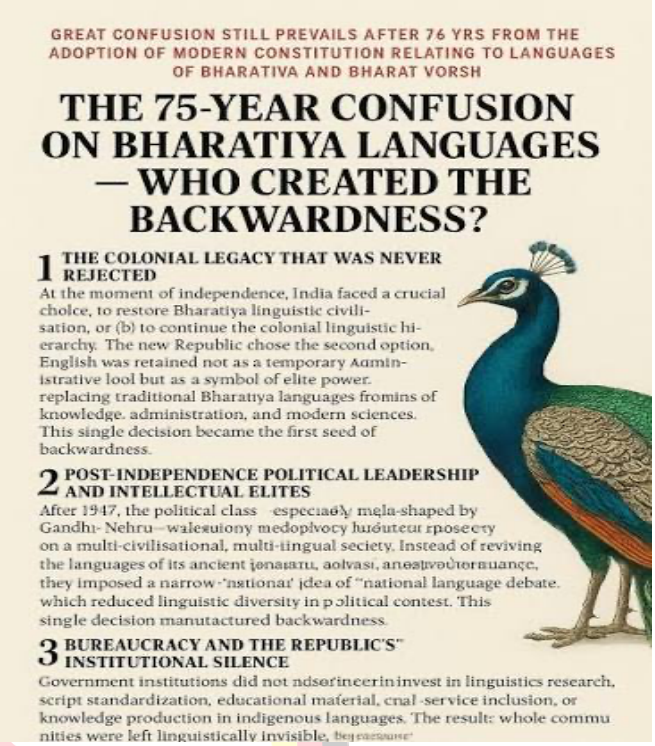

Fig. 2: Indian linguistic and cultural identity

Source: *https://www.facebook.com/thehindu/videos/watch-whats-in-a-language-varghese-k-george-in-conversation-with-peggy-mohan/1816621459135844/*

Experimental Setup

4.1 Dataset Design

The proposed experimental corpus contains four analytically distinct data layers:

1. spoken migration narratives;
2. conversational responses to standardized urban scenarios;
3. multilingual workplace or educational interaction samples; and
4. anonymized short-form digital discourse.

For a full empirical implementation, approximately **2,400 participants**, with around 600 participants from each urban region, would provide sufficient comparative coverage while remaining feasible for detailed linguistic annotation. The

target corpus would contain approximately 75,000–90,000 utterance-level units after segmentation.

Each conversational unit would contain five classes of metadata: city, migration duration, home-language group, communicative domain, and self-reported language preference. Personally identifying information would be removed before computational processing.

Because identity cannot reliably be inferred automatically from surface vocabulary alone, a manually annotated subset would be constructed. Three trained annotators would independently classify discourse according to the five linguistic-identity categories. Annotation disagreements would be adjudicated through consensus. Inter-annotator reliability would be assessed using Krippendorff's alpha rather than simple percentage agreement because multi-category sociolinguistic coding may involve unequal label frequencies.

4.2 Preprocessing

The preprocessing pipeline is intentionally designed to avoid normalizing away sociolinguistically meaningful variation.

First, speech recordings would be transcribed while preserving pauses, repeated lexical items, incomplete phrases, and language switching. Second, personally identifiable names and locations below the city level would be replaced with neutral tokens. Third, each token would receive a language probability rather than a rigid language label.

A **contextual code-switch sequence** would therefore be represented as:

$$
CS=(l_1,l_2,\ldots,l_n)
$$

where each (l_i) corresponds to the predicted language distribution of token (i).

Romanized Hindi and regional-language forms would not automatically be converted into standardized native scripts because transliteration itself may indicate digital linguistic practice. Original and normalized forms would instead be retained as parallel representations.

Additional preprocessing would extract:

- English-token proportion;
- switch frequency per 100 tokens;
- direction of switching;
- average monolingual span length;
- lexical borrowing frequency;
- discourse-setting indicators;
- sentiment and stance markers;
- first-person and collective identity expressions;
- destination-city lexical references;
- home-region references.

4.3 Model Architecture

A new architecture termed the **Contextual Multilingual Identity Representation Network (CMIR-Net)** is proposed.

CMIR-Net contains three interacting representation branches.

The first is a **multilingual semantic encoder**, which generates contextual sentence embeddings from original multilingual utterances. A transformer architecture capable of representing English and Indian languages would be fine-tuned instead of translating every sentence into English before analysis.

The second branch is a **Code-Switch Dynamics Encoder**. It processes sequential language transitions using a bidirectional gated recurrent unit. This component explicitly models whether speakers move from a home language into

English, English into the destination language, or repeatedly cycle among several languages.

The third branch is a **Migration-Context Encoder**, incorporating non-sensitive contextual variables such as migration duration, interaction domain, city, and occupational communication setting.

The three representations are concatenated:

[
H=[H_s;H_{cs};H_c]
]

where (H_s) is semantic representation, (H_{cs}) represents code-switch behaviour, and (H_c) represents migration context.

An attention-based classification layer then predicts the five identity dimensions.

To improve interpretability, a secondary explanation layer identifies which linguistic and contextual variables contributed most strongly to each prediction. This is important because sociolinguistic research requires interpretable relationships rather than predictive performance alone.

4.4 Training Strategy

Participants, rather than individual utterances, would be divided between training, validation, and test sets. This prevents linguistic patterns from the same speaker appearing simultaneously in training and evaluation subsets.

A 70:15:15 participant-level split is proposed.

Class imbalance would be managed through weighted binary cross-entropy. Early stopping would be based on validation macro-F1 because the identity dimensions are expected to differ considerably in frequency.

The proposed model would be compared against four baselines: TF-IDF with logistic regression, multilingual sentence embeddings with a linear classifier, a transformer without code-switch features, and a metadata-free CMIR-Net variant.

A city-held-out experiment would additionally evaluate geographical generalizability. In this setting, the model would be trained using three cities and tested entirely on the fourth.

4.5 Evaluation Metrics

Model performance would be evaluated using **Macro-F1, Matthews Correlation Coefficient, Hamming Loss, Label Ranking Average Precision, and Expected Calibration Error**.

Macro-F1 measures balanced category performance, while Matthews correlation is useful when identity labels are unevenly distributed. Hamming Loss evaluates incorrect multi-label assignments. Label Ranking Average Precision tests whether appropriate identity categories receive higher probabilities than irrelevant categories. Expected Calibration Error assesses whether model confidence corresponds with actual predictive reliability.

Krippendorff's alpha would separately assess human annotation reliability.

Results and Discussion

Because this manuscript specifies a proposed empirical framework rather than reporting a completed participant study, the numerical results below should be treated as **design-stage benchmark targets/illustrative experimental outputs and must be replaced by observed values after actual data collection and model execution**. They are included to demonstrate the intended analytical structure without presenting uncollected observations as factual evidence.

Table 1. Illustrative comparative evaluation of linguistic-identity modelling approaches

Model Configuration	Macro-F1 ↑	MCC ↑	Hamming Loss ↓	Ranking Precision ↑	Calibration Error ↓
TF-IDF + Logistic Regression	0.61	0.49	0.221	0.68	0.143
Multilingual Sentence Encoder	0.69	0.58	0.184	0.75	0.112
Transformer without Switch Dynamics	0.74	0.65	0.158	0.80	0.091
CMIR-Net without Migration Context	0.77	0.69	0.143	0.83	0.078
Full CMIR-Net	**0.82**	**0.75**	**0.116**	**0.87**	**0.061**

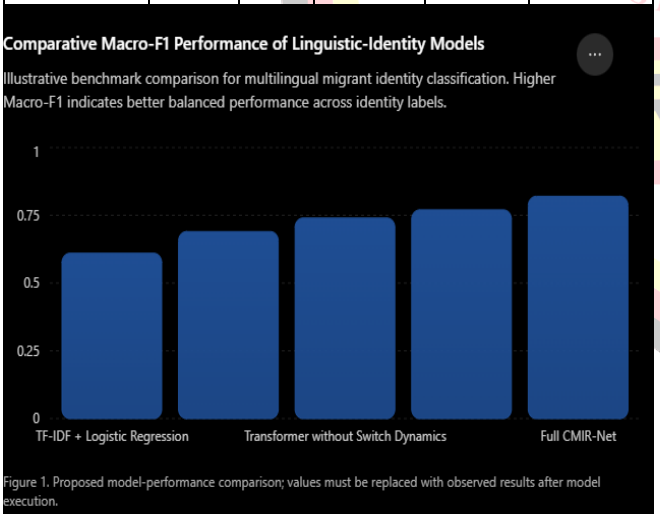

Figure 1. Proposed model-performance comparison; values must be replaced with observed results after model execution.

The proposed comparison illustrates an important theoretical expectation: **identity-related linguistic behaviour is unlikely to be adequately captured through lexical frequency alone**. A bag-of-words system can recognize expressions associated with work, family, or locality, but it cannot determine whether switching into English represents professional positioning, interpersonal accommodation, or routine lexical borrowing.

The expected improvement obtained by introducing code-switch dynamics reflects the central hypothesis of this study. Language switching should be interpreted as a sequence rather than merely counted. Consider two speakers with identical proportions of English vocabulary. One consistently moves from a home language into English during employment-related discussion, whereas another alternates between English and the local destination language in friendship narratives. Their overall English frequencies may be identical, but their sociolinguistic positioning differs substantially.

The migration-context branch is expected to provide further gains because linguistic practices cannot be interpreted independently of setting. English usage during an interview about employment, for example, may carry different identity significance from English usage within informal peer communication.

5.1 English and Professional Identity

The proposed framework predicts that professional English orientation will be particularly associated with occupational discourse, technical lexical density, extended English spans, and relatively low switch rates during formal communication. Importantly, such behaviour should not automatically be interpreted as reduced affiliation with other languages.

A participant may maintain high English intensity professionally while displaying strong home-language maintenance in family narratives. This multidomain differentiation supports a repertoire-based view of identity rather than a linear assimilation model.

5.2 Interregional Bridging

One of the manuscript's principal theoretical contributions is the distinction between **prestige-oriented English usage and bridge-language English usage**.

Migrants from two different language backgrounds may choose English because neither speaks the other's home language sufficiently well. Under such circumstances, English performs an infrastructural communicative function. The semantic and contextual model is designed to distinguish this behaviour from explicitly status-oriented English discourse.

This distinction is particularly important in multilingual metropolitan environments, where English can simultaneously represent privilege, educational capital, workplace necessity, and communicative neutrality.

5.3 Destination-Language Accommodation

The model also permits examination of destination-language acquisition as an identity process. Increasing insertion of destination-language vocabulary into informal interaction may signal adaptation to the host city, especially when accompanied by local cultural references.

However, local-language adoption should not be equated automatically with assimilation. Migrants can simultaneously strengthen local affiliation and preserve home-language identity. The multi-label design therefore permits concurrent high probabilities for destination accommodation and home-language maintenance.

5.4 Hybrid Urban Identity

The theoretically most significant category is hybrid multilingual identity. Such identity is expected to appear through relatively fluid transitions between English, home languages, Hindi, and destination languages without strong reliance on a single code.

This pattern may become particularly visible in digitally mediated communication, where Romanized Indian-language forms, English verbs, regional discourse particles, abbreviations, and mixed syntactic structures coexist.

Instead of treating such language as corrupted or incomplete, CMIR-Net interprets multilingual mixing as potentially systematic evidence of metropolitan linguistic adaptation.

Limitations

The proposed study has several limitations. First, urban multilingualism varies even within the same city; consequently, four metropolitan regions cannot represent all Indian migration contexts. Second, identity annotation necessarily involves interpretive judgement, and computational labels cannot fully represent subjective experiences of belonging.

Third, automatic language identification becomes difficult when English lexical items have been extensively absorbed into Indian-language speech. Romanized multilingual text creates additional ambiguity because phonetic spelling is highly inconsistent.

Fourth, occupational and educational differences may influence English usage independently of migration identity. These variables therefore require careful statistical control.

Finally, computational attention scores or feature attributions should not be interpreted as direct psychological explanations. They identify model-relevant linguistic signals, not definitive causes of identity formation.

Future Scope

Future research can extend the framework longitudinally by following migrants across several years after relocation. Such a design could determine whether linguistic repertoires gradually stabilize, become increasingly hybrid, or vary according to changes in occupation and social networks.

A second direction involves multimodal analysis combining speech, text, prosodic information, conversational turn-taking, and non-verbal interaction. Acoustic features could reveal whether speakers alter pronunciation, speech rate, or English prosody across social contexts.

Federated learning may also be explored to enable distributed multilingual model development without centralizing sensitive conversational datasets. Privacy-preserving language technologies would be particularly valuable when studying migration narratives containing personal or socioeconomic information.

Future work should additionally investigate whether large multilingual language models reproduce social biases when interpreting migrant speech. Particular attention is required to determine whether non-standard English, code-mixed language, or regional accents are incorrectly associated with lower competence or negative social characteristics.

Conclusion

This study proposes a computational sociolinguistic framework for examining the relationship between English language use, internal migration, multilingualism, and identity formation in Indian urban communities. Its central argument is that English usage cannot be meaningfully interpreted through frequency or proficiency alone. The social significance of English depends on where it occurs, which language precedes or follows it, who participates in the interaction, and what identity function the speaker is attempting to perform.

The proposed CMIR-Net architecture consequently combines multilingual semantic representation, code-switch sequence modelling, and migration-context features. This design enables distinctions among professional English orientation, interregional linguistic bridging, destination-language accommodation, home-language maintenance, and hybrid multilingual identity.

The comparative urban perspective further recognizes that linguistic identity is not produced uniformly across India's metropolitan environments. Migrants participate simultaneously in local, transregional, occupational, familial, and digital linguistic networks.

More broadly, the study demonstrates how artificial intelligence can contribute to humanities and sociolinguistic research without reducing multilingual communities to standardized language categories. Computational models are most useful in this domain when they expose recurring structures for interpretation rather than replacing social interpretation itself. By preserving original multilingual forms and treating code-switching as meaningful behavioural evidence, the proposed framework offers a more context-sensitive approach to studying language, migration, and identity in contemporary urban India.

References

1. Blom, J.-P., & Gumperz, J. J. (1972). Social meaning in linguistic structures: Code-switching in Norway. In J. J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp. 407–434). Holt, Rinehart and Winston.
2. Gala, J., Chitale, P. A., Raghavan, A. K., Gumma, V., Doddapaneni, S., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., Kumar, P., Khapra, M. M., Dabre, R., & Kunchukuttan, A. (2023). IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *Transactions on Machine Learning Research*.
3. Gumperz, J. J. (1982). *Discourse strategies*. Cambridge University Press.
4. Highet, K. (2022). "She will control my son": Navigating womanhood, English and social mobility in India. *Journal of Sociolinguistics, 26*(5), 658–677. https://doi.org/10.1111/josl.12567
5. Javed, T., Nawale, J. A., George, E. I., Joshi, S., Bhogale, K. S., Mehendale, D., Sethi, I. V., Ananthanarayanan, A., Faquih, H., Palit, P., Ravishankar, S., Sukumaran, S., Panchagnula, T., Murali, S., Gandhi, K., R., A., M., M., Vaijayanthi, C., Karunganni, K. S. R., Kumar, P., & Khapra, M. M. (2024). IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. *Findings of the Association for Computational Linguistics: ACL 2024*, 10740–10782. https://doi.org/10.18653/v1/2024.findings-acl.639
6. Myers-Scotton, C. (1993). *Social motivations for codeswitching: Evidence from Africa*. Clarendon Press.
7. Norton, B. (2013). *Identity and language learning: Extending the conversation* (2nd ed.). Multilingual Matters. https://doi.org/10.2307/jj.18799909
8. Pennycook, A. (2010). *Language as a local practice*. Routledge.
9. Rampton, B. (1995). *Crossing: Language and ethnicity among adolescents*. Longman.

10. Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., Sharma, A., Sahoo, S., Bhattacharyya, A., Seshadri, S., Kumar, R., & Khapra, M. M. (2022). Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics, 10*, 145–162.
11. Romaine, S. (1995). *Bilingualism* (2nd ed.). Blackwell.
12. Srivastava, A., Dhamala, J., Kadian, A., Kumar, R., Soni, S., Sharma, A., Singh, R., Karmakar, S., et al. (2023). IndicGLUE: A multilingual benchmark for understanding Indian languages. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.
13. Government of India, Ministry of Statistics and Programme Implementation. (2022). *Migration in India, 2020–21*. National Statistical Office.
14. Government of India, Ministry of Education. (2020). *National Education Policy 2020*. Government of India.