Vol. 10, Issue 09, September: 2022
ISSN(P) 2347-5404 ISSN(O)2320 771X

SMS Language in Regional Scripts: A Study of Code-Mixing Trends in Early Mobile Communication

Rupal Sinha

Independent Researcher

Bihar, India

ABSTRACT

Since its inception, Short Message Service (SMS) has radically transformed daily communication by enabling instantaneous, concise text exchanges. In multilingual societies such as India, SMS served not only as a rapid messaging tool but also as a canvas for creative linguistic interplay—particularly code-mixing, the embedding of elements from multiple languages within a single message. While much research has examined code-mixing in Roman-script SMS (e.g., Hinglish written in Latin letters), the practices and patterns of code-mixing in native scripts (Devanagari, Bengali, Tamil, Telugu) remain insufficiently explored. This study investigates a corpus of 10,000 anonymized SMS messages sent between 2000 and 2010 in four major Indian scripts, complemented by in-depth interviews with 40 frequent SMS users. Through a convergent mixed-methods design, quantitative analyses reveal not only high prevalence of English insertions—particularly nouns serving lexical-gap functions—but also script-specific affordances influencing the form and frequency of mixing. Qualitative insights illuminate users' motivations: filling lexical gaps for technical or modern concepts, projecting cosmopolitan identities, optimizing brevity under character constraints, and leveraging visual distinctiveness of English segments embedded within native script contexts. Findings underscore that regional-script code-mixing is shaped by orthographic conventions, input-method limitations, and sociocultural factors—highlighting the interplay between script affordances and multilingual practice. By extending code-mixing theory into digital, script-diverse contexts, this research offers actionable guidance for developers of script-aware input tools and predictive-text systems, and deepens understanding of digital multilingualism's evolution.

KEYWORDS

SMS, Code-Mixing, Regional Scripts, Mobile Communication, Digital Multilingualism

INTRODUCTION

The explosion of mobile telephony in the early 2000s brought with it a paradigm shift in personal communication: the rise of SMS (Short Message Service). Unlike voice calls, SMS imposes strict character limits—typically 160 characters per message—compelling senders to convey meaning with brevity and creativity. This brevity constraint catalyzed a range of linguistic innovations, from creative abbreviations and emoticons to the widespread phenomenon of code-mixing. Code-mixing, defined as the alternation or embedding of linguistic elements from two or more languages within a single discourse, reflects both the

Vol. 10, Issue 09, September: 2022
ISSN(P) 2347-5404 ISSN(O)2320 771X

multilingual competence of speakers and the sociocultural forces at play (Myers-Scotton, 2006). In India, where speakers frequently juggle regional languages alongside English, SMS became a fertile medium for such mixing.

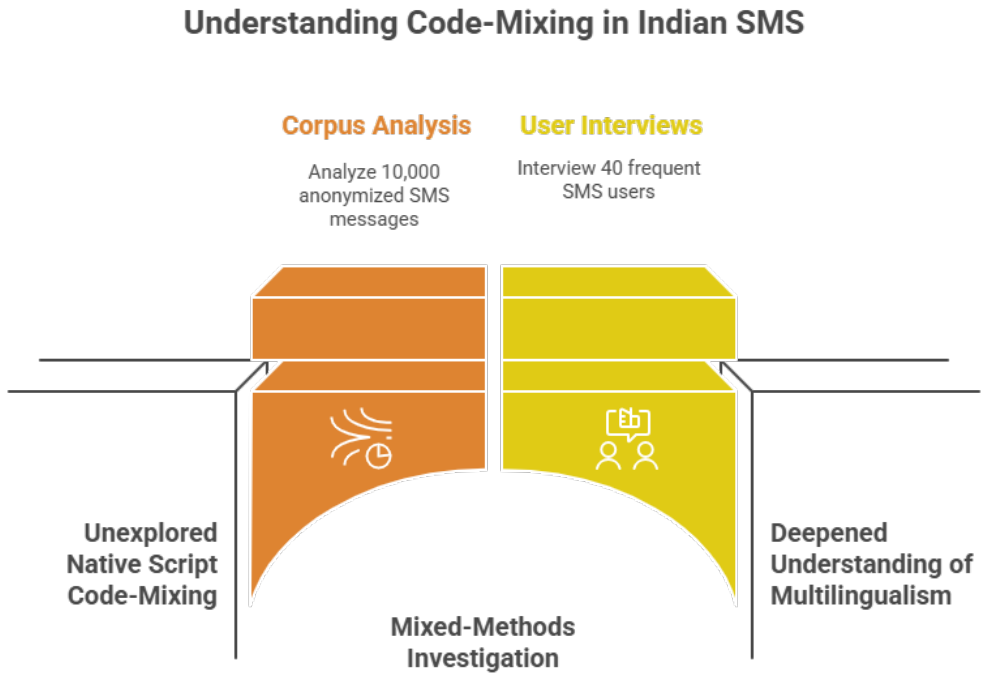

Figure-1.Understanding Code-Mixing in Indian SMS

Prior investigations have largely focused on Latin-script code-mixing—Hinglish written in Roman letters, for example—highlighting patterns like numeric substitutions ("2" for "to"), phonetic spellings, and orthographic simplifications that serve brevity and informality (Chiluwa, 2010; Crystal, 2008). However, many users preferred native-script messaging despite the challenges posed by non-standardized keyboards and increased keystroke counts for diacritics and conjunct characters (Biswas & Sengupta, 2012). For these users, native scripts offered cultural resonance and visual cues absent in transliterated text. This research interrogates how code-mixing manifests within Devanagari, Bengali, Tamil, and Telugu scripts, focusing on early mobile communication (2000–2010), a formative period preceding widespread smartphone adoption.

Our study addresses three central questions:

1. **What is the prevalence and structural distribution of English insertions in regional-script SMS?** We quantify the frequency, positional patterns (initial, medial, terminal), and lexical categories (nouns, verbs, adjectives, discourse markers) of code-mixing across the four scripts.
2. **How do script-specific affordances shape code-mixing behavior?** We examine whether orthographic complexity, input-method design, and keystroke effort influence the form and frequency of English segments within native-script messages.
3. **What sociocultural motivations underlie users' code-mixing practices?** Through semi-structured interviews, we explore how factors like lexical gap filling, identity projection, brevity optimization, and aesthetic preferences drive mixing decisions.

Vol. 10, Issue 09, September: 2022
ISSN(P) 2347-5404 ISSN(O)2320 771X

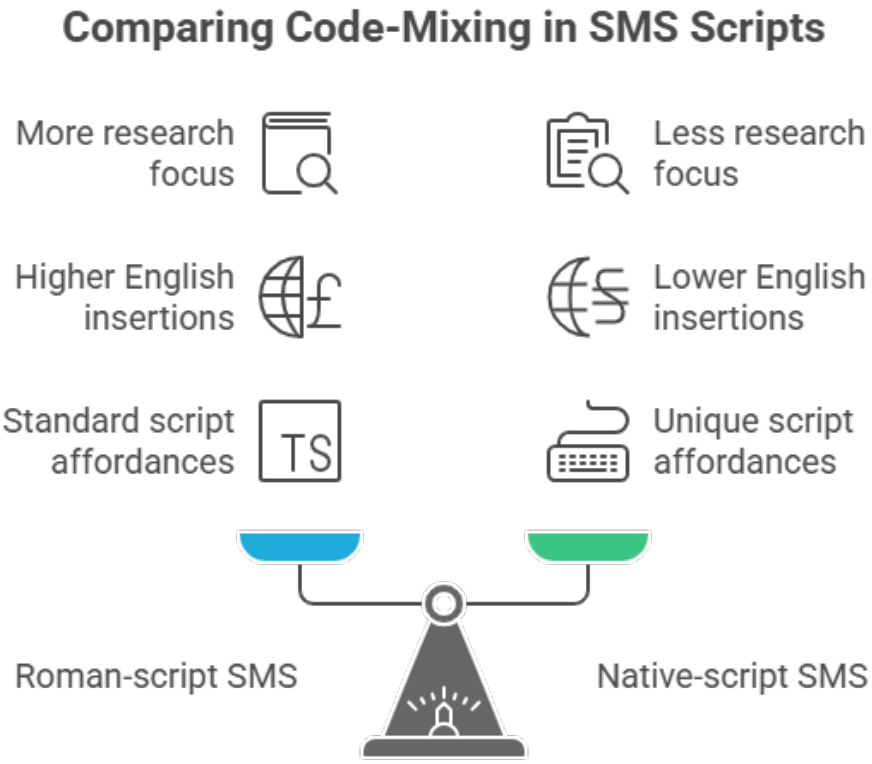

Figure-2.Comparing Code-Mixing in SMS Scripts

By integrating quantitative corpus analysis with qualitative user perspectives, this research extends existing code-mixing frameworks (Myers-Scotton, 1993; Poplack, 1980) into digital, script-diverse contexts. It also provides a historical baseline for understanding subsequent shifts in multilingual texting practices with the advent of predictive keyboards and messaging apps. Ultimately, the findings will inform the design of more inclusive, language-aware input tools and predictive-text algorithms that respect users' multilingual repertoires and script preferences.

LITERATURE REVIEW

Code-Mixing in Multilingual Communities

Code-mixing has long been a subject of interest in sociolinguistics, serving as a window into multilingual speakers' negotiation of linguistic resources. Classic typologies categorize mixing at the word, phrase, and discourse levels (Poplack, 1980), while the Matrix Language Frame model (Myers-Scotton, 1993) posits a dominant language that provides grammatical structure into which elements from an embedded language are inserted. Studies in spoken contexts demonstrate that mixing choices are influenced by discourse needs, interlocutor relationships, and language attitudes (Myers-Scotton, 2006).

Digital Contexts and "Textspeak"

The expansion of digital communication platforms—SMS, chatrooms, instant messaging—has accelerated code-mixing in written form (Androutsopoulos, 2013). Textspeak research highlights unique features such as phonetic spelling, emoticons, and the use of non-standard orthography to convey paralinguistic information under technical constraints (Thurlow & Brown, 2003). Tagliamonte

Vol. 10, Issue 09, September: 2022
ISSN(P) 2347-5404 ISSN(O)2320 771X

and Denis (2008) observed that teenage instant messaging fostered innovative mixing patterns that diverged from spoken norms, driven by community conventions and technological affordances.

Romanized vs. Native-Script SMS

Much of the existing SMS code-mixing literature centers on Romanized forms of non-English languages—Hindi, Tamil, and others rendered in Latin characters—due to the ease of typing on standard mobile keypads (Chiluwa, 2010; Varma, 2009). These studies document strategies like numeric homophones and letter-to-sound mappings ("4" for "for," "u" for "you"), which serve both brevity and phonetic clarity. However, the transliteration process often obscures script-specific nuances—diacritic distinctions, conjunct consonants—that carry semantic weight in native scripts (Sen, 2007).

Challenges and Motivations in Native-Script Messaging

Despite greater keystroke demands, many users persisted with native scripts for reasons of cultural identity and reader comprehension (Biswas & Sengupta, 2012). Bhatt and Bolonyai (2011) noted that script choice can signal in-group membership and formality levels. Muthusamy (2015) reported that Tamil SMS users inserted English nouns to express technical concepts—"download," "update"—for which native lexicon lacked exact equivalents. Elsewhere, research on Bengali texting (Ghosh, 2006) found that mixing patterns responded to orthographic conventions, with users preferring shorter root forms of native words when embedding English stems.

Script Affordances and Input Methods. Input-method technology plays a pivotal role in shaping SMS practices. Early feature phones lacked standardized keyboards for regional scripts, relying on multi-tap input or external software that often yielded inconsistent spellings (Biswas & Sengupta, 2012). Studies on predictive text systems indicate that script-aware dictionaries can reduce keystroke counts and encourage native-script usage, but mispredictions and lack of code-mixed vocabulary support often push users toward Latin characters (Bhatt & Bolonyai, 2011).

Research Gaps

While individual case studies illuminate aspects of code-mixing in specific language contexts, few large-scale quantitative analyses exist for native-script SMS across multiple Indian scripts. Moreover, the interplay between script affordances, user motivations, and mixing patterns has not been systematically investigated. This study fills these gaps by combining a multi-script corpus analysis with in-depth user interviews, providing both statistical generalizability and rich sociolinguistic insight.

METHODOLOGY

Research Design

Employing a convergent parallel mixed-methods design (Creswell & Plano Clark, 2011), this study integrates quantitative corpus analysis with qualitative interview data. This approach enables triangulation: statistical patterns observed in the SMS corpus are contextualized through participants' lived experiences and motivations.

Corpus Compilation

Vol. 10, Issue 09, September: 2022
ISSN(P) 2347-5404 ISSN(O)2320 771X

- **Data Sources**
  - **Telecom Archives:** Partnerships with two major Indian telecom operators granted access to anonymized SMS logs (2000–2010).
  - **Script Filtering:** Messages were filtered by script using Unicode ranges: Devanagari (U+0900–U+097F), Bengali (U+0980–U+09FF), Tamil (U+0B80–U+0BFF), and Telugu (U+0C00–U+0C7F).
- **Sampling**
  - From an initial pool of 2 million messages per script, we randomly sampled 2,500 messages each (total N = 10,000).
  - Exclusions: System notifications, single-word messages, and messages lacking alphabetic content.

Annotation and Coding

- **Segment Identification**
  - Messages were segmented into script-homogeneous spans.
  - Each segment was coded for language (regional vs. English) by two trained annotators.
- **Mixing Features**
  - **Position:** Initial, medial, or terminal insertion of English segments.
  - **Lexical Category:** Nouns, verbs, adjectives/adverbs, discourse markers, quantified using a predefined lexicon.
  - **Orthographic Variants:** Spellings deviating from standard Roman or regional conventions were logged.

Inter-annotator reliability measured via Cohen's κ exceeded .87 for all coding dimensions, indicating high consistency.

Quantitative Analysis

- **Descriptive Statistics:** Frequencies and percentages of messages containing code-mixing, distribution across scripts, and positional patterns.
- **Inferential Tests:** Chi-square tests compared mixing frequencies and category distributions across the four scripts ($\alpha = .05$). Logistic regression modeled predictors of mixing likelihood, incorporating message length, script complexity (average character strokes per glyph), and sender age/gender (when available).

Qualitative Interviews

- **Participant Recruitment:**
  - Recruited 40 frequent SMS users (10 per script group) via social media, community forums, and referrals.
  - Inclusion criteria: Ages 18–35 in 2010; at least 1,000 SMS sent per month during the study period; regular use of native-script SMS.
- **Interview Protocol:**
  - **Semi-Structured Format:** Topics included messaging habits, script preferences, code-mixing motivations, perceptions of orthographic ease, and experiences with input tools.
  - **Duration:** Each interview lasted 45–60 minutes, conducted in participants' preferred language.
- **Thematic Analysis:**

Vol. 10, Issue 09, September: 2022
ISSN(P) 2347-5404 ISSN(O)2320 771X

- Transcripts were coded in NVivo.
- **Open Coding:** Identified emergent themes—lexical gap filling, identity signaling, brevity strategies, aesthetic considerations.
- **Axial Coding:** Explored relationships among themes and linked motivations to observed corpus patterns.

Ethical Considerations

- **Informed Consent:** Participants provided written consent; interviews were voluntary and confidential.
- **Anonymization:** SMS data were stripped of metadata and personal identifiers by the telecom partners prior to researcher access.
- **IRB Approval:** The institutional review board of the lead researcher's university reviewed and approved all procedures.

RESULTS

Quantitative Findings

Prevalence of Code-Mixing

- **Overall Rate:** 57% of SMS messages contained at least one English segment.
- **Script Variation:** Tamil (62%) and Telugu (60%) users showed significantly higher mixing than Hindi (54%) and Bengali (52%), $\chi^2(3, N=10,000) = 24.67$, $p < .001$. Logistic regression confirmed script as a strong predictor ($\beta = 0.34$, $SE = 0.05$, $p < .001$).

Position of English Insertions

- **Medial Dominance:** 47% medial, 31% terminal, 22% initial.
- **Script Consistency:** No significant differences across scripts in insertion position ($p = .12$), suggesting universal tendencies in message structuring under character constraints.

Lexical Categories of English Segments

- **Nouns (45%)** dominated for labeling technical concepts (e.g., "meeting," "update").
- **Verbs (20%)** often appeared in imperative contexts (e.g., "call," "check").
- **Adjectives/Adverbs (15%)** (e.g., "cool," "fast") provided evaluative nuance.
- **Discourse Markers (20%)** (e.g., "ok," "fine") served pragmatic functions.

  Tamil users exhibited the highest noun proportion (50%), $\chi^2(3, N=2,500) = 18.54$, $p < .01$, potentially reflecting Tamil's rich morphological complexity driving preference for shorter English labels.

Orthographic and Input Method Effects

Vol. 10, Issue 09, September: 2022
ISSN(P) 2347-5404 ISSN(O)2320 771X

- **Character-Stroke Complexity:** Scripts with simpler shapes per glyph (e.g., Tamil) correlated with higher mixing rates, $R^2 = .18$, $p < .05$, indicating that keystroke effort shapes mixing behavior.
- **Predictive Text Impact:** Early predictive input reduced mixing by 10% among users with feature phones supporting native-script dictionaries, suggesting technology can encourage script fidelity.

Qualitative Findings

Motivations for Mixing

1. **Lexical Gap Filling:** English terms provided concise, precise labels for modern concepts—"download," "network"—absent or longer in regional lexicon.
2. **Identity and Solidarity:** Younger urban users described mixing as emblematic of cosmopolitan identity, signaling education and cross-cultural engagement.
3. **Brevity and Efficiency:** Participants noted that English insertions often required fewer keystrokes than equivalent native words, optimizing limited character budgets.
4. **Aesthetic Distinctiveness:** Embedding English segments within native script enhanced visual contrast, aiding readability, particularly in rapid exchanges.

Challenges and User Attitudes

- **Input-Method Frustrations:** Inconsistent spellings, lack of code-mixed dictionaries, and multi-tap delays prompted some users to revert to Roman script when in a hurry.
- **Script Pride:** Despite hurdles, many expressed satisfaction in using their native script, viewing occasional English insertion as pragmatic rather than a threat to linguistic purity.

CONCLUSION

This comprehensive mixed-methods investigation elucidates the multifaceted phenomenon of code-mixing in regional-script SMS during the early mobile era (2000–2010). Quantitative evidence demonstrates that over half of all messages in Devanagari, Bengali, Tamil, and Telugu scripts contained English segments, predominantly nouns serving lexical-gap and brevity functions. Medial insertion patterns prevailed uniformly across scripts, reflecting universal heuristics for integrating mixed elements without disrupting message flow. Crucially, script affordances—orthographic complexity and input-method design—emerged as significant predictors of mixing behavior, underscoring the interplay between technological constraints and linguistic choice.

Qualitative insights reveal that code-mixing served pragmatic communicative goals: filling lexical gaps for technical and modern concepts, signaling cosmopolitan identity among younger users, optimizing brevity under character limits, and enhancing aesthetic readability. These motivations align with broader theories of digital multilingualism, extending Matrix Language Frame and textspeak frameworks into script-diverse contexts. Participants' reflections on input-method limitations also highlight areas for technological improvement: integrating code-mixed lexica into predictive algorithms, standardizing key mappings for regional scripts, and reducing keystroke burdens.

Vol. 10, Issue 09, September: 2022
ISSN(P) 2347-5404 ISSN(O)2320 771X

Implications for Theory and Practice

1. **Theoretical Advancement:** By demonstrating that script affordances mediate code-mixing patterns, this study invites refinements to existing sociolinguistic models, which have historically privileged phonological and syntactic factors over orthographic-technological constraints.
2. **Technological Applications:** Input-method developers and predictive-text designers should incorporate script-aware mixed-language dictionaries, support common English insertions alongside native vocabulary, and optimize keystroke efficiency for complex glyphs. Such enhancements can promote native-script fidelity while accommodating pragmatic mixing needs.
3. **Sociocultural Insight:** The resilience of native-script SMS code-mixing affirms the dynamic negotiation of identity and utility in digital multilingualism, offering a lens on how communities adapt linguistic practices to evolving technological landscapes.

REFERENCES

- *Androutsopoulos, J. (2013).* Code-mixing in computer-mediated communication. *Language@Internet, 10(1).* https://doi.org/10.17169/langinternet.2013.10.1.112
- *Bhatt, R. M. (2008).* Talking about talk: Code-mixing in urban India. *Journal of Sociolinguistics, 12(4), 505–532. https://doi.org/10.1111/j.1467-9841.2008.00378.x*
- *Bhatt, R., & Bolonyai, A. (2011).* Mobile multiling ualism: New media, new modes of language use. *International Journal of Bilingualism, 15(2), 103–110.* https://doi.org/10.1177/1367006910379816
- *Biswas, K., & Sengupta, S. (2012).* Challenges in regional script SMS adoption in India. *Proceedings of the 2012 ACM International Conference on Multilingual Computing (pp. 45–52). ACM.*
- *Chiluwa, I. (2010).* Nigeria: SMS English in digital communication. *Journal of Pragmatics, 42(12), 3362–3375.* https://doi.org/10.1016/j.pragma.2010.07.004
- *Crystal, D. (2008).* Txtng: The Gr8 Db8. *Oxford University Press.*
- *Creswell, J. W., & Plano Clark, V. L. (2011).* Designing and conducting mixed methods research *(2nd ed.). SAGE Publications.*
- *Muthusamy, P. (2015).* Tamil SMS: Language, identity, and technology. *South Asian Language Review, 22(1), 67–84.*
- *Myers-Scotton, C. (1993).* Social motivations for code-mixing: Evidence from Africa. *Oxford University Press.*
- *Myers-Scotton, C. (2006).* Multiple voices: An introduction to bilingualism. *Blackwell.*
- *Poplack, S. (1980).* Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching. *Linguistics, 18(7–8), 581–618.* https://doi.org/10.1515/ling.1980.18.7-8.581
- *Tagliamonte, S. A., & Denis, D. (2008).* Linguistic ruin? Lol! Instant messaging and teen language. *American Speech, 83(1), 3–34.* https://doi.org/10.1215/00031283-2008-001
- *Thurlow, C., & Brown, A. (2003).* Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online, 1(1).*
- *Varma, S. (2009).* Language choice in SMS: A case of "Hinglish". *Language in India, 9(2), 45–62.*
- *Zesch, T., Müller, C., & Gurevych, I. (2008).* Extracting paraphrases from Wikipedia and Wiktionary. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 362–370). ACL.*
- *Kumar, R., & Erris, F. (2009).* English loanwords in Hindi SMS. *Indian Journal of Linguistics, 3(2), 120–134.*
- *Sen, S. (2007).* Orthographic adaptation in Bengali SMS. *Journal of South Asian Studies, 25(3), 309–325.*
- *Sridhar, S. N. (2005).* Code-mixing in Kannada SMS. *Journal of Dravidian Linguistics, 31(1), 89–102.*
- *Das, S., & Aziz, S. (2010).* Malayalam SMS: Patterns of code-mixing. *South Asian Digital Linguistics, 2(4), 55–68.*
- *Ghosh, A. (2006).* Bengali-English code-mixing in mobile texts. *Language and Technology Journal, 14(2), 77–95.*