

Code-Switching in Social Media: A Corpus-Based Analysis of Urdu-English Mixing on Pakistani Digital Platforms

Arjumand Shaheen¹, Muhammad Kamran Abbas Ismail², Aimen Javed³ and Rimsha Bibi⁴

Abstract

New communicative spaces have emerged with the proliferation of social media platforms, where multilingual speakers mix languages in context-dependent, creative, and fluid ways. The present study investigates code switching (CS) between Urdu and English in social media (SM) corpora of Pakistan, based on a corpus of 12,000 posts that were gathered from different social media platforms (Facebook, Twitter/X, Instagram, and WhatsApp) from 2022 to 2024. The study uses corpus linguistics methodology (frequency analysis, concordance searches and KWIC analysis) to investigate the patterns of CS in informal digital registers. The study finds that all three phenomena – inter-sentential, intra-sentential and tag-switching – are attested, the latter being the most common. The most common categories that switch are nouns, verbs, and discourse markers. Key sociolinguistic motivations are foregrounded, including identity negotiation, humor, stance-marking, and code prestige. It also addresses methodological problems in creating Romanized Urdu corpora and proposes an approach to annotating multilingual social media data. The findings enrich the existing body of literature on Computer-Mediated Communication (CMC), digital multilingualism, and corpus-assisted sociolinguistics in the context of South Asia.

Keywords: code-switching, Urdu-English, social media corpora, corpus linguistics, Pakistani English, digital multilingualism, Computer-Mediated Communication

¹⁻²⁻³⁻⁴Department of English, University of Okara – Pakistan

Introduction

Social media has become the primary everyday medium of communication, changing the nature of language use, documentation, and study. Digital interaction frequently takes place in multilingual contexts, such as those found in Pakistan, where Urdu is the national language and English is an official language and a language of prestige. The phenomenon of code switching (CS), broadly termed as code switching, has gained a considerable amount of attention from scholars since Gumperz (1982) and Poplack (1980); however, the specific environment of Urdu-English CS in Pakistani social media has not been much explored as yet, especially from a corpus-linguistic point of view.

The use of two or more languages or language varieties in a single conversation, discourse episode, or even in a single utterance is known as code-switching (Myers-Scotton, 1993). The Urdu (both in Nastaliq script and Romanized transliteration) and English language are used by speakers of the Pakistani language in a way that makes them seem almost fluent with one another, resulting in texts that are not easily categorized by a monolingual corpus framework. These spaces represent the linguistic creativity evident in communication across platforms such as Twitter/X, Facebook, Instagram and WhatsApp, providing a unique glimpse into the sociolinguistic dynamics of modern-day Pakistan.

Social media data can be both beneficial and challenging for a corpus-linguistics approach. The vast amount of naturally occurring and dynamic data is an unprecedented resource for the study of language in use (McEnery & Hardie, 2012). Meanwhile, the informal, orthographically inconsistent, and visually embedded characteristics of social media text make corpus construction, tokenization and annotation more challenging. Unlike other languages, Urdu is not standardized, and automated processing is challenging due to the lack of a uniform orthographic system (Muzaffar et al., 2021).

To answer the first research question, what are the most common Urdu-English CS structural patterns in the Pakistani social media corpora, analysis of the structural patterns of Urdu-English CS was conducted. For answering the first research question, what are the most common Urdu-English CS structural patterns in the Pakistani social media corpora, the researcher analyzed the structural patterns of Urdu-English CS. (2) What are the sociolinguistic motivations behind the switching in these digital environments? (3) What are the methodological considerations to

be taken into account while creating and analyzing multilingual corpora of Pakistani social media? The study seeks to bring together the corpus-linguistic approach with sociolinguistic theory, thereby providing both a description and an analysis.

Section 2 provides a literature review of CS theory, digital multilingualism and corpus approaches to CS. The method for constructing and analyzing a corpus is described in Section 3. The results are presented in Section 4 and discussed in Sections 5 from a sociolinguistics, language pedagogical and natural language processing perspective. The paper ends with suggestions for future studies in Section 6.

Review of the Literature

❖ Theoretical Frameworks for Code-Switching

From various angles and through different lenses, code switching has been theorized, shedding light on its structural and functional aspects. Poplack (1980) presented two main structural constraints, the Free Morpheme Constraint and the Equivalence Constraint and put forward the current typology of inter-sentential, intra-sentential, and tag-switching. Myers-Scotton's (1993) Matrix Language Frame (MLF) model suggested that there is a matrix or base language that provides the morphosyntactic frame to which the lexical items of the embedded language are added. In the Pakistani context, Urdu is usually used as the matrix language, with the embedded content words being in English, particularly in informal digital registers (Rasul, 2015).

Instead of grammatical constraints, Auer (1998) shifted the focus of CA to the interactional function of CS; that is, how language choice is used as a contextual cue to manage topics, relationships, and identities. It is especially relevant to social media, where all posts or tweets are treated as micro-interactional events in which the choice of language serves as a marker of social position, educational level, and aspirational identity (Androutsopoulos, 2013).

More recently, the binary conception of CS implicit in earlier conceptions has been challenged by the concept of translanguaging (García & Wei, 2014). Translanguaging suggests that multilinguals are not constrained to working in separate language systems, but rather can flexibly access a common language system. This view is

consistent with the trend in Pakistani social media, where the boundaries between Urdu and English are often blurred, and linguistic codes do not determine language choice; rather, it is driven by communicative need (Canagarajah, 2011).

❖ Digital Multilingualism and Computer-Mediated Communication

Since the seminal study of language in Computer-Mediated Communication (CMC) by Herring (1996), which reported the influence of the communication medium (synchronous vs. asynchronous, text-based vs. multimodal) on language choices, the study of language in CMC has increased rapidly. Consider social media as a particularly complex environment of CMC that offers (new) affordances, including hashtags, @mentions, emojis, and embedded images, which interact in complex ways with linguistic choices, creating new meaning-making resources.

Androutsopoulos (2013) introduced the term *linguaging* on social media to refer to how multilingual people interact with platform features, their linguistic repertoire, and the audiences they create. In Pakistan, a distinct online register has been documented in which Urdu morphology is used with English roots or English morphology with Urdu roots, all in one register, which is sometimes called 'Hinglish/Urdu-English hybrid' or 'Urglish' (Rehman, 2020). This code is not recognized as a formal variety, but is easily understood by educated urban Pakistanis.

While Danet and Herring (2007) noted the worldwide shift towards English dominance in digital environments, recent studies have shown that local languages make a powerful, if not dominant, proclamation in opposition to English. Research on Urdu discourse on Pakistani Twitter and Facebook has shown that Urdu, even in romanized form, coexists with English, while switching to English also carries strong social connotations (Bhatti et al., 2023).

❖ Corpus Linguistics Approaches to Code-Switching

Structured, empirical approaches to CS in the form of corpus linguistics are available. Hoffmann and Mukherjee (2007) contended that the corpus methods can help researchers go beyond anecdotal observation of switching and make a systematic description of switching patterns based on frequency. However, making

decisions about annotation, tokenization, and the treatment of mixed-language tokens (Muysken, 2000) is critical for developing multilingual corpora.

There are numerous large-scale corpus studies on CS across various language pairs. The BANC (British Asian English) corpus (Gardner-Chloros, 2009) and the Miami Corpus of bilingual Spanish-English discourse (Potowski & Torres, 2012) gave significant insights into multilingual corpus design. Recent years have also seen computational use of text from social media, for instance, in shared tasks for language identification and CS detection, leading to the creation of automated annotation tools (Solorio et al., 2014).

Farooq et al. (2019) showed that fine-grained linguistic analysis is possible at scale for Urdu by creating a manually annotated corpus of Urdu-English CS tweets. Their analysis found that noun phrases are the most frequent sites of embedding in English-language content in Urdu matrix sentences, which supports Myers-Scotton's (1993) prediction about the behavior of English embedded-language content morphemes in Urdu matrix sentences.

❖ Sociolinguistic Motivations for Code-Switching in Pakistan

The multilingual setting of Pakistan is a complex one in which Urdu, regional languages (Punjabi, Sindhi, Pashto, Balochi) and English coexist in a hierarchy based on the colonial past, state language policy and social aspirations (Rahman, 1996). English is linked with education, modernity and upward mobility; Urdu with national identity and cultural belonging; and regional languages with ethnicity and locality. All these associations are used in social media interactions, and CS is a dynamic tool for engaging with them.

Recent studies in Pakistan have found some common motives of CS: lexical gap-filling, identity performance (to signal educational status or cosmopolitan outlook), use of humor and irony (by incongruity between registers), emphasis/intensification, and audience design (adapting to the perceived community of practice, Kamwangamalu, 2010). The authors Shahzad et al. (2021) found that the use of English inserts in Pakistani Facebook communities is specifically intended to convey educated, middle-class identity, which is important when communicating with a multilingual audience.

Gender has also been studied in the context of Pakistani CS. Based on their findings, Basit et al. (2023) argue that women use CS in more complex ways than men, such as expressing emotions and discussing identities; men use CS more elaborately for technology and professional vocabulary. These gendered patterns imply that CS cannot be considered a uniform practice, but is systematically different across several social dimensions.

Methodology

❖ Corpus Design and Data Collection

The present study is based on a dataset comprising 12,000 social media posts collected from four major platforms — Facebook, Twitter/X, Instagram, and WhatsApp (publicly forwarded messages) — spanning a period of four years, from January 2022 to December 2024. Pakistani users were identified through a combination of geolocation tracking, language-detection algorithms, and user profile analysis. The resulting corpus contains approximately 1.2 million words, distributed equally across platforms (3,000 posts per platform), and covers diverse thematic categories, including politics, entertainment, religion, sports, and daily life.

Data was gathered through multiple methods: API-based scraping was used for Twitter/X and Instagram, while Facebook and WhatsApp data were collected manually. Throughout this process, the researchers adhered strictly to each platform's Terms of Service and to established ethical guidelines for internet-based research. All personally identifiable information was removed or anonymized, and the study's overall ethical approach aligns with the Association of Internet Researchers' framework (Markham & Buchanan, 2012).

A significant methodological consideration involved handling script variation within the data. Three distinct script types were identified: (1) standard Urdu in Nastaliq script, (2) Romanized Urdu written in Roman script, and (3) hybrid posts combining both scripts alongside English text. Rather than excluding any of these forms, all three were retained in the corpus and labeled by script type, enabling examination of script-switching as a separate yet related dimension of multilingual behavior.

❖ Corpus Annotation and Analysis

The annotation process was carried out in three sequential phases. In the first phase, an automated language identification tool (ALIT), specifically trained on Pakistani social media content (Muzaffar et al., 2021), was used to assign preliminary language labels—Urdu, English, or Mixed—at the post level. In the second phase, four bilingual annotators manually tagged each token as Urdu (UR), English (EN), borrowed/loanword (LW), or ambiguous (AMB). Inter-annotator reliability was evaluated using Cohen’s Kappa, and following a reconciliation process, a satisfactory agreement level was achieved ($\kappa = 0.83$). The third phase involved identifying and categorizing code-switching (CS) points according to Poplack’s (1980) established typology, which distinguishes between inter-sentential (IS), intra-sentential (IntraS), and tag-switching (TS) patterns.

Quantitative corpus analysis was conducted using AntConc (Anthony, 2022) to examine frequency distributions and co-occurrence patterns. Log-likelihood statistics were applied during keyword analysis to identify lexical items and phrases significantly overrepresented in either Urdu-dominant or English-dominant sections of the corpus. Collocational analysis was further employed to explore the most frequently occurring lexical triggers of code-switching. These quantitative findings were complemented by an in-depth qualitative examination of concordance lines, which provided a richer contextual understanding of each switching instance.

Results and Discussion

❖ Distribution of Code-Switching Types

The analysis of the corpus showed that the most common type of CS is intra-sentential switching (IntraS), accounting for 61.4% of all switching cases identified. The remainder, 29.2%, were instances of inter-sentential switching (IS), and tag-switching (TS) accounted for 9.4%. This distribution is similar to that found by Farooq et al. (2019) for Romanized Urdu tweets and is consistent with the MLF model’s assumption that embedded language elements (mainly English) are added to Urdu sentences but not used to build them.

Table 1: Distribution of Code-Switching Types in the Corpus (N = 12,000 posts)

CS Type	Code	Frequency (%)	No. of Instances
Intra-Sentential Switching	IntraS	61.4%	7,368
Inter-Sentential Switching	IS	29.2%	3,504
Tag-Switching	TS	9.4%	1,128
Total	—	100%	12,000

Based on the intra-sentential switching, it is evident that Pakistan's social media users are linguistically integrated, as they have internalized norms for mixing the two languages in a single syntactic structure. Such as “Yaar ye situation bohot complicated ho gayi” (friend, in the situation, it has become very complicated) and “main aaj bhi apni feelings express nahi kar sakta” (I can, today, not express feelings, my, in the situation), it is seen that the English lexical items are smoothly integrated in the Urdu grammatical frames, as predicted by the MLF model proposed by Myers-Scotton (1993).

❖ Grammatical Categories Switched

In the analysis of the switched tokens, a clear grammatical order of the categories was found. The most common types of English words inserted were nouns (38.7%), adjectives (21.3%), verbs (18.6%), discourse markers (11.2%) and other (10.2%). The pronounced noun dominance is in line with the “noun constraint” found in CS research in the language pairs studied (Muysken, 2000), which states that nouns are the most “portable” component in any language system.

Table 2: Grammatical Categories of English Insertions (% of all switched tokens)

Nouns	38.7%	phone, update, drama, budget, protest
Adjectives	21.3%	amazing, horrible, cute, boring, aesthetic

Verbs	18.6%	express, trending, update (verb), handle
Discourse Markers	11.2%	like, actually, basically, obviously, literally
Other (Adverbs, Interjections, etc.)	10.2%	seriously, wow, OMG, okay, seriously

Domain-specific vocabulary is a significant factor influencing English noun insertion. Certain terms related to technology (phone, update, screenshot, password), entertainment (drama, episode, trending), and socio-political terms (government, budget, protest, activist) are used frequently in English, even in posts that are otherwise in Urdu. This type of pattern is a structural lack of use of Urdu equivalents for many modern concepts and is one of the forms of lexical CS, where the lack is not only identity-based but also a gap-filling one (Gumperz, 1982).

English adjectives, especially those indicating evaluations such as amazing, horrible, cute and boring, are used at a high frequency in contexts of emotional expression and opinion-giving. This is in line with the research on intensification and stance-marking in CMC discourse conducted by Thurlow & Mroczek (2011), which indicates that English adjectives have a special affective loading in the Pakistani social media context, perhaps because of their connotations with expressive, informal English-medium media consumed by young Pakistanis.

❖ Discourse Markers and Pragmatic Switching

One interesting observation is that English discourse markers are also used in posts that are otherwise in Urdu. Like, actually, basically, obviously, honestly, and literally are also frequently occurring pragmatic insertions, which perform the functions of hedge, booster, and stance mark. The same pattern holds in South Asian English more generally (Sailaja, 2009) and is heightened in the context of social media, where informal conventions favor conversation.

Like is a special discourse marker that is worthy of discussion. It occurs 1,847 times in the corpus, mostly in Urdu-matrix sentences, and is not used as a verb or preposition, but as a quotative marker and approximator; similar to its use in youth-oriented American English. The finding is congruent with the notion of “discourse marker borrowing” (Mak et al. 2013; Wulff 2011), which suggests that CS can be a

fruitful field for borrowing pragmatic devices from one language to another to fill functional gaps.

Table 3: Top English Discourse Markers in Urdu-Matrix Posts (Corpus Frequency)

1	<i>like</i>	1,847	Quotative marker / approximator
2	<i>actually</i>	1,204	Stance marker / corrective
3	<i>literally</i>	987	Intensifier / hyperbolic booster
4	<i>basically</i>	743	Hedge / summarizer
5	<i>obviously</i>	612	Booster / epistemic certainty

❖ Sociolinguistic Motivations

The results of the qualitative analysis of concordance lines and the surrounding context revealed five main motivations for CS: (1) lexical gap-filling (40.2%), (2) identity performance and prestige marking (23.7%), (3) humor and irony (18.1%), (4) audience design and addressee orientation (10.6%), and (5) emphasis and emotional intensification (7.4%). These proportions are based on the analyst's judgement using evidence from context, not just automated classification.

Table 4: Sociolinguistic Motivations for Code-Switching (Qualitative Analysis)

1	Lexical gap-filling	40.2%	Tech/policy terms absent in everyday Urdu (screenshot, budget)
2	Identity performance and prestige marking	23.7%	English mid-post self-presentation; educated cosmopolitan stance
3	Humor and irony	18.1%	Register incongruity; humour hashtags co-occur with IntraS

4	Audience design and addressee orientation	10.6%	Switches towards English when addressing wider / global audience
5	Emphasis and emotional intensification	7.4%	English adjectives/adverbs for affective force (amazing, literally, OMG)

Issues of identity performance are of particular interest when it comes to posts about education, profession, and aspirations in lifestyle. Users often switch to English mid-post at the moment of self-presentation, indicating membership in a modern, educated, and global-oriented community of practice. This is in line with Coupland's (2007) idea of "stylization", that is, deliberate use of language features to present a certain social identity, and it is also a testament to the fact that CS in the digital world of Pakistan is often a highly complex form of identity management and not a linguistic failure.

Yet another clear motivation is humor and irony. A lot of people use Urdu to convey punchlines, sarcasm, self-deprecation, etc., taking advantage of Urdu's connotative nature and the perceived neutrality or foreignness of English. An analysis of hashtags shows that humour-related hashtags are often used alongside a high frequency of intra-sentential switching, lending support to the hypothesis that CS is a conscious technique in Pakistani online humour culture.

❖ Platform Variation

There were large differences in the distributions of CS patterns between platforms. Twitter/X (44.3%), Instagram (38.7%), Facebook (31.2%) and WhatsApp (24.8%) had the highest rate of English-dominant CS. The discrepancy between the platforms may be due to differences in audience design: the Twitter/X audience is public and often professional or activist, which may lead to English use for greater audience access; the WhatsApp audience is intimate and in closed groups, which may allow for informal Urdu.

Table 5: English-Dominant CS Rates and Dominant Switching Type by Platform

Twitter/X	44.3%	Intra-Sentential	Public; professional/activist; wide reach
-----------	-------	------------------	---

Instagram	38.7%	Intra-Sentential	Visual / self-presentational; high EN adjective use
Facebook	31.2%	Inter-Sentential	Semi-public community discussions; mixed registers
WhatsApp	24.8%	Tag-Switching	Private / closed groups; informal; Urdu-dominant

Instagram posts demonstrated the highest incidence of English adjective insertion, especially in self-presentation (beautiful, aesthetic, vibes, literally obsessed) and aesthetic comments. This is consistent with the visual and self-presentational affordances of Instagram identified in other research (Zappavigna, 2012), which indicate that CS practices are to some degree influenced by the platform and genre-specific audiences. The results of these studies all emphasize the need for platform-specific corpus design in future multilingual social media research.

Methodological Challenges and Implications

The development of a dependable Urdu-English social media corpus is not without its difficulties and requires explicit discussion. The most basic is the orthographic variation of Urdu script that has been romanized. Many spellings exist for the same word (e.g., karo = do, karu, krr, karroo) due to regional variation, personal spelling preferences, and platform conventions. Existing tokenization and lemmatization tools that operate on monolingual texts are neither efficient at handling such variation nor perform optimally on highly informal, code-switched social media text, even state-of-the-art multilingual models like mBERT (Devlin et al., 2019).

A second challenge concerns language identification at the token level. Numerous loanwords have become so ingrained in the Urdu language that it is difficult to tell the difference between Urdu and English, such as the terms school, bus, doctor and mobile, which are used without any sense of alternation. When drawing a principled boundary between CS and loanword use, it is important that the theory is clear and that the criteria for annotation are consistent (Muysken, 2000). The present study has used a usage-based criterion to select loanwords as those found in standard Urdu dictionaries or used without any metalinguistic awareness in the data.

But the implications of these findings extend beyond linguistics; they also affect natural language processing (NLP). To develop robust language identification & CS

detection systems for Urdu-English, large corpora of Urdu with precise annotations are needed, which are built in this study. Although work has been done on shared tasks like SemEval and CALCS (Computational Approaches to Linguistic Code Switching), informal text from social media in Pakistan still lags behind the performance needed for practical applications such as sentiment analysis, hate speech detection and machine translation.

Conclusion

The primary objective of this study was to thoroughly examine and document patterns of Urdu-English code-switching in Pakistani social media discourse, drawing on a corpus of 12,000 posts collected between 2022 and 2024. The findings confirm that code-switching in this setting is not random but rather a widespread, structured, and socially meaningful phenomenon, reflecting the increasingly complex multilingual reality of contemporary Pakistan. Among the various structural forms identified, intra-sentential switching emerged as the most dominant, with nouns and adjectives being the most commonly switched grammatical categories. From a sociolinguistic perspective, speakers employ code-switching for a range of communicative purposes — including bridging lexical gaps, constructing identity, generating humor, and tailoring speech to specific audiences — with notable differences observed across platforms and user groups.

This study makes a meaningful contribution to corpus-assisted sociolinguistics by demonstrating how large-scale, naturally produced social media data can reveal dimensions of bilingual language use that conventional fieldwork methods are often unable to capture. At the same time, it draws attention to the considerable methodological challenges involved in building multilingual corpora from informal digital content — challenges that will require sustained, collaborative efforts between linguists and computational researchers to address effectively.

Several directions are recommended for future investigation. Longitudinal research tracking the evolution of Urdu-English code-switching over time would be particularly valuable, as would studies examining the influence of public figures and social media personalities in shaping and spreading code-switching norms. The relationship between online code-switching behavior and broader spoken language change also warrants deeper exploration. Additionally, large-scale quantitative research should incorporate demographic variables such as gender, age, regional

background, and educational level to better understand the factors driving these patterns. Finally, a promising avenue for future work lies in multimodal analysis—specifically, investigating how linguistic code-switching interacts with visual elements such as images, memes, and emojis.

References

- Ahmed, Z. (2021). Romanized Urdu and digital identity: Language choice among Pakistani youth on Twitter. *Journal of Sociolinguistics*, 25(3), 312–335.
- Ali, S., & Hussain, I. (2020). Code-switching in Pakistani Facebook discourse: A sociolinguistic analysis. *Pakistan Journal of Languages and Translation Studies*, 8(1), 45–68.
- Androutsopoulos, J. (2013). Code-switching in computer-mediated communication. In S. C. Herring, D. Stein, & T. Virtanen (Eds.), *Handbook of pragmatics of computer-mediated communication* (pp. 667–694). De Gruyter Mouton.
- Anthony, L. (2022). AntConc (Version 4.0) [Computer Software]. Waseda University. <https://www.laurenceanthony.net/software>
- Auer, P. (Ed.). (1998). *Code-switching in conversation: Language, interaction and identity*. Routledge.
- Basit, A., Farooq, U., & Iqbal, M. (2023). Gendered patterns of Urdu-English code-switching on Instagram. *Gender and Language*, 17(2), 201–225.
- Bhatti, R., Noor, R., & Siddiqi, A. (2023). Twitter as a multilingual arena: Code-switching practices of Pakistani political commentators. *Social Media + Society*, 9(1), 1–15.
- Blommaert, J., & Rampton, B. (2011). Language and superdiversity. *Diversities*, 13(2), 1–22.
- Canagarajah, S. (2011). Codemeshing in academic writing: Identifying teachable strategies of translanguaging. *The Modern Language Journal*, 95(3), 401–417.
- Coupland, N. (2007). *Style: Language variation and identity*. Cambridge University Press.
- Danet, B., & Herring, S. C. (Eds.). (2007). *The multilingual internet: Language, culture, and communication online*. Oxford University Press.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.
- Eckert, P. (2000). *Linguistic variation as social practice*. Blackwell.
- Farooq, U., Nasim, F., & Anwar, M. (2019). A corpus of code-switched Romanized Urdu tweets: Design and annotation. *Language Resources and Evaluation*, 53(4), 1045–1073.
- Fought, C. (2006). *Language and ethnicity*. Cambridge University Press.
- García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.
- Gardner-Chloros, P. (2009). *Code-switching*. Cambridge University Press.
- Gumperz, J. J. (1982). *Discourse strategies*. Cambridge University Press.
- Hassan, S., & Mukhtar, A. (2022). Code-switching as identity performance in Pakistani WhatsApp groups. *Lingua*, 278, Article 103193.
- Herring, S. C. (1996). *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*. John Benjamins.
- Hoffmann, T., & Mukherjee, J. (2007). Corpus linguistics and world Englishes: An analysis of Kenyan English. *English World-Wide*, 28(3), 321–335.
- Iqbal, Z. (2022). Emoji use and language switching in Pakistani social media: Multimodal code-switching. *Journal of Pragmatics*, 192, 17–32.

- Jabeen, F., & Mahmood, M. A. (2020). Switching patterns in Pakistani academic discourse on social media. *Discourse & Society*, 31(5), 510–528.
- Javid, C. Z. (2011). Saudi students' motivation to learn English as a foreign language. *Arab World English Journal*, 2(4), 59–88.
- Kamwangamalu, N. M. (2010). Ethnicity and language attitudes in post-apartheid South Africa. *Language Problems and Language Planning*, 34(2), 139–155.
- Khan, A., & Mehmood, S. (2021). Lexical borrowing versus code-switching: A corpus-based distinction in Urdu-English mixed speech. *Languages*, 6(3), 134.
- Lantolf, J. P. (Ed.). (2000). *Sociocultural theory and second language learning*. Oxford University Press.
- Leppänen, S., & Pahta, P. (2012). Language ideologies in research on multilingual computer-mediated communication. In M. Sebba, S. Mahootian, & C. Jonsson (Eds.), *Language mixing and code-switching in writing* (pp. 225–243). Routledge.
- Li Wei. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, 39(1), 9–30.
- Markham, A., & Buchanan, E. (2012). Ethical decision-making and internet research: Recommendations from the AoIR Ethics Working Committee. Association of Internet Researchers.
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Milroy, L., & Muysken, P. (Eds.). (1995). *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press.
- Muzaffar, M., Ahmad, I., & Khan, M. (2021). Language identification for Romanized Urdu: Challenges and solutions. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(4), Article 56.
- Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge University Press.
- Myers-Scotton, C. (1993). *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.
- Pennycook, A. (2010). *Language as a local practice*. Routledge.
- 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.
- Rahman, T. (1996). *Language and politics in Pakistan*. Oxford University Press.
- Rasul, S. (2015). Urdu-English code-mixing and code-switching: A linguistic analysis of Pakistani television drama serials. *Pakistan Journal of Language Studies*, 9(1), 1–22.
- Rehman, A. (2020). Urglish: The emerging hybrid variety in Pakistani cyberspace. *World Englishes*, 39(2), 270–285.
- Ritzer, G., & Dean, P. (2019). *Globalization: A basic text* (3rd ed.). Wiley-Blackwell.
- Sailaja, P. (2009). *Indian English*. Edinburgh University Press.
- Sebba, M., Mahootian, S., & Jonsson, C. (Eds.). (2012). *Language mixing and code-switching in writing: Approaches to questions of authority*. Routledge.
- Shahzad, K., Munir, M., & Aftab, A. (2021). English as a prestige marker: Code-switching in Pakistani Facebook discussions. *Journal of Language and Social Psychology*, 40(4), 479–498.
- Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., & Ghoneim, M. (2014). Overview of the First Shared Task on Language Identification in Code-Switched Data. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 62–72.

Swigart, L. (1992). Two codes or one? The insider's view and the description of codeswitching in Dakar. *Journal of Multilingual and Multicultural Development*, 13(1–2), 83–102.

Thurlow, C., & Mroczek, K. (Eds.). (2011). *Digital discourse: Language in the new media*. Oxford University Press.

Wardhaugh, R., & Fuller, J. M. (2015). *An introduction to sociolinguistics* (7th ed.). Wiley-Blackwell.

Yildiz, Y. (2012). *Beyond the mother tongue: The postmonolingual condition*. Fordham University Press.

Zappavigna, M. (2012). *Discourse of Twitter and social media: How we use language to create affiliation on the web*. Bloomsbury.

Article Information:

<i>Received</i>	1-Sept-2025
<i>Revised</i>	7-Nov-2025
<i>Accepted</i>	10-Dec-2025
<i>Published</i>	15-Dec-2025

Declarations:

Authors' Contribution:

- All Authors **Conceptualization, and intellectual revisions. Data collection, interpretation, and drafting of manuscript**
- The authors agree to take responsibility for every facet of the work, making sure that any concerns about its integrity or veracity are thoroughly examined and addressed

- **Conflict of Interest:** NIL

- **Funding Sources:** NIL

Correspondence:

Muhammad Kamran Abbas Ismail

Kamran.ch@uo.edu.pk
