AUTOMATIC AI-BASED ANALYSIS OF CODE-SWITCHING AND CODE-MIXING INVOLVING ENGLISH

Isomova Feruza Nizom qizi
Master's Alumna of Webster University in Tashkent

Annotation: *This article examines the growing phenomenon of Uzbek–English code-switching and code-mixing in digital, educational, and social contexts in Uzbekistan. It highlights how globalization, technological expansion, and increasing English usage have accelerated bilingual communication among young speakers. The study emphasizes the role of Artificial Intelligence and Natural Language Processing in analyzing linguistic boundaries, morphological integration, and pragmatic functions in mixed-language discourse. Particular attention is given to the challenges posed by the agglutinative structure of Uzbek and the limited availability of annotated corpora. The article also reviews how modern AI models process code-mixed inputs and the extent to which they capture sociolinguistic motivations behind language alternation.*

Key words: *Uzbek-English code-switching; code-mixing; bilingualism; Artificial Intelligence (AI); Natural Language Processing (NLP); morphological adaptation; sociolinguistics; annotated corpora; machine translation.*

The phenomenon of code-switching and code-mixing between Uzbek and English has intensified in recent years, driven by globalization, the expansion of digital communication, and the increasing presence of English in educational, professional, and social domains in Uzbekistan. As bilingual practices become more widespread—especially among students, young professionals, and active social media users—there is a growing need to systematically study how and why speakers alternate between the two languages. Traditionally, research on Uzbek-English bilingualism has been primarily qualitative, focusing on linguistic patterns and sociocultural motivations. However, the rapid advancement of artificial intelligence (AI) and Natural Language Processing (NLP) technologies has introduced new possibilities for large-scale, data-driven investigation of bilingual speech and text.

AI-based approaches offer significant advantages in analyzing code-switching: they enable the automated identification of language boundaries, the detection of morphologically integrated loanwords, and the interpretation of pragmatic markers across languages. Nevertheless, the Uzbek-English bilingual context presents unique challenges due to the typological differences between the languages and the dynamic nature of their interaction in real-life communication. The existing lack of large, annotated corpora further complicates computational analysis, requiring researchers to develop innovative data collection strategies and tailored models. Despite these challenges, ongoing work in this field demonstrates that AI-driven methodologies can substantially deepen our understanding of both linguistic structure and social meaning in bilingual communication.

Automatic AI-based analysis of code-switching and code-mixing between Uzbek and English has recently become a significant area of inquiry, particularly as digital

communication expands across educational and social platforms in Uzbekistan. As bilingual practices intensify among younger speakers, scholars increasingly turn to artificial intelligence (AI) and Natural Language Processing (NLP) to systematically analyze the linguistic, sociolinguistic, and computational dimensions of these phenomena. Code-mixing in Uzbek-English contexts is especially rich because of the typological differences between the two languages—agglutinative Uzbek and analytic English—which create unique linguistic patterns that AI systems must learn to process accurately.

One of the most frequently observed patterns is intra-sentential code-mixing, where an English lexical item is inserted into an Uzbek grammatical structure. For instance, in the sentence "Men assignmentni kechqurun topshirdim," the English noun assignment receives the Uzbek accusative case marker -ni, demonstrating productive morphological integration. This type of insertion is challenging for AI systems because it requires the simultaneous recognition of the English lexical root and the Uzbek inflectional morphology. Similarly, hybrid sentences like "I need to study tonight, chunki test ertaga bo'ladi" reflect fluid bilingual competence, where speakers select the language that best expresses a particular semantic or pragmatic function. AI models analyzing such structures must consider both syntactic boundaries and pragmatic motivations to correctly segment and classify tokens.

Inter-sentential switching, by contrast, poses fewer structural difficulties for computational systems. In examples such as "Bugun juda charchadim. I think I need a break," the boundary between Uzbek and English occurs between independent clauses, allowing AI to process each segment with distinct language parameters. However, even in these cases, context-sensitive interpretation remains essential, particularly when assessing discourse cohesion, speaker intention, or stylistic motivations. Another common pattern is tag-switching, illustrated by expressions such as "Bu juda qiyin, you know?" where the English tag functions as a discourse marker expressing shared understanding or seeking affirmation. For AI systems, accurate identification of such tags requires pragmatic sensitivity and an understanding of how discourse markers operate across languages.

A major barrier to advanced AI research on Uzbek-English code-switching is the lack of large, annotated corpora. Existing studies often rely on manually collected datasets from Telegram channels, university classrooms, social media posts, or semi-structured interviews. The scarcity of high-quality training data limits the performance of machine learning models, especially for tasks such as automatic segmentation, morphological analysis, and code-switch prediction. Nonetheless, ongoing efforts to build bilingual and code-mixed corpora show promise, particularly as researchers adopt semi-automatic annotation tools and incorporate user-generated digital content.

AI-based methodologies have also been applied to machine translation (MT). While multilingual models such as MBART, NLLB, and various transformer-based LLMs demonstrate strong performance on monolingual data, they often struggle with code-mixed input. Their errors typically involve incorrect morphological handling, inconsistent syntactic mapping, or the over-normalization of English borrowings into Uzbek. These shortcomings reveal the need for specialized, code-switch-aware models that can recognize

blended morphosyntactic structures and preserve the nuanced distribution of languages within a sentence.

To enhance processing accuracy, researchers train NLP models to detect a range of linguistic features. These include lexical motivations for switching—often triggered by gaps in Uzbek terminology in fields such as information technology, business, or science—where English technical terms are more accessible to speakers.

AI systems also learn to identify morphological adaptation, a hallmark of Uzbek-English mixing, in forms like meetingga or assignmentni, where English roots attach to Uzbek suffixes to maintain grammatical coherence.

Additionally, contextual embeddings help models determine the likely language of a token based on surrounding grammatical cues. Beyond linguistic factors, AI tools increasingly perform sociolinguistic analyses, quantifying how bilingual practices index identity, prestige, modernity, and group membership among youth. These computational findings often align with sociolinguistic theory, which views English as a symbol of educational aspiration, global connectivity, and technological competency in contemporary Uzbekistan.

In summary, the AI-based study of Uzbek-English code-switching and code-mixing constitutes a rapidly evolving research domain that bridges linguistics, computer science, and sociocultural analysis. The structural complexity arising from the combination of an agglutinative and an analytic language presents unique challenges for NLP models, particularly in identifying intra-sentential mixing and morphologically adapted English borrowings.

Despite the current limitations posed by insufficient annotated corpora, emerging computational techniques—combined with expanding digital communication environments—provide increasing opportunities for robust and scalable analysis.

Ultimately, advancements in AI methodologies will not only improve the automatic processing of bilingual data but also deepen scholarly understanding of how language practices reflect social identities, educational contexts, and cultural transformations in Uzbekistan.

REFERENCES:

1. Myers-Scotton, C. (1993). Social Motivations for Codeswitching: Evidence from Africa. Oxford University Press.
2. Fishman, J. A. (1972). Sociolinguistics: A brief introduction. Rowley: Newbury House.
3. Yusupova, N. (2025). Sotsiollingvistika va zamonaviy lingvistikning kesishgan nuqtasi: ko‘p tilli O‘zbekiston va kod-almashinuv holatlari. JOURNAL OF “MAMUN SCIENCE”. Volume3, Issue 3.
4. Abdulvokhidova, N. (2025). O‘zbek va ingliz tillari uchun kod-almashtirish (code-switching)ni aniqlovchi model. Educational Research in Universal Sciences, 4(7)
5. Egamberganova, G. A. (2025). Analyzing code-switching in Uzbek-English classroom interactions. Ilmiy Tadqiqotlar va Ularning Yechimlari. Vol. 6 No. 02

6. Shermirzayeva, M. (2025). Code-switching among bilingual English learners in Uzbekistan. International Journal of Science-Innovative Research. Vol. 1 No. 2

7. Jabborov, M. M. (2025). Comparative study of code-switching in bilingual Uzbek-English speakers. Sustainability of Education, Socio-Economic Science Theory. Vol. 3 No. 30.