# Tweets and Tags: A Study of Code-Switching in Online Contexts.

Aahad Vakani

DePauw University

Abdul Aahad Vakani

HONR 300

Professor Amity Reading

April 28th, 2025

Code-switching is broadly defined as “the use of several languages or dialects within the same conversation or utterance” (Gardner-Chloros, 2009, p. 4), and it is very prevalent in most bilingual and international communities. Online forums and communities offer a very interesting perspective into how code-switching works, in addition to where and when people code-switch the most. “Code-switching is not merely linguistic behavior but also an expression of social identity and cultural belonging” (Bullock & Toribio, 2012, p. 9), highlighting how code-switching is more complicated than just changing how one speaks to fit how others around them are speaking. This paper argues that code-switching is a key strategy for bilingual speakers to perform and negotiate their identities online, with variation between public and private platforms.

There are many different theories about the existence of code-switching and its importance and relevance in bilingual communities and, specifically, in their linguistic and social domains. Peter Auer says that “Code-switching serves communicative purposes, marking social and interactional boundaries” (Auer, 1998, p. 15). Similarly, Myers-Scotton (2006) points out that “speakers switch codes to index relationships, identities, and situations” (p. 75). This implies that bilingual speakers code-switch not just for clarity and aiding comprehension, but also to aid social signals and to manage their identities. Digital discourse is an interesting dimension to examine code-switching because of the added layers of adaptation and innovation. Sebba, Mahootian, and Jonsson (2012) point out that "online environments enable innovative forms of written code-switching that mirror spoken patterns but also show distinct digital traits" (p. 37). This intersectionality of conversation underscores the characteristics of code-switching that is inherent to language use in online forms.

Advancements in understanding code-switching through translation tools and machine learning have been useful, but despite those, there are still quite a few challenges, particularly in multilingual data within computational frameworks. Jose et al. (2020) observed that "traditional NLP models struggle with code-switching due to its multilingual and non-standard nature, highlighting a significant research challenge" (p. 3). In an attempt to address this issue, I utilized the L3Cube-HingCorpus dataset, described by Nayak and Joshi (2022) as containing "over 52.9 million sentences and 1.04 billion tokens collected primarily from Twitter" (p. 1). I chose this dataset due to its scale, specificity, and its detailed linguistic attributes in addition to the presence of naturally occurring bilingual discourse in Hindi-English. Every word in the dataset is annotated to describe whether it's an English word or if it's a Hindi word. Other key variables analyzed in this paper include the frequency of switches per 100 words, common placements where the switching happened (beginning, middle, or end), and types of switches (intra-sentential, inter-sentential, and tag level). The sheer size of the dataset helps detect patterns at a macro-level in bilingual textual conversations, however, its sole focus on Twitter as the platform it employs to gather data could present a bias in the results.

Following preprocessing of the HingCorpus data, results indicated substantial patterns in bilingual communication on Twitter. The average switches per 100 words was approximately 31, almost 1 switch every 3 words, indicating high language fluidity. Most switches were mid-sentence (35,690 instances), followed by beginning (11,031 instances) and at the end

(10,329 instances). The most common switch type was tag-level (6,420 switches), most of these being interjections or emotionally charged phrases. All of these metrics were summarized using word counts, switch placement, and switch type classification to ensure methodological transparency.

The observed prevalence of mid-sentence or intra-sentential switching supports theories of online communication being more informal, where code-switching in these places serves a more expressive purpose, and helps translate social identities. Existing research trends like those outlined by Jose et al (2020) are reinforced by these findings, demonstrating that "online code-switching exhibits frequent mid-sentence transitions and linguistic blending, characteristics increasingly prevalent in informal digital communication" (p. 7). This has a parallel to in-person code-switching in bilingual communities, where people tend to switch mid-sentence to re-emphasize their identity to people from similar backgrounds as them. Code-switching is thus deliberate, patterned, and holds social value. The dominant mid-sentence switching and tag-level switches indicate that bilingual code-switchers consciously use it as a strategy to negotiate their identities online. This aligns with Auer's (1998) theories of code-switching being an interactional boundary marker and Bullock and Toribio's (2010) arguments regarding the role that code-switching plays in social identity performance.

While being comprehensive, this study is not without its limitations; the sole focus on Hindi-English code-switching restricts its broader applicability. Additionally, as stated before, the sole focus on Twitter may neglect diverse contexts and demographic variances. The dataset also lacks detailed metadata regarding users' intents or emotional tones. Despite these limitations, the findings from this study contribute to the growing body of research that positions code-switching as an intentional, identity-driven linguistic practice rather than a random or purely utilitarian phenomenon. In online environments, particularly informal ones like Twitter, users engage in strategic language mixing to express emotions, claim cultural affiliation, and establish solidarity with specific audiences. These communicative acts underscore that digital code-switching mirrors many of the same social motivations found in face-to-face bilingual interaction. Twitter, being a public platform and the primary source for the HingCorpus dataset, is often a performative platform. This shapes users' language choices and the content of their discourses. As Androutsopoulos (2015) notes, public digital writing is deeply entangled with audience design and identity performance. Thus, private platforms like WhatsApp may show a different picture when it comes to code-switching observations.

On private messaging platforms, code-switching is theorized to serve a utilitarian purpose. Efficiency, intimacy, or solidarity may be signalled through a code-switch, while bilingual speakers may still default to English in academic contexts. Semi-anonymous platforms like Reddit, code-switching might serve a different purpose again, users may switch without fears of misinterpretation. Comparative studies, such as those by Danet and Herring (2007), suggest that platform affordances (e.g., character limits, visibility, threading) materially shape linguistic behavior, including the frequency and nature of code-switching.

Building on this, the prominence of tag-level switches offers further insight into the emotional and cultural signals behind code-switching. These short and embedded phrases function as emotional cues or expressions of solidarity. For example, interjections such as “bas” or “please yaar” are scattered throughout Hinglish tweets and act as markers that reassert cultural authenticity, and signal to audiences of the person they are conversing with being familiar because of their shared cultural roots. These tag-switches carry a high indexical load, signalling shared cultural reference points in addition to bilingual competence. These types of switches can bypass grammatical constraints and serve discourse functions, while also functioning as pragmatic markers that can soften commands, convey sarcasm, and strengthen emotional appeal, which may otherwise be lost in monolingual discourse. Many NLP (Natural Language Processing) tools may fail to interpret such short tokens correctly or misclassify them due to their brevity and informal usage. Building sentiment-aware, culturally sensitive parsers will be essential if we are to accurately model how these switches operate in real-world discourse.

These linguistic patterns also reveal important gaps in how current NLP systems process and interpret bilingual data. Most models are trained on Western-centric monolingual corpora and tend to underperform with code-mixed inputs. This is also validated by social media data, where language mixing follows no particular structure. It is full of emojis, abbreviations, and spelling variations. In certain examples from the HingCorpus, a single sentence might contain English root words with a Hindi prefix or a suffix. For example, “movie dekhne gaya tha”, here English and Hindi words blend into a single sentence that makes sense to bilingual speakers and follows their grammatical rules, but could be a challenge for traditional NLP models. Researchers like Solorio et al. (2014) have called for the creation of code-switching-specific NLP pipelines, including language identification at the word level, multilingual embedding models, and context-aware decoders. In the current study, the L3Cube-HingCorpus offers word-level language tags, but further preprocessing was still required to ensure consistent classification. Finally, training pipelines that understand the social context cues that come with code-switching remain the biggest challenge. Models can detect switches and some can even understand the surface-level meaning behind them, but understanding why a switch occurs is a task for hybrid models that combine NLP and sociolinguistic theories.

Given these complexities, a multidisciplinary approach is vital to understanding code-switching in online contexts. Linguistics provides the ground-level framework required to understand the functions of code-switching, while sociology and cultural studies help frame the identity and performativity of the switches. Computational advancements allow for pattern recognition, but without the theoretical backgrounds being contextualized, this analysis will remain shallow. Future research could benefit from a comparative approach that examines code-switching across multiple platforms, such as WhatsApp, Reddit, or YouTube comments, where context, anonymity, and audience design vary considerably. Additionally, including multilingual datasets beyond Hindi-English would help capture a broader picture of code-switching patterns and motivations across different linguistic and cultural groups. Integrating sentiment analysis or user profiling (while respecting privacy norms) might also offer deeper insight into the affective and social dimensions of code-switching in digital discourse.

The insights that this study provides can offer real-world applications. Educators can use this to understand how students use code-switching to generate inclusive language policies and curriculum. Institutions can use it as a resource for learning and identity exploration, both online and offline. Developers behind these LLMs can use these findings to build and train inclusive and culturally aware systems. They need to account for code-switching as an expressive linguistic mode, rather than noise that needs to be ignored. Understanding code-switching is a marker of sophistication for these AI models, making them universal, rather than being strictly Western-centric. Finally, for the bilingual speakers in question, this research affirms their practices of code-switching as not irregular but as patterned and worthy of scholarly research. Code-switching is not a sign of confusion in bilingual speakers; it is instead a sign of dexterity and cultural negotiation.

In conclusion, this study affirms that code-switching is more than a byproduct of bilingualism; it is a meaningful social act. Bilingual users online do not merely alternate between languages for convenience but often do so to negotiate identity, reinforce in-group belonging, and navigate the fluid boundaries of digital interaction. The findings underscore the need for both computational tools and linguistic theories to adapt to the evolving nature of multilingual communication in the digital age. As code-switching becomes more prevalent and more visible, especially in online settings, its role as a linguistic and cultural resource will only grow in relevance.

# References

Auer, P. (1998). *Code-switching in conversation: Language, interaction, and identity. Routledge*.

Bullock, B. E., & Toribio, A. J. (Eds.). (2012). *The Cambridge handbook of linguistic code-switching.* Cambridge University Press.

Gardner-Chloros, P. (2009). *Code-switching*. Cambridge University Press.

Jose, N., Gupta, D., & Shrivastava, M. (2020). *A survey of current datasets for code-switching research.* In 2020 IEEE International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 136-141). IEEE. https://doi.org/10.1109/ICACCS48705.2020.9074205

Myers-Scotton, C. (2006). *Multiple voices: An introduction to bilingualism*. Wiley-Blackwell.

Nayak, R., & Joshi, M. (2022). *L3Cube-HingCorpus: A corpus for Hinglish code-mixed language modeling*. L3Cube Pune Research Group. Retrieved from https://github.com/l3cube-pune/code-mixed-nlp

Sebba, M., Mahootian, S., & Jonsson, C. (Eds.). (2012). Language mixing and code-switching in writing: Approaches to mixed-language written discourse. Routledge.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Hirschberg, J., ... & Liu, Y. (2014). Overview for the First Shared Task on Language Identification in Code-Switched Data. *Proceedings of the First Workshop on Computational Approaches to Code Switching*.

Androutsopoulos, J. (2015). Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism*, 19(2), 185–205.

Danet, B., & Herring, S. C. (Eds.). (2007). *The multilingual internet: Language, culture, and communication online*. Oxford University Press.