# THE IMPACT OF AI ON SOCIOLINGUISTIC EFFORTS TO ANALYSE CODE-SWITCHING IN MULTILINGUAL CONVERSATIONS

**PRADHAN, MELISSA**
*English Language and Linguistics, College of Arts and Humanities*

**ABSTRACT**

The introduction of Artificial Intelligence (AI) in sociolinguistic analysis has helped to improve the analysis of code-switching (C-S) in multilingual conversations. This paper investigates the effectiveness of various AI programmes, such as Natural Language Processing (NLP) and Machine Learning (ML), in finding and analysing the linguistic phenomena of C-S. AI-powered models enable the automatic recognition and processing of many languages from a single utterance, improving the accuracy and efficiency of sociolinguistic research. AI in sociolinguistics allows for a more in-depth understanding of language dynamics, cultural exchanges, intricate dialects, and accent variations. Nonetheless, there are challenges when applying AI in sociolinguistic research, such as algorithm bias, the requirement for diverse linguistic data, and ethical considerations. Thus, this article looks at recent research on AI's role in sociolinguistic analysis to interpret the positive and negative effects of AI on C-S analysis. It aims to raise awareness about AI's revolutionary role in enhancing sociolinguistic efforts to understand and analyse C-S in multilingual situations, with a focus on C-S in Indian communities.

## INTRODUCTION

Code-switching (C-S) is a common linguistic phenomenon defined as the constant alteration between two or more languages within one conversation (Yilmaz et al., 2016, p.1). C-S is seen in speech and writing, where people switch between various languages 'within an utterance or between consecutive utterances' (Deng et al., 2022, p.527). In linguistics, utterances refer to continuous parts of speech occurring in-between breaths or pauses (Glossary of Linguistic Terms, 2015). As such, a C-S utterance in writing is representative of a textual transcription of C-S patterns found in speech. C-S is especially prevalent among multilingual individuals, who demonstrate complex linguistic adaptability by fluently alternating between languages. C-S analysis is a new field of exploration that allows sociolinguists to gain insight into cultural dynamics, power relations, and people's social identity within multilingual communities. In recent years, Artificial Intelligence (AI) has gradually emerged as a tool to aid C-S analysis. AI-powered technologies like Natural Language Processing (NLP), Machine Learning (ML), Deep Neural Networks (DNN), and Automatic Speech Recognition (ASR) help to process and analyse large multilingual datasets through increased abilities in classification, logical reasoning, information extraction, finding C-S patterns, etc. Hence, they enable sociolinguists to understand the complexities of language use at previously inaccessible levels.

However, while AI offers essential advancement in understanding language dynamics, dialect variations, and cultural exchanges through its classifications of C-S data, it has its limitations in sociolinguistic studies. Use of AI for C-S in sociolinguistics is a relatively novel approach, resulting in many concerns such as algorithmic biases, the need for diverse linguistic data, inaccuracy in results, and ethical considerations. As such, this study explores how AI contributes to the field of sociolinguistics by enhancing C-S analysis and multilingual communications, while addressing its inherent limitations, complexities, and the expected responsibilities, such as the ethics of data collection for C-S research.

## ADVANTAGES AND LIMITATIONS OF AI IN C-S ANALYSIS

With AI's integration into sociolinguistics, there is better facilitation of data collection and processing, enabling faster analysis of extensive datasets. Use of AI models offers significant advantages in terms of speed, efficiency, and scalability. Through automated data processing, AI allows for the rapid analysis of vast multilingual datasets, which would have otherwise been time-consuming and resource intensive. Previously, sociolinguists manually collected data, limiting them to much smaller datasets and other logistical constraints (Gallegos et al., 2024). However, with the introduction of AI tools like NLP, and Multilingual DNNs, linguists can analyse large volumes of data from real-world multilingual interactions found both offline and online, because they can train AI models to accumulate related and corresponding C-S labelled data (Deng et al., 2022, p.528). For example, such AI models access data found on social media platforms like X (formerly Twitter) and Facebook to analyse the frequency and trends of C-S found in bilingual tweets or captions, both in formal and informal settings (Das and Gambäck,

2013). AI models are also used in analysing bilingual spoken discourse in linguistics experiments, where researchers record spoken utterances of C-S and use the AI models, like NLP, to closely analyse the trends found in their speech and replicate those results to create greater datasets for further research (Gallegos et al., 2024). The extensive dataset accessibility hence widens the scope for C-S analysis, allowing sociolinguists to study C-S in greater depth than before. Additionally, the availability of remote diverse datasets allows sociolinguists to capture a more comprehensive pattern of C-S across various multilingual communities and languages, ultimately deepening their understanding of multilingual communication. For instance, with the help of AI models, Agarwal et al. (2017) found that in Hindi-English C-S, Hindi is generally preferred to express negative emotions, such as using Hindi swear words (Doğruöz et al., 2021, p.1660). Thus, data processing done by computational programs helps researchers uncover patterns in C-S practices, reflecting social, cultural, and contextual nuances.

However, the accuracy of AI tools for analysis can largely differ when relying on isolated and selective data. This selective data usage often overlooks important elements, such as by excluding essential information about cultural and social subtleties in multilingual communities (Doğruöz et al., 2021, p.1658). In many multilingual societies, some languages are treated as being superior to others. For example, in India, Indo-Aryan languages like Konkani are considered more prestigious than Dravidian languages like Kannada (Doğruöz et al., 2021, p.1658). Nadkarni (1975) shows that despite being fluent in both Kannada and Konkani, Saraswat Brahmins avoid C-S between the two languages. Thus, C-S between prestigious and less prestigious languages is historically unseen because of the social bias against inferior or less prestigious languages in multilingual societies (Doğruöz et al., 2021, p.1658). Unfortunately, AI models are unable to accurately understand the social and historical influence in multilingual communities, especially when dealing with 'prestige' in languages (Doğruöz et al., 2021, p.1658). AI models also struggle with sentence structures that do not conform to preset or straightforward linguistic boundaries, such as sentences with short forms, word play, and Romanised spelling of words in different languages. This challenges the processing of more complex C-S utterances in real-life because these do not adhere to the preset boundaries of different languages inputted during training of AI models (Doğruöz et al., 2021, p.1655).

Moreover, the current inability of AI models to perceive social and historical influences also leads it to accumulate data that is independent of its socio-cultural factors, leading to inaccuracies in analysing social patterns in C-S sentences. This limitation arises from the underdeveloped state of AI models that excludes many languages from multilingual analysis. Despite C-S being a focal point in sociolinguistics, the field often prioritises well-documented or widely spoken languages due to the easily available data (Doğruöz et al., 2021, p.1661). As a result, the scarcity of comprehensive linguistic datasets for less commonly spoken or less prestigious languages hinders the advancements in AI-driven C-S analysis. Thus, the limited academic interest associated with these languages causes them to be frequently overlooked, resulting in minimal progress in creating AI models capable of accurately analysing C-S in such contexts. Moreover, while AI models can be trained to interpret C-S sentences, the training data rarely mirrors the real-life diversity of multilingual interactions in society (Deng et al., 2022, p.528). These AI tools, like ASR and DNNs, are primarily developed for monolingual sentence analysis and are adapted to read C-S data (Yilmaz et al., 2016, p.5). Thus, they are still prototype models and can fail to recognise every type of C-S that may occur in an utterance (Yilmaz et al., 2016). As a result, since AI frequently misses the subtle linguistic cues characterising the complexity of C-S, the analysis of C-S data remains far less precise than monolingual data (Doğruöz et al., 2021, pp.1659-1660; Deng et al., 2022, p.530). Therefore, this emphasises the limitations of current AI tools in completely capturing the sociocultural richness that is an integral part of C-S practices.

Nonetheless, AI models have proven to be beneficial in advancing Automatic Language Identification (LID) and segmentation, especially in C-S analysis. LID is the process of using computational models to automatically determine the language of any provided written or spoken source (Qafmolla, 2017, p.140). In linguistics, segmentation complements LID by analysing language as a composition of basic and distinct contrasting elements made of meaningful segments and simple sound units that form words (Kluender, n.d.). Through the utilisation of complex trained algorithms based on deep-learning, such as ASR and NLP algorithms, sociolinguists can detect the precise points of transition between languages within an utterance or text (Shah and Sitaram, 2019; Yilmaz et al., 2016). As such, AI analyses C-S contexts and switching points automatically, reducing the requirement for manual segmentation by researchers, giving insights into linguistic choices and social contexts. Moreover, deep-learning AI models like NLP can also detect the patterns across different categories of C-S sentences, such as intra-sentencing and inter-sentencing switches, helping researchers to explore when and why multilingual speakers may prefer certain C-S styles (Yilmaz et al., 2016). 'Intra-sentencing' involves one primary language with culturally significant words from another, while 'inter-sentencing' balances multiple languages within a sentence (Deng et al., 2022, p.1655). Hence, AI gives sociolinguists the enhanced ability to process and interpret large multilingual datasets, allowing them to deepen their understanding and categorisation of different C-S patterns and practices across diverse cultural, social, and linguistic settings.

## EVALUATING AI'S CONTRIBUTION TO LINGUISTIC DATA COLLECTION AND ANALYSIS IN C-S RESEARCH

The integration of AI in C-S analysis offers a balanced approach to both quantitative and qualitative insights within sociolinguistics. Through the automation of large-scale data analysis, AI enables researchers to identify trends in C-S patterns across different geographical and social contexts. For instance, ML models help sociolinguists examine the difference between C-S within immigrant communities and speakers in their home countries, such as C-S within Non-Resident Indian (NRI) communities and local Indian communities. Often researchers use quantitative data, such as the frequency of C-S switches within an utterance, to interpret cultural and social influences for different types of C-S used amongst largely multilingual communities (Das and Gambäck, 2013, p.45). Quantitative data collected and interpreted through AI can help sociolinguists create speech patterns to statistically generate C-S sentences for future studies. For example, a new AI model called Translation for Code-Switching (TCS) is trained with quantitative data of Hindi-English C-S sentences and monolingual Hindi sentences (Tarunesh et al., 2021). Once fully trained, TCS can generate realistic Hindi-English C-S sentences using monolingual Hindi sentences, which allows for further studies in understanding C-S within Hindi-English bilingual communities (Tarunesh et al., 2021, p.3154). As previously stated, ASR and NLP models also allow sociolinguists to analyse subtle shifts in language use, like contexts where NRIs and local Indian speakers employ intra-sentence or inter-sentence switching. For example, AI-driven analysis demonstrated that while NRIs usually use inter-sentencing C-S, local Indian speakers tend to use intra-sentencing C-S (Dey and Fung, 2014; Das and Gambäck, 2013). Additionally, sociolinguists find similarities through AI, such as both communities tending to use it more in an informal setting than a formal scenario (Das and Gambäck, 2013). AI models can pinpoint specific social situations or phrases that most commonly prompt C-S, enriching sociolinguistic understanding of language choices which reflect social identity and cultural adaptation, resulting in a more qualitative analysis of the given data. Thus, AI models support both statistical patterns and the more intricate cultural insights essential in sociolinguistic studies, often providing insights into the adaptive strategies of diaspora communities.

Yet, despite these advancements, the real-world application of AI is limited in accuracy, especially in diverse linguistic contexts. AI models can effectively analyse popular language pairs, such as English-Bengali, English-Chinese, English-Hindi, etc., due to the extensive bilingual data available. For less common language pairs, like Frisian-Dutch or Frisian-English, there is limited accurate data available, thus hindering AI analysis (Yilmaz et al., 2016, p.4). This scarcity results in AI models being unable to accurately recognise precise C-S patterns, triggers, and the subtleties of language choice when dealing with rare language pairs. Moreover, despite DNNs and LID models being able to recognise C-S in controlled settings, such as when optimal and sufficient training data is provided, these models are unable to adapt and process the situational anomalies and the diversified instances of C-S that occur spontaneously in real-world conversations (Doğruöz et al., 2021, p.1660). As Deng et al. (2022) note, these models 'need to be able to adapt to users' linguistic styles, while also being capable of determining when and how to code-switch'. In many multilingual communities, speakers often code-switch based on context, social cues, and the audience they interact with. Often with shifting social cues, no person uses C-S the same way as another, resulting in many of the AI-generated patterns, which use training data, becoming inaccurate and sometimes void in the real-world. For example, as previously mentioned, AI models flagged that NRIs tend to use inter-sentencing C-S relatively more than other C-S forms. However, these models fail to note more socially complex C-S patterns related to back-flagging i.e. when a speaker reverts to a prior used language or dialect (Deng et al., 2022). The study by Dey and Fung (2014) shows that many NRIs are inclined to use 'backflagging' more frequently, especially in informal settings with other NRIs who share similar backgrounds or fluency in the switched languages (Deng et al., 2022). As such, these complex dynamics pose a great challenge to AI's predictive accuracy, emphasising its limited generalisability to all real-life sociolinguistics contexts. Therefore, while AI has the potential to be a powerful tool for analysing mainstream bilingualism, its application to more culturally nuanced, localised, or underrepresented languages continues to be a key hinderance for its success in sociolinguistic studies.

## ETHICAL CONSIDERATIONS FOR USE OF AI IN SOCIOLINGUISTIC RESEARCH

Finally, the most significant limitation for current AI models in C-S analysis is their subjection to possible ethical questioning, particularly regarding data usage, bias, and representation. As mentioned earlier, AI models often rely on publicly available data from social media platforms for their analysis of patterns and autorecognition of C-S (Das and Gambäck, 2013). However, this results in concerns over privacy, consent, and data ownership. Individuals may be unaware that their posts or tweets are being analysed, potentially violating principles that are outlined in the General Data Protection Regulation (GDPR). The GDPR highlights the need for transparency and confidentiality in the usage of personal information, raising ethical and legal questions concerning data mining from social media platforms (Data Protection Commission, 2024). Additionally, the automation of C-S analysis leads to questions about the replacement of qualitative insights traditionally derived from human-led studies, which puts studies at risk of providing culturally biased or stereotypical results. A notable example is the earlier cited Konkani-Kannada language pair, where the understanding of 'prestige' languages was achieved primarily through human-led studies (Doğruöz et al., 2021, p.1658). In contrast, AI models tend to generalise such pairs under more popularly analysed Indian C-S patterns, such as Hindi-Bengali or Hindi-English, failing to capture their unique dynamics (Doğruöz et al., 2021). Hence, the potential for AI results to reinforce stereotypes stresses the need for transparent, inclusive, and ethical practices in AI-driven sociolinguistic research, while preserving the ethical collection of data.

## CONCLUSION: THE FUTURE OF AI IN C-S RESEARCH AND SOCIOLINGUISTICS

In conclusion, the application of AI for analysing C-S is an emerging field within sociolinguistic research. Despite its potential, significant steps are still necessary to refine AI models for more accurate and ethically responsible usage. Currently, the application of AI in C-S analysis, while promising, brings challenges in terms of model bias, sensitivity to linguistic variation, and limitations in processing accuracy of lesser-known language pairings. These limitations may lead to inaccurate conclusions or unintentional marginalisation of underrepresented language communities. Moving forward, while AI has the potential to transform C-S analysis in sociolinguistics, future work must address issues of bias, accuracy, and inclusivity. There is a need for greater investment in diverse datasets and malleable AI models to better capture the richness of language use in the real world. As AI tools evolve, interdisciplinary collaboration between sociolinguists, computational linguists, and AI specialists will be essential to develop ethical and contextually accurate models (Doğruöz et al., 2021, p.1655). This evolution promises more comprehensive insights into C-S and offers a pathway for AI to contribute more broadly to inclusive and equitable sociolinguistic research.

## REFERENCES


Das, A. and Gambäck, B. 2013. Code-mixing in social media text. *Traitement Automatique des Langues*. **54**(3), pp.41-64.

Deng, S., Li, C., Bai, J., Zhang, Q., Zhang, W.-Q., Yang, R., Cheng, G., Zhang, P. and Yan, Y. 2022. Summary on the ISCSLP 2022 Chinese-English code-switching ASR challenge. *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. pp.527–531. [Online]. [Accessed 10 November 2024]. Available from: https://arxiv.org/pdf/2210.06091.

Dey, A. and Fung, P. 2014. A Hindi-English code-switching corpus. *European Language Resources Association (ELRA)*. pp.2410-2413.

Doğruöz, A. S., Sitaram, S., Bullock, B. E. and Toribio, A. J. 2021. A survey of code-switching: linguistic and social perspectives for language technologies. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. **1**, pp.1654–1666.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R. and Ahmed, N. K. 2024. Bias and fairness in large language models: a survey. *Computational Linguistics*. **50** (3), pp.1097–1179.

Glossary of Linguistic Terms. 2015. *Utterance*. [Online]. [Accessed 9 December 2024]. Available from: https://glossary.sil.org/term/utterance.

Data Protection Commission. 2019. *Guidance on the principles of data protection_Oct19*. [Online]. [Accessed 9 December 2024]. Available from: https://www.dataprotection.ie/sites/default/files/uploads/2019-11/Guidance%20on%20the%20Principles%20of%20Data%20Protection_Oct19.pdf.

Kluender, R. [no date]. Center for Academic Research and Training in Anthropogeny (CARTA). *Language segmentation*. [Online]. [Accessed 10 November 2024]. Available from: https://carta.anthropogeny.org/moca/topics/language-segmentation.

Nadkarni, M. V. 1975. Bilingualism and syntactic change in Konkani. *Language*. **51**(3), pp.672–683.

Shah, S. and Sitaram, S. 2019. Using monolingual speech recognition for spoken term detection in code-switched Hindi-English speech. *2019 International Conference on Data Mining Workshops (ICDMW)*. pp.1126-1130.

Qafmolla, M. A. N. 2017. Automatic language identification. *European Journal of Language and Literature Studies*. **7**(1), pp.140-150.

Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A. and Fung, P. 2021. Are multilingual models effective in code-switching? *Center for Artificial Intelligence Research (CAiRE)*.

Tarunesh, I., Kumar, S. and Jyothi, P. 2021. From machine translation to code-switching: generating high-quality code-switched text. In: Zong, C., Xia, F., Li, W. and Navigli, R. eds. *Proceedings of the 59th Annual Meeting of the*

*Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. **1**, pp.3154-3169.

Treffers-Daller, J. 2022. Code-switching among bilingual and trilingual Children. In: Stavans, A. and Jessner-Schmid, U. ed., *The Cambridge handbook of childhood multilingualism*. 1st ed. Cambridge University Press, pp.190–214. [Online]. [Accessed 16 November 2024]. Available from: https://www.cambridge.org/core/product/identifier/9781108669771%23CN-bp-8/type/book_part.

Yilmaz, E., Van Den Heuvel, H. and Van Leeuwen, D. 2016. Code-switching detection using multilingual DNNS. *2016 IEEE Spoken Language Technology Workshop (SLT)*. pp.610-616.