Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

AI-Driven Analysis of Code-Switching from Local European Languages to English in Media and Literary Discourse

Tutova E. V., Ebzeeva Y. N., Smirnova Y. B., Gishkaeva L. N.

Acknowledgements: This publication has been supported by the RUDN University named after Patrice Lumumba Scientific Projects Grant System, project no. 050738-0-000

ARTICLE INFO

Received: 25 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

English has become a dominant global language, reshaping linguistic practices throughout Europe. In media and literature, code-switching—the alternation between local languages and English—has grown increasingly common. This study employs advanced AI-driven Natural Language Processing techniques to analyze the patterns, motivations, and sociolinguistic implications of this phenomenon.

Our research draws on diverse sources, including newspapers, social media posts, television scripts, films, and digital literature from several European regions. By using state-of-the-art machine learning models, we automatically detect instances of code-switching and examine the contexts in which these shifts occur. This approach allows us to classify language blending based on cultural, social, and genre-specific factors. Early results indicate that code-switching is often employed to convey modernity, express technical ideas, or create a particular social tone, with variations observed across different media and regions. Furthermore, our study integrates quantitative analysis with qualitative insights by reviewing selected case studies in depth. This mixed-methods strategy enriches our understanding of the relationship between language choice and cultural identity. The findings shed light on how the increasing use of English influences local linguistic landscapes and may signal broader trends in language evolution. Our research not only highlights the transformative role of AI in sociolinguistic studies but also offers practical implications for educators, media professionals, and policymakers. By clarifying the dynamics of code-switching, we hope to contribute to efforts aimed at preserving linguistic diversity while embracing the benefits of global communication. In sum, this work demonstrates that AI is a powerful tool for uncovering complex language phenomena, offering fresh perspectives on how European languages adapt in a rapidly globalizing media environment. This comprehensive AI-driven analysis not only deepens our understanding of multilingual dynamics but also informs future strategies for language education and media production, ensuring that cultural nuance is maintained amid global linguistic shifts in diverse European contexts.

Keywords: multilingual, production, linguistic, policymakers

INTRODUCTION

The advent of globalization has dramatically reshaped the linguistic landscape of Europe, where the influence of English as a global lingua franca has become increasingly pervasive. In media and literary discourse, this influence is manifested through the phenomenon of code-switching—the practice of alternating between local European languages and English within a single communicative context. This introduction outlines the context, significance, and innovative approach of our study, which leverages AI-driven Natural Language Processing (NLP) techniques to analyze and interpret the patterns, motivations, and sociolinguistic implications behind this linguistic shift (Rogers et al.).

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Historically, code-switching has been studied as a naturally occurring linguistic practice among bilingual and multilingual speakers (Poplack et al.). Traditionally, researchers have focused on its pragmatic functions, such as signaling group identity, negotiating social relationships, or emphasizing particular points during communication. In Europe, however, the dynamics of code-switching are evolving. The rapid expansion of digital media, coupled with the omnipresence of English in academic, commercial, and pop cultural domains, has intensified the frequency and complexity of these language shifts. The interplay between local languages and English not only reflects changing communication practices but also raises important questions about cultural identity, linguistic preservation, and the future of European languages.

Media and literature serve as particularly rich contexts for observing these linguistic transformations. From newspapers and television scripts to online blogs and digital literature, diverse genres offer a multifaceted view of how code-switching operates within public discourse. In these spaces, language is not merely a tool for communication but also a marker of social status, modernity, and cultural capital. For example, a novel might seamlessly interweave English technical terms into a narrative originally penned in Italian or Spanish, subtly conveying a cosmopolitan identity. Similarly, in digital journalism, the strategic insertion of English phrases can serve to emphasize immediacy and global relevance. Despite the ubiquity of this phenomenon, traditional qualitative methodologies face challenges in capturing its full scope and nuance, particularly given the vast and varied data produced by today's digital media.

Recent advances in artificial intelligence, particularly in NLP, have opened new avenues for linguistic research. AI-driven models can process large datasets with high accuracy, enabling researchers to detect patterns that might be imperceptible through manual analysis. By employing machine learning algorithms—such as Transformer-based models that excel in multilingual understanding—this study aims to systematically detect and categorize instances of code-switching across a broad corpus of media and literary texts. This approach not only enhances our ability to quantify code-switching occurrences but also allows for a deeper exploration of the contextual factors that trigger these language shifts.

Our research addresses several key questions: How can AI effectively identify and classify code-switching instances from local European languages to English? What linguistic and contextual cues does the AI model use to distinguish between casual borrowing and deliberate shifts in language? And, importantly, what do these patterns reveal about the broader sociolinguistic landscape of Europe? By answering these questions, our study seeks to contribute to a more nuanced understanding of how global linguistic trends influence local language practices.

Moreover, the integration of AI in sociolinguistic research carries significant practical implications. For educators, media professionals, and policymakers, insights gained from this study can inform strategies for preserving linguistic diversity while embracing the benefits of global communication. AI-driven analysis not only augments traditional sociolinguistic research but also provides a scalable, reproducible methodology that can be applied to other multilingual settings worldwide (Briva-Iglesias).

In reviewing prior research, we note that many studies have relied on limited corpora or manual coding, which, while insightful, are inherently constrained by the sheer volume and complexity of contemporary digital communication. By contrast, our methodology capitalizes on recent computational advances to process large-scale datasets, thus offering a more comprehensive view of code-switching phenomena. This innovative approach bridges the gap between traditional qualitative analyses and modern quantitative techniques, providing a robust framework for examining the linguistic interplay between local European languages and English.

In summary, this study presents an AI-driven investigation into code-switching in European media and literary discourse—a phenomenon that encapsulates the dynamic interaction between global and local cultures. By leveraging cutting-edge NLP tools, we aim to uncover the subtle nuances of language use, offering fresh insights into the ways in which European linguistic identities are evolving in an era of

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

globalization. Through our research, we aspire not only to advance academic understanding of multilingual communication but also to support efforts that balance the preservation of linguistic heritage with the inevitable march toward global integration.

METHODS AND MATERIALS

This study employs a mixed-methods approach that combines large-scale data processing with in-depth linguistic analysis to explore code-switching from local European languages to English in media and literary discourse. The following sections detail the data sources, preprocessing procedures, AI models, and evaluation metrics used in our analysis.

Our research is based on a comprehensive corpus gathered from multiple media and literary sources. We collected texts from major European newspapers, digital literature repositories, television and film transcripts, and social media platforms such as Twitter and Facebook. These sources were chosen to capture a broad spectrum of language use across both formal and informal contexts. The corpus spans several European languages, including French, German, Spanish, Italian, and Dutch, alongside the pervasive use of English. To ensure representativeness, data were sourced from a range of publication dates and geographic regions. Additionally, a subset of the corpus was manually annotated by expert linguists to create a reliable ground truth dataset for training and validation.

Preprocessing of the raw text data involved several steps. First, we applied language identification algorithms to accurately label the language of each text segment. We then performed tokenization, normalization, and removal of extraneous symbols using tools such as spaCy and NLTK. Manual annotation guidelines were developed to identify and mark instances of code-switching. Annotators noted not only the occurrence of English insertions but also contextual cues such as surrounding discourse markers and genre-specific language features.

To automatically detect and classify code-switching events, we fine-tuned Transformer-based models, including multilingual BERT and XLM-R. These models, pre-trained on extensive multilingual corpora, were selected for their proficiency in understanding and processing complex language patterns. Fine-tuning was performed on our annotated dataset using Python and the Hugging Face Transformers library. We set up sequence labeling tasks to pinpoint code-switch boundaries and employed unsupervised clustering techniques to group similar switching patterns based on context.

Our computational framework was implemented in Python, running on a high-performance computing cluster equipped with GPUs to manage the large-scale dataset efficiently. We evaluated the performance of our models using precision, recall, and F1-score metrics, applying cross-validation to ensure robustness. In addition to quantitative metrics, we conducted qualitative case studies to examine the socio-cultural contexts behind code-switching instances. Finally, to promote transparency and reproducibility, all code and processed datasets will be shared in an open-source repository, subject to any necessary data sharing agreements.

This integrative methodological framework thus combines advanced NLP techniques with traditional linguistic analysis, providing a comprehensive approach to understanding the dynamics of code-switching in European media and literary discourse.

DISCUSSION

This study set out to explore the complex phenomenon of code-switching from local European languages to English within media and literary discourse by leveraging advanced AI-driven Natural Language Processing techniques. The findings offer nuanced insights into the interplay between global linguistic trends and local cultural identities, revealing not only the frequency and context of English insertions but also their sociolinguistic underpinnings.

One of the most striking observations is that the use of English in European media and literature is not random; rather, it appears as a deliberate and context-sensitive tool. In digital journalism and social

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

media platforms, English is frequently employed to inject a sense of immediacy and modernity. Our AI models detected that in news articles and online commentaries, code-switching often occurs when discussing technology, business, or international affairs. This pattern suggests that English is perceived as a language of progress and global connectivity. In contrast, literary texts displayed a more varied pattern: while some authors seamlessly interwove English phrases to evoke cosmopolitanism or highlight technical concepts, others used code-switching as a stylistic device to enrich character dialogue or evoke specific cultural references. These differences underscore the idea that code-switching functions on multiple levels—communicative, stylistic, and symbolic.

The Transformer-based models, such as multilingual BERT and XLM-R, proved to be effective in identifying and categorizing code-switching events. The high precision and recall metrics obtained during evaluation confirm that AI can be a robust tool for large-scale linguistic analysis. Notably, the models were able to differentiate between casual borrowing—where English terms are integrated seamlessly into local syntactic structures—and more deliberate language shifts that signal a change in tone or context. The unsupervised clustering techniques further revealed that distinct clusters of code-switching could be associated with different genres and contexts, such as informal social media posts versus formal newspaper articles. These clusters provide a roadmap for future research, where each category could be examined in more detail to understand the underlying social motivations.

Despite these successes, our study also encountered several challenges and limitations. One primary challenge was the inherent complexity of multilingual texts, particularly when authors blend languages at various structural levels. Although our annotation guidelines and preprocessing methods were carefully designed, subtle nuances in meaning and connotation sometimes proved difficult to capture. For instance, idiomatic expressions and culturally loaded terms may not have been fully recognized by the models, leading to occasional misclassifications. This limitation suggests that while AI tools offer powerful methods for processing vast amounts of data, they must be complemented by traditional linguistic expertise to ensure that context and cultural subtleties are not overlooked.

Another limitation relates to the representativeness of the corpus. Although we assembled a diverse dataset spanning various media types and European languages, some sources remain underrepresented. For example, while mainstream media and widely circulated digital literature were extensively covered, niche publications and regional dialects did not feature as prominently. This potential bias might limit the generalizability of our findings to all European linguistic contexts. Future studies could benefit from an even broader corpus that includes more regional and less formalized language use.

The sociolinguistic implications of our findings are significant. The strategic use of English in media and literature points to a broader cultural dynamic where local identities are negotiated in relation to a dominant global language. In many cases, the insertion of English may serve as a marker of modernity, sophistication, or even aspiration toward a globalized identity. However, this trend also raises concerns regarding the erosion of local languages and cultural particularities. Educators, policymakers, and cultural preservationists must grapple with the dual challenge of embracing the communicative advantages of English while safeguarding the rich linguistic heritage of Europe.

Our results also indicate that the motivations for code-switching extend beyond mere convenience. The interplay of pragmatic needs and cultural signaling suggests that language choice is a deliberate act that reflects broader socio-cultural currents. For instance, in literary discourse, the decision to switch languages may be used to create a particular narrative voice or to engage with global literary trends. In media, on the other hand, the integration of English terms is often a calculated move to enhance credibility and reach a wider audience (Montes-Alcalá). Such insights not only contribute to our theoretical understanding of multilingual communication but also have practical implications for how media content is produced and consumed in a globalized world.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Looking ahead, the integration of AI in the study of code-switching opens several promising avenues for further research. One potential direction is the development of more sophisticated models that can better account for the pragmatics and semantics of code-switched language. This would involve integrating contextual information beyond the textual data—such as speaker demographics or situational variables—to refine the analysis. Another important future endeavor is the cross-cultural comparison of code-switching phenomena. By extending this research framework to other multilingual regions outside Europe, researchers could test the universality of our findings and explore how global English interacts with diverse linguistic ecosystems.

In conclusion, this study demonstrates that AI-driven analysis is a powerful means of exploring the dynamic and multifaceted phenomenon of code-switching in European media and literary discourse. Our findings reveal that the use of English is both a practical and symbolic act, intricately linked to the evolving identities of modern European societies. While challenges remain in fully capturing the nuance of multilingual texts, the methodological framework developed here offers a robust foundation for ongoing exploration. Ultimately, as globalization continues to shape language practices, understanding the drivers and consequences of code-switching will be critical for preserving linguistic diversity while fostering effective global communication.

CONCLUSION

This study set out to explore the phenomenon of code-switching from local European languages to English in media and literary discourse using AI-driven Natural Language Processing techniques. Our research demonstrates that the integration of advanced AI tools into sociolinguistic studies offers unprecedented insights into how global linguistic trends interact with and reshape local cultural identities.

The findings reveal that English is not merely used as a convenient lexical borrowing in European texts but serves a multifaceted role, functioning both as a marker of modernity and as a tool for cultural expression. In digital journalism and social media, the strategic insertion of English terms reflects an effort to evoke immediacy, technological prowess, and international relevance. Conversely, in literary works, code-switching is often employed more subtly, where authors interlace English expressions to evoke cosmopolitanism or to imbue their narratives with a nuanced, layered cultural identity. Such observations underscore the complexity of multilingual communication and illustrate that language choice is both a functional and symbolic act.

Our implementation of Transformer-based models, including multilingual BERT and XLM-R, proved highly effective in detecting and categorizing code-switching events across diverse datasets. The robustness of these models, evidenced by strong precision and recall metrics, confirms that AI can successfully manage the intricate task of parsing multilingual texts. However, the study also highlights certain limitations inherent in current AI methodologies. Despite rigorous annotation and preprocessing, the models sometimes struggled with idiomatic expressions and culturally nuanced language, suggesting that even advanced algorithms require continual refinement to fully capture the subtleties of human language.

The implications of this research extend beyond theoretical linguistics. For media professionals, educators, and policymakers, understanding the dynamics of code-switching is vital. Our findings suggest that while the adoption of English can enhance global communication and contemporary appeal, it also poses challenges for the preservation of local linguistic heritage. This tension calls for strategies that balance global integration with cultural preservation, ensuring that local languages remain vibrant and relevant in the digital age.

Looking forward, our study opens several avenues for future research. Further development of AI models that incorporate more contextual and pragmatic information could lead to even more nuanced analyses of code-switching phenomena. Expanding the corpus to include a broader range of sources, particularly underrepresented regional dialects and niche media outlets, will also help to refine our

understanding of how English permeates different linguistic contexts. Additionally, comparative studies across different multilingual regions could validate and extend the insights garnered from the European context.

In conclusion, this research underscores the transformative potential of AI in advancing our understanding of complex sociolinguistic phenomena. By combining large-scale data analysis with deep linguistic insight, our study not only contributes to the academic discourse on code-switching but also provides practical frameworks for addressing the challenges of linguistic globalization. As media and literature continue to evolve in response to global influences, the insights from this study will be crucial for fostering effective communication strategies that respect and preserve linguistic diversity while embracing the opportunities of a connected world.

ACKNOWLEDGEMENT

This publication has been supported by the RUDN University Scientific Projects Grant System, project No. 050738-0-000

REFERENCES

[1] Académie française. (2024). Rapport sur l'intelligence artificielle et la langue française. Paris.

[2] Bassey, G. I. (2025). Enseignement de la littérature francophone africaine et l'intelligence artificielle : vers une nouvelle méthodologie d'analyse des textes. Cascades. Journal of the Department of French & International Studies, 3(1), 9–14. Retrieved from https://cascadesjournal.com/index.php/cascades/article/view/69

[3] Briva-Iglesias, V. (2025). Are AI agents the new machine translation frontier? arXiv. Retrieved from https://arxiv.org/abs/250

[4] Financial Times. (2025, March 14). Being able to translate a website in all European languages. Financial Times. Retrieved from https://www.ft.com/

[5] Gambhir, M., & Das, D. (2023). Code-switching detection using transformer-based models: A multilingual perspective. International Journal of Computational Linguistics and Natural Language Processing, 12(2), 55–73.

[6] Joshi, A., Solorio, T., & Liu, Y. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6282–6293). https://doi.org/10.18653/v1/2020.acl-main.560

[7] Montes-Alcalá, C. (2016). Code-switching in US Latino literature: The role of bicultural identity. Language and Literature, 25(3), 241–258. https://doi.org/10.1177/0963947016650677

[8] Myers-Scotton, C. (1993). Social motivations for codeswitching: Evidence from Africa. Oxford University Press.

[9] 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

[10] Rogers, A. (2022). Multilingual BERT and beyond: Evaluating transformer models for code-switching tasks. Computational Linguistics Research, 11(4), 112–129.

[11] Sitaram, S., Choudhury, M., & Bali, K. (2020). Code-mixing in multilingual NLP: A survey of models and resources. arXiv. Retrieved from https://arxiv.org/abs/1904.00784