Board 209: Bridging Language Barriers in Healthcare Education: An Approach for Intelligent Tutoring Systems with Code-Switching Adaptation

Dr. Zechun Cao, Texas A&M University, San Antonio

Zechun Cao received his master's and Ph.D. degrees in computer science from the University of Houston. His research lies at the intersection of cybersecurity, privacy, and artificial intelligence (AI). His doctoral thesis centers around developing network and host intrusion detection methods leveraged by intelligent user behavior recognition. He also collaborates with economists and city planners on devising AI algorithms that result in long-lasting real-world impact. More recently, he has been passionate about designing algorithms and tools to keep users' private confidential data secure in an AI-driven world. Dr. Cao's work has been published in international conferences and journals. He is a member of ACM and IEEE and has served as a TPC member and reviewer for various journals and international conferences.

German Zavala Villafuerte
Ali Jalooli
Renu Balyan
Sanaz Rahimi Moosavi
Francisco Iacobelli, Northeastern Illinois University

Dr. Iacobelli is a Computer Scientist with a research focus at the intersection between human-computer interaction, natural language processing, education and artificial intelligence. He has been applying this research to healthcare and to bridge health disparities. Dr. Iacobelli is an associate professor in the Computer Science Department at Northeastern Illinois University where he has taught since 2011. He is also an associated faculty member of the Center for Advancing Safety in Machine Intelligence (CASMI) at Northwestern University.

Bridging Language Barriers in Healthcare Education: An Approach for Intelligent Tutoring Systems with Code-Switching Adaptation

Abstract: The recent rapid development in Natural Language Processing (NLP) has greatly enhanced the effectiveness of Intelligent Tutoring Systems (ITS) as tools for healthcare education. These systems hold the potential to improve health-related quality of life (HRQoL) outcomes, especially for populations with limited English reading and writing skills. However, despite the progress in pre-trained multilingual NLP models, there exists a noticeable research gap when it comes to code-switching within the medical context. Code-switching is a prevalent phenomenon in multilingual communities where individuals seamlessly transition between languages during conversations. This presents a distinctive challenge for healthcare ITS aimed at serving multilingual communities, as it demands a thorough understanding of and accurate adaptation to code-switching, which has thus far received limited attention in research.

The hypothesis of our work asserts that the development of an ITS for healthcare education, culturally appropriate to the Hispanic population with frequent code-switching practices, is both achievable and pragmatic. Given that text classification is a core problem to many tasks in ITS, like sentiment analysis, topic classification, and smart replies, we target text classification as the application domain to validate our hypothesis.

Our model relies on pre-trained word embeddings to offer rich representations for understanding code-switching medical contexts. However, training such word embeddings, especially within the medical domain, poses a significant challenge due to limited training corpora. In our approach to address this challenge, we identify distinct English and Spanish embeddings, each trained on medical corpora, and subsequently merge them into a unified vector space via space transformation. In our study, we demonstrate that singular value decomposition (SVD) can be used to learn a linear transformation (a matrix), which aligns monolingual vectors from two languages in a single meta-embedding. As an example, we assessed the similarity between the words “cat” and “gato” both before and after alignment, utilizing the cosine similarity metric. Prior to alignment, these words exhibited a similarity score of 0.52, whereas after alignment, the similarity score increased to 0.64. This example illustrates that aligning the word vectors in a meta-embedding enhances the similarity between these words, which share the same meaning in their respective languages. To assess the quality of the representations in our meta-embedding in the context of code-switching, we employed a neural network to conduct text classification tasks on code-switching datasets. Our results demonstrate that, compared to pre-trained multilingual models, our model can achieve high performance in text classification tasks while utilizing significantly fewer parameters.

Introduction

Intelligent tutoring systems (ITS) emulate human tutors on specific topics by using constructivist approaches to teaching. As such, they monitor user progress and ask challenging questions, eliciting deep explanations and providing feedback through conversation [1], [2]. ITSs have been used successfully in many domains, increasing learning performance and content recall after a medium to long term [1], [3]. ITSs can offer a scalable alternative for instruction for underserved populations [4], which can be of great help in health care-related fields, such as breast cancer, where education can greatly bridge inequity gaps [5], [6].

Breast cancer is the most common cancer diagnosis and the leading cause of cancer-related deaths among the Hispanic population [7]. Breast cancer survivors often experience ongoing diagnosis and treatment-related symptoms that negatively impact their health-related quality of life (HRQoL), including fatigue, depressive symptoms, and changes in sleep and sexual function. Hispanic Breast Cancer Survivors (BCS) are more likely to report poorer HRQoL than their White counterparts, even after adjustment for factors such as socioeconomic status [8], [9]. Previous research has shown that delivering cancer-related information, stress management, coping skills and increasing self-efficacy in communication, in a culturally appropriate intervention, can improve quality of life in particular related to health outcomes in the post-treatment survivorship phase [5], [10], [11].

Although there are some tutoring systems that delve into health topics, to the best of our knowledge there are none aimed at breast cancer survivorship [12]. Moreover, the target population for ITSs in general and for health-related ITSs in particular, has been college educated students that interact via textual interfaces (using a chatbot typing responses), and have not been built specifically for minorities, or for non-college low literacy populations [4]. Moreover, the content is rarely delivered at a reading level that is recommended by the National Institutes of Health (NIH)/ American Medical Association (AMA). Recent research shows that building ITSs to serve these minority populations is far from trivial [4].

In addition to the ITS related research gaps, another important and critical factor to be accounted for is privacy, because the interactions between a BCS and a tutoring system may involve personal, identifiable health information. The Health Insurance Portability and Accountability Act (HIPAA) of 1996 [13] requires appropriate safeguards to protect the privacy of medical records as well as other identifiable health information. However, that information needs to be utilized by the

system and be readily accessible. It is desirable that the underlying NLP models do not expose the private information contained in the training data. However, de-identification methods even when applied correctly yield data that sometimes retains the risks of identification [13]. Because most ITSs have not been geared towards patients, privacy of the models and user data has not been a major concern, therefore the research gap we intend to address is that of maintaining privacy while maintaining accuracy and low latency during NLP model training that may be computationally intensive.

2 Project Description

2.1 Project Framework

Our research endeavors to construct a framework for an Intelligent Tutoring System (ITS) that is both culturally appropriate and respects user privacy, with a specific focus on serving the Hispanic Breast Cancer Survivor (BCS) community with limited English reading and writing skills. The core of our initiative is to foster an ITS that not only comprehends but also effectively engages with this demographic through tailored interactions. This objective will be realized through the integration of two foundational components: firstly, the development of culturally appropriate language models. These models are designed to resonate with the linguistic patterns and cultural nuances characteristic of Hispanic populations with limited English reading and writing skills. This involves a meticulous process of data collection from interactions with Hispanic BCS, refining the tutoring content to address breast cancer survivorship by leveraging expert insights, and the subsequent training of NLP models with this curated data. Secondly, we aim to innovate privacy-preserving algorithms encapsulated within an API, to secure the tutoring system's conversational data. This will be achieved by applying homomorphic encryption techniques, enabling the execution of NLP tasks on encrypted data, thus ensuring the confidentiality and integrity of user interactions. A concise visual representation of our project's overarching framework and its components is provided in Figure 1.

2.2 Collaborating Institutions and Students Engagement

The project's collaboration framework is adeptly led by Northeastern Illinois University (NEIU), coordinating with three Hispanic Serving Institutions (HSIs) - California State University Dominguez Hills (CSUDH), SUNY Old Westbury (SUNYOW), and Texas A&M University-San Antonio (TAMUSA). NEIU spearheads data collection and system development. SUNYOW takes charge of NLP model training and evaluation, while CSUDH and TAMUSA focus on pioneering privacy-preserving algorithms and API development. Frequent virtual meetings and annual face-to-face gatherings, alongside focused team interactions, ensure effective collaboration and progress moni-

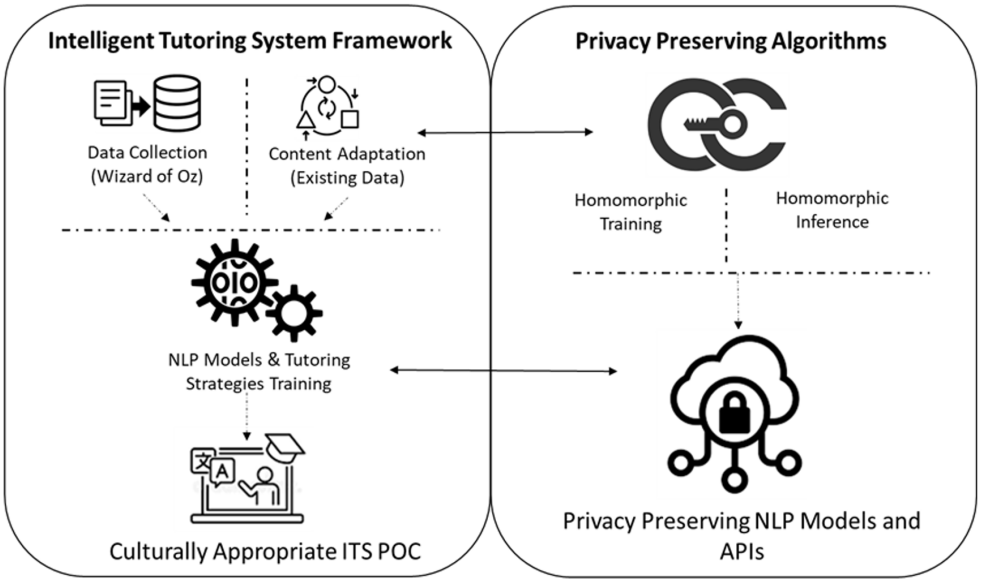

Figure 1: Our project is centered around two primary components: a culturally appropriate ITS framework and the development of privacy-protecting APIs to protect user privacy.

toring. This initiative cultivates a dynamic academic and research milieu, deeply engaging students in critical tasks, fostering cross-mentorship, and collaborative research, which not only aligns with the project's objectives but also primes students for future academic and research careers, emphasizing long-term sustainability and the prospect of future joint endeavors.

The project deeply engages students in crucial roles, fostering a vibrant academic environment. Students participate in content creation, model evaluation, and notably, disseminate their research findings at national conferences. This exposure not only enriches their academic experience but also enhances their professional development, preparing them for future careers in academia and research. The initiative's emphasis on cross-mentorship and collaborative research further strengthens the project's impact, setting a foundation for sustained collaboration and future endeavors across institutions.

3 Adapting Spanish-English Code-Switching in ITS

3.1 Background

The term "code-switching" is used both inside and outside the field of linguistics. Informally, code-switching is sometimes used to refer to relatively stable informal mixtures of two languages, such as Spanglish, Taglish, or Hinglish [14]. Some scholars of literature use the term to describe literary styles that include elements from more than one language, as in novels by Chinese-American,

Anglo-Indian, or Latino writers [15]. Code-switching can take on various forms, but in this paper, we define it as the use of both Spanish and English in bilingual communication. This showcases the intricate linguistic dance that bilingual speakers engage in, which reflects a blend of linguistic choice and cultural narrative. This phenomenon is especially prominent in communities where both languages are woven into the social and cultural fabric, allowing individuals to fluidly navigate their bilingual identities [16], [17]. Beyond simple language mixing, code-switching incorporates a sophisticated amalgamation of grammatical structures, cultural cues, and contextual relevance, highlighting the cognitive dexterity of bilingual speakers in enriching their communication and expression [18].

In educational contexts, the presence of Spanish-English code-switching reveals valuable insights into pedagogical strategies that can support bilingual learners, particularly in challenging traditional monolingual frameworks. By acknowledging code-switching as a valid linguistic and cultural expression, educational systems can foster more inclusive and affirming spaces for bilingual students [19]. Such an approach, which values the natural bilingual practices of students, including code-switching, can enhance learning by offering cognitive and linguistic support, thereby validating students' linguistic identities and leveraging their bilingualism as an asset in the educational process [20]. According to a study by Shafi et al. [21], students better learned and understood their non-dominant language when the teacher practiced code-switching.

Extending the concept of code-switching to ITS presents a promising avenue for enriching education. For Hispanic breast cancer survivors, an ITS that adeptly incorporates Spanish-English code-switching can significantly improve understanding and engagement with crucial health information. This personalized, culturally attuned educational tool can bridge language barriers, facilitating better health management and informed decision-making. Moreover, the ability of an ITS to navigate code-switching effectively can bridge communication gaps in patient education, especially in areas with high bilingual populations. For instance, in explaining complex medical conditions or treatment plans, patients may find it easier to understand and retain information presented in a mix of both languages, aligning with their everyday language use [22]. This bilingual approach can lead to better patient engagement, comprehension, and adherence to treatment, ultimately improving health outcomes. Therefore, the incorporation of Spanish-English code-switching capabilities in ITS not only enhances linguistic accessibility but also embodies a patient-centered approach, acknowledging and respecting the linguistic diversity of the user base.

In this paper, we delineate our endeavors in integrating Spanish-English Code-Switching within NLP models. These initiatives culminate in the creation of culturally appropriate language models

that are employed within ITS.

3.2 Data Collection and Adaptation

We are building a Wizard of Oz (WOZ) interface version of the ITS – a prototype that simulates a fully operational system, yet certain modules are manually operated by an individual (the "wizard"). This approach enables researchers to examine user interactions without the need for a completely autonomous ITS. We are developing a tutoring script to manage the WOZ system. The wizard, who is a trained research assistant, actively prompts users for in-depth reflections and responses on the subjects, guided by this script, aiming to elicit extensive dialogue. The speech collected during these interactions is being transcribed by a speech recognition engine, which will serve to refine natural language processing algorithms, enhancing their ability to capture the unique dynamics of interactions between our target demographic and the ITS.

We aim to recruit at least 30 Hispanic women, in stage III breast cancer remission for over a year, for a study involving a two-hour session that includes surveys, a survivorship skills video, and system interaction training. Data, including text from speech recognition and screen/voice recordings, will be anonymized for analysis to refine a tutoring system's language model and identify interaction patterns, such as pauses and self-corrections. We anticipate discovering specific linguistic features and potential errors, like word misappropriation and code-switching, to enhance NLP model training.

Following the data collection, we adapt and tailor educational content specifically for Hispanic BCS, ensuring cultural appropriateness and simplicity for effective comprehension. This involves repurposing existing content, previously developed for a mobile app, into a format suitable for ITS-based tutoring. The adaptation process includes creating a tutoring script that breaks down the content into manageable subtopics, formulating questions designed to elicit detailed responses, and providing varied feedback to user inputs. This task aims to transform the content into an interactive learning experience that resonates with the users' cultural and linguistic background, enhancing the ITS's effectiveness in delivering health education tailored to Hispanic BCS.

3.3 Multilingual Meta-Embeddings for Code-Switching

To overcome the code-switching problem, we adopt a multilingual meta-embedding model learned from different languages. Our approach can be seen as a method to create a universal multilingual meta-embedding learned in a supervised way with code-switching contexts by gathering information from monolingual sources. Concurrently, this is a language-agnostic approach where it does

not require any language information of each word. We show the possibility of transferring information from multiple languages to unseen languages, and this approach can also be useful for a low-resource setting. To effectively leverage the embeddings, we use FastText subwords information to solve out-of-vocabulary (OOV) issues.

The application of Singular Value Decomposition (SVD) in learning word meta-embeddings offers a sophisticated approach for integrating diverse pre-trained word embeddings into a unified representation, particularly useful for code-switching scenarios. SVD, a matrix factorization technique, decomposes a matrix A into three distinct matrices U, Σ, and V^T, capturing the essential properties of the original data in a reduced-dimensional space. In the context of meta-embeddings, SVD is applied to a matrix that aggregates multiple word embeddings, aiming to derive a singular embedding space that retains the salient features of each original set.

Mathematically, SVD decomposes the matrix A into U, Σ, and V^T, where $A = U\Sigma V^T$. Here, U and V are orthogonal matrices representing the left and right singular vectors, respectively, and Σ is a diagonal matrix containing the singular values. To utilize SVD for meta-embedding learning, the matrix Σ is truncated to keep only the top k singular values, along with the corresponding vectors in U and V. This results in a reduced meta-embedding matrix $A_k = U_k\Sigma_k V_k^T$, which encapsulates the core semantic dimensions from the combined embeddings, enhancing the processing of code-switched text.

This approach, as investigated by Bamman et al. [23] and Yin et al. [24], allows for the effective integration of linguistic information from multiple embedding sources, offering significant improvements in NLP models' handling of code-switched data.

3.4 Preliminary Results and Discussion

In the experimental setup of our study, we use pre-trained FastText English (EN) and Spanish (ES) word embeddings [25] as our primary language embeddings, specifically targeting English and Spanish to evaluate the effectiveness of SVD in aligning linguistic representations. We employed SVD to derive a linear transformation matrix that harmonizes the English vectors with their Spanish equivalents within an integrated meta-embedding space. This approach is intended to reduce the disparity between corresponding word pairs across the languages while maintaining the intrinsic semantic relationships among words within each language. Our methodology entailed a comparative analysis between the SVD alignment technique and a baseline method that relied on distinct monolingual embeddings for English and Spanish. We utilized cosine similarity, ranging

from 0 to 1, as a measure to ascertain the semantic closeness of word pairs across the two languages. A score closer to 0 suggests a lack of semantic similarity, indicative of orthogonal word vectors, whereas a score nearing 1 signals a strong semantic similarity. For example, the semantic correlation between "cat" in English and its Spanish counterpart "gato" was initially measured at 0.52 using the baseline approach but exhibited a noticeable increase to 0.64 with the implementation of SVD alignment. This improvement underscores a more coherent semantic congruence between the word pairs within the integrated meta-embedding space facilitated by SVD.

Table 1: Cosine Similarity of Medical Terminology Pairs

English	**Spanish**	**Baseline**	**Meta-Embedding**
fracture	fractura	0.58	0.74
deaf	sordo	0.33	0.47
blind	siego	0.07	0.24
hypertension	hipertension	0.41	0.70
pneumonia	neumonia	0.44	0.65
fever	fiebre	0.56	0.75
cough	toz	0.07	0.20
chills	escalofrios	0.37	0.57
thrombus	trombo	0.28	0.53
dizziness	mareo	0.33	0.55
bronchitis	bronquitis	0.48	0.69
cataract	catarata	0.49	0.63
arthritis	artritis	0.42	0.65
appendicitis	apendicitis	0.50	0.71
leukemia	leucemia	0.55	0.75
lymphoma	linfoma	0.46	0.64

To deepen our analysis of the meta-embedding's efficacy, we curated a selection of medical terminology pairs in both Spanish and English. This allowed us to meticulously compare the cosine similarity scores between the baseline method and our meta-embedding approach, providing a focused lens on the performance within a specialized vocabulary context. As shown in Table 1, our initial findings indicate that the meta-embedding approach consistently improves semantic correlations between English-Spanish medical term pairs. We observe an overall elevation in cosine similarity by an average of 19.26% across 16 term pairs, demonstrating the alignment's effectiveness in crafting a cohesive semantic space that bridges multiple languages. This bolstered representation is likely to enhance the sophistication and accuracy of the ITS when handling texts featuring Spanish-English code-switching. The implications of these findings extend beyond immediate results, catalyzing a broader discourse on the adaptability of this approach to other languages and the nuanced complexities of various linguistic forms beyond simple nouns. Consequently, this paves

the way for comprehensive research efforts to delineate the full potential and constraints of SVD applied to multilingual meta-embedding scenarios.

4 Future Work

Our preliminary results led us to hypothesize that the improved alignment of monolingual vectors through SVD and the subsequent creation of meta-embeddings could significantly enhance performance in text classification tasks within ITS. Given the nuanced nature of language learning and the diverse linguistic backgrounds of learners, ITS can greatly benefit from a more sophisticated understanding of language semantics that transcends linguistic boundaries. Future work will explore the application of these meta-embeddings in classifying text across various educational content and learner interactions. By leveraging the enriched semantic space that meta-embeddings provide, we aim to improve the accuracy of classifying learner responses, feedback personalization, and content recommendation in ITS, thereby facilitating a more adaptive and responsive learning environment.

In addition to applying meta-embeddings to text classification, we plan to utilize this method to train NLP models on the dataset we have collected and adapted as mentioned in the previous section. This dataset, enriched with aligned multilingual vectors, presents an opportunity to develop NLP models that are inherently more robust to the linguistic variability inherent in code-switched data. By training models on this enriched dataset, we aim to advance the state-of-the-art in processing mixed-language content, particularly in educational settings where code-switching is prevalent. This endeavor will not only contribute to the field of multilingual NLP but also pave the way for more inclusive and linguistically aware educational technologies.

5 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2219586, No. 2219587, No. 2219588, No. 2219589. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, "Autotutor: an intelligent tutoring system with mixed-initiative dialogue," *IEEE Trans. Educ.*, vol. 48, no. 4, pp. 612–618, 2005. [Online]. Available: http://dblp.uni-trier.de/db/journals/te/te48.html#GraesserCHO05

[2] J. A. Kulik and J. D. Fletcher, "Effectiveness of intelligent tutoring systems: A meta-analytic review," *Review of Educational Research*, vol. 86, no. 1, pp. 42–78, 2016.

[3] K.-I. Malatesta, P. Wiemer-Hastings, and J. Robertson, "Beyond the short answer question with research methods tutor," in *ITS*. London, UK: Springer-Verlag, 2002, pp. 562–573. [Online]. Available: http://portal.acm.org/citation.cfm?id=648031.743916

[4] Y. Fang, A. Lippert, Z. Cai, S. Chen, J. Frijters, D. Greenberg, and A. Graesser, "Patterns of adults with low literacy skills interacting with an intelligent tutoring system," *International Journal of Artificial Intelligence in Education*, 2021.

[5] G. Juarez, L. Mayorga, A. Hurria, and B. Ferrell, "Survivorship education for latina breast cancer survivors: Empowering survivors through education," *Psicooncologia*, vol. 10, no. 1, pp. 57–68, 2013, pMID: 24416043.

[6] S. Williams and A. Schreier, "The effect of education in managing side effects in women receiving chemotherapy for treatment of breast cancer," *Oncology Nursing Forum*, vol. 31, no. 1, pp. E16–E23, 2004.

[7] American Cancer Society, "Cancer facts and figures for hispanic/latinos 2018-2020," https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-facts-and-figures-for-hispanics-and-latinos/cancer-facts-and-figures-for-hispanics-and-latinos-2018-2020.pdf, 2020, accessed: 2024-02-06.

[8] B. Yanez, E. H. Thompson, and A. L. Stanton, "Quality of life among latina breast cancer patients: a systematic review of the literature," *Journal of Cancer Survivorship*, vol. 5, no. 2, pp. 191–207, 2011.

[9] T. Luckett, D. Goldstein, P. N. Butow, V. Gebski, L. J. Aldridge, J. McGrane, W. Ng, and M. T. King, "Psychological morbidity and quality of life of ethnic minority patients with cancer: A systematic review and meta-analysis," *Lancet Oncology*, vol. 12, no. 13, pp. 1240–1248, 2011.

[10] B. Yanez, M. Maggard Gibbons, P. I. Moreno, A. Jorge, and A. L. Stanton, "Predictors of psychological outcomes in a longitudinal study of latina breast cancer survivors," *Psychology and Health*, vol. 31, no. 11, pp. 1359–1374, 2016.

[11] K. D. Graves, R. E. Jensen, J. Cañar, M. Perret-Gentil, K.-G. Leventhal, F. Gonzalez, L. Caicedo, L. Jandorf, S. Kelly, and J. Mandelblatt, "Through the lens of culture: quality of life among latina breast cancer survivors," *Breast Cancer Research and Treatment*, vol. 136, no. 2, pp. 603–613, 2012.

[12] C. R. Wolfe, V. F. Reyna, C. L. Widmer, E. M. Cedillos, C. R. Fisher, P. G. Brust-Renck, and A. M. Weil, "Efficacy of a web-based intelligent tutoring system for communicating genetic risk of breast cancer: a fuzzy-trace theory approach," *Medical Decision Making*, vol. 35, no. 1, pp. 46–59, 2015, accessed: 2024-02-05.

[13] U.S. Department of Health & Human Services, "Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule," https://www.hhs.gov/hipaa/index.html, 2012.

[14] A. C. Zentella, *Growing Up Bilingual: Puerto Rican Children in New York*. Malden, MA: Blackwell Publishers, 1997.

[15] L. Torres, "In the contact zone: Code-switching strategies by latino/a writers," *MELUS*, vol. 32, no. 1, pp. 75–96, 2007. [Online]. Available: https://www.jstor.org/stable/30029707

[16] F. Grosjean, *Life with Two Languages: An Introduction to Bilingualism*. Harvard University Press, 1982.

[17] S. Poplack, "Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching," *Linguistics*, vol. 18, no. 7-8, pp. 581–618, 1980.

[18] C. Myers-Scotton, *Social Motivations for Code-Switching: Evidence from Africa*. Oxford University Press, 1993.

[19] O. García and L. Wei, *Translanguaging: Language, Bilingualism and Education*. Palgrave Macmillan, 2014.

[20] M. Martin-Beltrán, "The role of language brokering in the learning experiences of bilingual students," *Journal of Language, Identity & Education*, vol. 13, no. 2, pp. 93–112, 2014.

[21] S. Shafi *et al.*, "Benefits of code-switching in language learning classroom at university of education lahore," *International Research Journal of Management, IT and Social Sciences*, vol. 7, no. 1, pp. 227–234, Jan. 2020. [Online]. Available: https://doi.org/10.21744/irjmis.v7n1.842

[22] L. J. Rodríguez-Fuentes and M. Fuentes, "Code-switching in health its: Bridging the language gap in healthcare," *Journal of Health Informatics*, vol. 21, no. 3, pp. 205–218, 2015.

[23] D. Bamman and N. A. Smith, "Distributed representations of geographically situated language," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 828–838, 2014.

[24] W. Yin and H. Schütze, "Learning word meta-embeddings," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016, pp. 1351–1360.

[25] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.