On the communicative utility of code-switching

Yanting Li, Gregory Scontras, and Richard Futrell
Department of Language Science
University of California, Irvine
{yantil5, g.scontras, rfutrell}@uci.edu

Abstract

In the multilingual world we live in, code-switching (CS) is becoming more natural and more common. Why do bilingual language users CS from one language (the source language) to another (the target language) during communication, and how do they decide the CS point? In this corpus study, we investigate the hypothesis that it is harder to accurately express the meaning represented by the CS words in the source language. We analyzed sentences containing CS from Chinese–English bilingual corpora and found evidence for our hypothesis: compared to non-CS words, the English CS words are farther away from their closest Chinese word neighbors in a bilingual meaning space. This result supports the idea that bilinguals are using CS as a communication strategy to express their intended meanings accurately and efficiently.

1 Introduction

Code-switching (CS) refers to the scenario where a language user switches from one language to another during communication (Solorio et al., 2014; Adel et al., 2015; Zhou et al., 2020; Beatty-Martínez et al., 2020; Tomić and Valdés Kroff, 2022). The phenomenon is widely observed, both in spoken (e.g. Fricke and Kootstra, 2016; Heredia and Altarriba, 2001; Deuchar et al., 2014; Nguyen and Bryant, 2020) and written (e.g. Calvillo et al., 2020; Chang and Lin, 2014; Feldman et al., 2021; Chakravarthi et al., 2020) language use. Globalization has built stronger connections between countries and cultures; for English alone, there are over 1 billion people speaking it as a second language. The increase in multilingual speakers, together with the global status of English, has made CS involving English more and more common (Nakayama et al., 2018; Chakravarthi et al., 2020). As language scientists, we are charged with looking deeper into the process behind CS to better understand the communicative strategy of multilingual speakers.

Why do people code-switch? More specifically, what factors influence the choice to switch at certain words of an utterance but not others? Previous research has approached this question from different angles. Several factors have been shown to play a role in determining the CS point. For instance, word length: the longer a word, the more likely you are to switch to another language (where it may be shorter) to express that meaning (Myslín and Levy, 2015; Calvillo et al., 2020; Bhattacharya and van Schijndel, 2023). The syntactic role of the word is another factor: nouns are more likely to be CSed than verbs, function words, etc. (Myslín and Levy, 2015; Calvillo et al., 2020; Bhattacharya and van Schijndel, 2023). Semantic factors such as concreteness also play a role: more concrete words are more likely to be CSed (Myslín and Levy, 2015).

Another widely-discussed factor is predictability as operationalized by surprisal, the negative log probability of a word given context (Hale, 2001; Levy, 2008; Hale, 2016): CS words tend to have higher surprisal, meaning that these words are relatively less predictable from the context (Myslín and Levy, 2015; Calvillo et al., 2020; Bhattacharya and van Schijndel, 2023). There are two potential explanations for the role of surprisal in CS: according to a speaker-oriented explanation, words with higher surprisal impose more difficulty for production, and since speakers have limited cognitive resource, this will result in a weaker inhibition on the target language, letting words from that language "slip out" (Calvillo et al., 2020). Meanwhile, Myslín and Levy (2015) proposed an audience-oriented explanation: the words with higher surprisal need more attention from the listener, so the speaker will switch to a less frequent, and thus more salient language to alert the listeners of upcoming information peaks.

In this paper, we explore another aspect of efficiency: the communicative utility of CS words. Intuitively, the CS word in the language we switch

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 343-349.
Irvine, California, June 27-29, 2024

into might better express our intended meaning, as the source language may not have a word that expresses exactly the same meaning, even when there is a direct translation. For instance, 地下室 *dìxiàshì* in Mandarin Chinese is officially equivalent to English *basement*. However, the housing situation in China is very different from that of North America—there are far more tall apartment buildings than single-family homes in China. Because of this, when a Chinese–English bilingual hears the English word *basement*, the picture they have in mind might be different from the picture triggered by the Chinese word 地下室 *dìxiàshì*. Therefore in a Chinese conversation among Chinese–English bilinguals, when talking about the basement of a single-family home in the US, the speaker might consider switching into English for this word to achieve greater accuracy. In contrast, the English word *cat* expresses nearly exactly the same meaning as the Chinese word 猫 *māo* and so bilingual speakers may be less likely to CS for such a word. Similarly, Heredia and Altarriba (2001) provided an example in Spanish-English bilingual communication: the Spanish word *cariño* implies a combination of liking and affection, which cannot be expressed by an English word alone. Therefore, if a Spanish-English bilingual wish to refer to this concept, they would consider using Spanish to achieve a greater level of understanding.

In this research, we aim to test this hypothesis: people code switch when it is harder to express their intended meaning accurately in the source language.

2 Method

To see if a language has a vocabulary item that allows its speakers to express a certain meaning, we rely on word vectors, which help us locate words in a meaning space (Mikolov et al., 2013b,a; Bojanowski et al., 2017). In the meaning space, words with similar or related meanings are located close to each other while words with distant or unrelated meanings are located far away from each other. If there is a *bilingual* meaning space where words in both English and Chinese can be found, then for our hypothesis to be true, the English CS words should be located far away from any Chinese words in such a space, meaning no Chinese word has a meaning close enough to the CS words. To turn our hypothesis into something measurable, we choose to look for the closest Chinese word neighbor of each CS word and calculate the 1) distance and 2) cosine similarity between the two. We will then do the same for the English translation of comparable non-CS words. We predict that, compared to non-CS words, the CS words have 1) longer distance to and 2) smaller cosine similarity with their closest Chinese word neighbor.

2.1 Materials

In order to conduct the above comparison, we need a bilingual meaning space for English and Chinese words. We also need a number of CS and non-CS words from natural language production.

Bilingual meaning space We use aligned word vectors to create the bilingual meaning space. While word vectors of a specific language can be used to locate words in the meaning space of that language, *aligned* word vectors are pre-trained to align meaning spaces of multiple languages (Smith et al., 2017; Conneau et al., 2018), so words from these languages can exist in the same space. We used the aligned word vectors of Chinese and English created by Bojanowski et al. (2017) and Joulin et al. (2018) based on the pre-trained vectors computed on Wikipedia. As the aligned word vectors are sorted by frequency, the top 150k English vectors and the top 150k Chinese ones are taken out and combined to create a bilingual vector space with 300k words. For any two word vectors, Chinese or English, in this space, their distance and cosine similarity tell us about how similar their meanings are to each other.

Code-switching corpora Two Chinese–English bilingual corpora are used: one written corpus and one spoken corpus. The written one consists of posts on Chinese international student forums of three universities in Pittsburgh (Calvillo et al., 2020). The content is mainly about housing, schooling, and life in Pittsburgh. The spoken corpus, on the other hand, is built on spontaneous multi-turn conversational dialogue sources collected in Hong Kong (Lovenia et al., 2022), covering topics on education, persona, philosophy, sports, and technology. In both corpora, native speakers of Mandarin Chinese (who also happen to be bilingual speakers of English) are communicating with each other, yet they choose to CS into their second language, English, at certain points.

In the written corpus, each CS sentence is paired with a structurally similar monolingual Chinese sentence. For instance:

CS sentence:
客厅还有一个小的balcony。
The living room also has a small balcony.

Matching sentence:
厨房面积大，还有一个小的吧台。
The kitchen size is big, and also has a small bar.

The two sentences have at least a 40% Levenshtein similarity of their POS sequences, and the matching sentence contains the same POS trigram as the CS point and the words before and after it (Calvillo et al., 2020). In this example, the word *balcony* and 吧台 *bātái* "bar" have the same POS tag and appear in a similar syntactic environment, but one is CSed while the other one is not, allowing us to make a close comparison of the word pair. Following the above two criteria, we found matching sentences for all CS sentences in the spoken corpus as well. If none of the monolingual Chinese sentences fulfilled both criteria for a CS sentence, the sentence was excluded from the analysis.

2.2 Procedure

We make the simplifying assumption that the words used in the actual language production, whichever language they are in, best express the intended meaning of the speaker. Based on this assumption, we extracted three groups of words from the corpora:

CS nouns While some instances of CS involve short phrases or compound words, we limited our focus to single-worded instances, specifically nouns. This is because only single words can be found in the bilingual meaning space, and nouns are the most likely to get CSed (Myslín and Levy, 2015; Calvillo et al., 2020; Bhattacharya and van Schijndel, 2023). We found 199 CS nouns in the written corpus and 531 in the spoken corpus that can be located within our bilingual meaning space.

English translations of matching non-CS nouns As previously shown, each CS sentence in the corpora is paired with a syntactically similar monolingual Chinese sentence. This is to say, each CS noun (e.g., *balcony*) has a matching noun in the monolingual Chinese sentence (e.g., 吧台 *bātái* "bar"). We used `googletrans` (Han, 2020) to translate all these matching non-CS nouns into English. If a CS noun appears multiple times in the corpus, resulting in multiple matching non-CS nouns, we kept all that have a single-worded English translation that can be found in the bilingual meaning space. If none of the matching words of a CS noun has a single-worded English translation, or none of the translations can be found in the meaning space, the CS noun was excluded. Take the word *basement* as an example: it appeared as a CS word in 8 different CS sentences in the written corpus, each matched with a different monolingual Chinese sentence. Therefore, there are 8 different matching non-CS nouns, namely 车 "car", 客厅 "living room", 里面 "inside", 存储 "storage", 屋内 "indoor", 门口 "entrance", 兼职 "part time" and 学校 "school". Among these 8 non-CS nouns, only 6 have a single-worded English translation, and all 6 can be found in the bilingual meaning space, so these 6 words are kept as the matching nouns for *basement*. Meanwhile, for the word *balcony*, since it only appeared once in the whole corpus, it only has one match, which is *bar*. We ended up with 176 CS nouns in the written corpus and 477 in the spoken corpus with at least one matching non-CS noun.

English translations of random non-CS words To create a larger pool of non-CS words that are not limited to nouns, we gathered all words that appear in the monolingual Chinese sentences from each corpus and kept the ones with single-worded English translations (according to `googletrans`; Han, 2020) that can be found in the bilingual meaning space. 1425 non-CS words remained for the written corpus and 2181 for the spoken corpus.

For each English word in the above three groups, we located the word in the bilingual meaning space and found the word in simplified Chinese located closest to it. We then used the vectors of both words to calculate their Euclidean distance as well as cosine similarity.

3 Analysis

CS nouns vs. non-CS nouns We conducted paired t-tests between the CS vs. non-CS noun pairs (e.g., *balcony* and *bar* in the example earlier). As some CS nouns appear multiple times in one corpus (e.g., *basement*), resulting in multiple matching non-CS nouns, five samples were randomly selected for a paired t-test. The CS nouns are the same across the samples, while the matching nouns may be different. This is to say, for *basement*, its matching noun could be *school* in sample 1, *storage* in sample 2, *car* for sample 3, etc. For both corpora, between the CS nouns and

Corpus	Sample	Distance	t Statistic	p-value	Cosine Similarity	t Statistic	p-value
written	CS	1.062	—	—	0.434	—	—
	non-CS 1	1.030	4.404	<0.001	0.468	-4.422	<0.001
	non-CS 2	1.030	4.335	<0.001	0.468	-4.368	<0.001
	non-CS 3	1.028	4.698	<0.001	0.469	-4.526	<0.001
	non-CS 4	1.029	4.559	<0.001	0.468	-4.556	<0.001
	non-CS 5	1.030	4.443	<0.001	0.468	-4.454	<0.001
spoken	CS	1.048	—	—	0.448	—	—
	non-CS 1	1.036	2.739	0.006	0.461	-2.821	0.005
	non-CS 2	1.036	2.886	0.004	0.461	-2.930	0.004
	non-CS 3	1.038	2.323	0.021	0.459	-2.349	0.019
	non-CS 4	1.034	3.179	0.002	0.463	-3.196	0.001
	non-CS 5	1.038	2.235	0.026	0.459	-2.269	0.024

Table 1: Mean distances and cosine similarities from English words to their nearest equivalents in Chinese. We show statistics from paired t-tests, comparing the actually-produced CS nouns against the non-CS nouns, for both measures. The labels of non-CS 1 through 5 represent the five samples of matching non-CS nouns that are randomly drawn. The $df = 476$ for the spoken corpus and $df = 175$ for the written corpus.

their closest Chinese word neighbors, the mean distance is significantly larger than that of non-CS nouns; the mean cosine similarity is significantly smaller (Table 1).

CS nouns vs. non-CS words In addition to the paired comparison between CS and matching non-CS nouns, we are also curious about whether CS nouns are different from non-CS words in general. Therefore, we used the `boot` library in R (Canty and Ripley, 2022; Davison and Hinkley, 1997) to bootstrap the 95% confidence interval of the mean distance and mean cosine similarity using the data of the English translations of non-CS words from both corpora ($n = 1425$ for the written corpus and $n = 2181$ for the spoken one). We then calculated the mean values of the CS nouns from each corpus ($n = 199$ for the written corpus and $n = 531$ for the spoken one) and examined whether they fall outside of the confidence intervals. The results are shown in Table 2 and visualized in Fig. 1. As we can see, the mean values of the CS nouns (the red dots in Fig. 1) are all outside of their corresponding 95% confidence interval.

4 Discussion

In this paper, we aimed to investigate why bilingual language users code switch during natural communication. We proposed that it is because of the communicative utility of CS and hypothesized that people choose to switch when it is harder to express their intended meaning accurately in the source language—there may not be a salient word in the source language that means the same as the CS word. While this may be a clear intuition for many bilingual speakers, we are not aware of any existing studies that measure this using naturalistic language production data. Here we proposed a way to quantitatively measure the communicative utility of CS. We tested our hypothesis by locating words from both languages in the same meaning space; the CS words in the target language should be far away from any words in the source language. Conversely, the cosine similarity between the CS word and its closest word neighbor in the source language should be small.

Our comparisons between the CS nouns vs. matching non-CS nouns and between the CS nouns vs. non-CS words in general show evidence

Corpus	Measure	Mean of CS nouns	95% confidence interval of non-CS words
written	Distance	1.064	(1.036, 1.043)
	Cosine Similarity	0.432	(0.454, 0.461)
spoken	Distance	1.047	(1.040, 1.045)
	Cosine Similarity	0.450	(0.452, 0.457)

Table 2: Mean distance and cosine similarity of CS nouns to their closest Chinese word neighbor in comparison to the bootstrapped 95% confidence interval of non-CS words. The data is visualized in Fig 1.

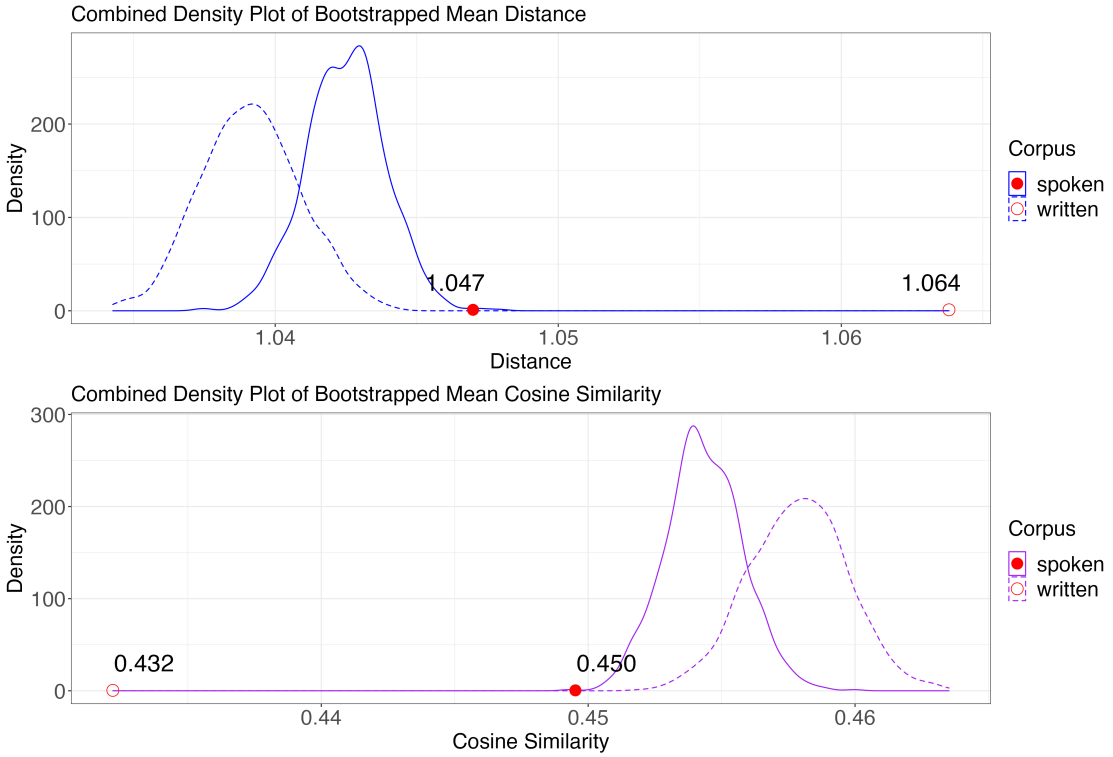

Figure 1: Density plots of the mean distance (top) and mean cosine similarity (bottom) bootstrapped from the non-CS words from the two corpora. The dashed line indicates data from the written corpus ($n = 1425$); the solid line indicates data from the spoken corpus ($n = 2181$). The red dots represent the mean values of the CS nouns, with the hollow dots for the written corpus ($n = 199$) and solid dots for the spoken corpus ($n = 531$).

supporting our hypothesis. Between the CS nouns in English and their closest Chinese neighbors, the distance is significantly larger and the cosine similarity is significantly smaller. This suggests that it is harder to pick a Chinese word to express the exact meaning of the English CS word. This is not to say that the meaning cannot be expressed accuratly using Chinese at all—it might be possible if the speaker uses a combination of multiple Chinese words. However, CS is perhaps a faster, shorter, and therefore more efficient choice to achieve the communication goal.

It is worth noticing that the difference between the CS-nouns and non-CS nouns or words are consistently smaller for the spoken corpus when compared to those for the written corpus. One potential explanation for this trend is that people are under more time pressure when having a real-time spoken conversation compared to writing forum posts. This pressure means that when an English word expresses the intended meaning most accurately, even when there are Chinese words nearby in the meaning space, the speaker may not have enough time to search for such words. As a result, they are more likely to produce CS. This is consistent with what was proposed by Calvillo et al. (2020), i.e. spoken language production allows CS to happen more frequently, although they see it as a result from the decreased cognitive resources to inhibit the alternative language. Another factor making CS more likely in spoken as opposed to written communication is that the switch cost is likely to be higher when typing than speaking, as it usually involves a switch of input keyboard. This cost will potentially create more resistance against CS, so typers are more motivated to search carefully in the meaning space around the English CS word for a Chinese equivalent, and only switch when it is sufficiently difficult to find anything with a close-enough meaning. These two factors, namely time pressure and switching cost, work in the same direction towards the difference we observed between the two corpora. This suggests that the mode of communication could affect the weight we assign to the communicative utility when making CS decisions.

Despite the above difference in effect size, the results from both corpora show consistent results that support our hypothesis. We thus contribute to the existing literature by identifying one more factor—the difficulty to accurately express a certain meaning in the source language—that may influence people's decision on whether or not to CS, as well as where to switch. With CS becoming more popular all over the world, we hope to better explain this phenomenon and better understand CS as a communicative strategy that bilinguals utilize to achieve communication goals effectively and efficiently.

Acknowledgements

We thank Debasmita Bhattacharya and Marten van Schijndel for helpful discussion. We also thank the audience of the 2024 California Meeting on Psycholinguistics for their comments and feedback.

References

Heike Adel, Ngoc Thang Vu, Katrin Kirchhoff, Dominic Telaar, and Tanja Schultz. 2015. Syntactic and semantic features for code-switching factored language models. *IEEE/ACM transactions on audio, speech, and language Processing*, 23(1):431–440.

Anne L Beatty-Martínez, Christian A Navarro-Torres, and Paola E Dussias. 2020. Codeswitching: A bilingual toolkit for opportunistic speech planning. *Frontiers in Psychology*, 11:1699.

Debasmita Bhattacharya and Marten van Schijndel. 2023. Code-switching in online posts reveals information-theoretic audience design. In *Human Sentence Processing 2023*, Pittsburgh, PA.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jesús Calvillo, Le Fang, Jeremy Cole, and David Reitter. 2020. Surprisal predicts code-switching in chinese-english bilingual text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4029–4039.

Angelo Canty and B. D. Ripley. 2022. *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-28.1.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. *arXiv preprint arXiv:2006.00206*.

Joseph Chee Chang and Chu-Cheng Lin. 2014. Recurrent-neural-network for language detection on twitter code-switching corpus. *CoRR*, abs/1412.4314.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

A. C. Davison and D. V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge. ISBN 0-521-57391-2.

Margaret Deuchar, Peredur Davies, Jon Herring, M Carmen Parafita Couto, and Diana Carter. 2014. Building bilingual corpora. *Advances in the Study of Bilingualism*, pages 93–111.

Laurie Beth Feldman, Vidhushini Srinivasan, Rachel B. Fernandes, and Samira Shaikh. 2021. Insights into codeswitching from online communication: Effects of language preference and conditions arising from vocabulary richness. *Bilingualism: Language and Cognition*, 24(4):791–797.

Melinda Fricke and Gerrit Jan Kootstra. 2016. Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, 91:181–201.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1–8. Association for Computational Linguistics.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

SuHun Han. 2020. googletrans 3.0.0.

Roberto R Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science*, 10(5):164–168.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Xu Yan, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J Barezi, et al. 2022. Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Mark Myslín and Roger Levy. 2015. Code-switching and predictability of meaning in discourse. *Language*, pages 871–905.

Sahoko Nakayama, Takatomo Kano, Quoc Truong Do, Sakriani Sakti, and Satoshi Nakamura. 2018. Japanese-english code-switching speech data construction. In *2018 Oriental COCOSDA - International Conference on Speech Database and Assessments*, pages 67–71.

Li Nguyen and Christopher Bryant. 2020. CanVEC - the canberra Vietnamese-English code-switching natural speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4121–4129, Marseille, France. European Language Resources Association.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Aleksandra Tomić and Jorge R. Valdés Kroff. 2022. Expecting the unexpected: Codeswitching as a facilitatory cue in online sentence processing. *Bilingualism: Language and Cognition*, 25(1):81–92.

Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, and Haizhou Li. 2020. End-to-end code-switching tts with cross-lingual language model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7614–7618.