# The meaning behind a code-switch

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

Why do people code-switch (CS)? This corpus study presents evidence that, compared to non-CS words, it is harder to express the meaning represented by the CS words accurately in the source language. This finding supports the idea that people use CS as a communication strategy to express their intended meanings accurately and efficiently.

**Introduction** Code-switching refers to the scenario where a language user switches from one language (the source language) to another (the target language) during communication [1]. Why would we CS at certain words of an utterance but not the others? Previous research has identified factors including the length, syntactic role, concreteness, as well as surprisal of the word [2][3][4]. In this paper, we explore the communicative utility of CS. We hypothesize that people CS when it is harder to express their intended meaning accurately in the source language.

**Method** To see whether the source language has a vocabulary item that has the same meaning of a certain CS word, we use pre-trained aligned word vectors which help us locate words in *different* languages in the *same* vector space [5][6]. Words with similar meanings locate closer to each other, with a larger cosine similarity. We predict that CS words are located far away from its closest word neighbor in the source language, with a smaller cosine similarity, suggesting that its meaning cannot be accurately expressed in the source language.
We first combined the most frequent 150k aligned word vectors of English and Chinese [7][8] to create a bilingual vector space. We then found two corpora containing CS: a written one extracted from university forums mainly about housing [3], and a spoken one consisting of spontaneous multi-turn conversations on education, persona, philosophy, sports, and tech [9]. As CS words are usually nouns [3][4], we obtained from each corpus: (1) a pool of English **CS nouns**, (2) a pool of Chinese **non-CS nouns**, (3) a pool of Chinese **non-CS words** with various syntactic roles (see Figure 1 for details), and used `googletrans` [10] to translate the words in (2) and (3) into English. For each English word in all three pools, we located the word and its closet Chinese word neighbor in the bilingual vector space, and calculated the distance and the cosine similarity between the pair.

**Result** **CS nouns vs. non-CS words:** For both corpora, we used the English translations of non-CS words (pool 3) to bootstrap the 95% confidence interval of the mean distance and mean cosine similarity between the English word and its closet Chinese word neighbor (density plotted in Fig 2). The mean values of the CS nouns (pool 1) from each corpus (the red dots) are well outside of their corresponding interval. **CS nouns vs. non-CS nouns:** Paired t-tests were conducted between the CS vs. non-CS noun pairs (e.g. “balcony” and “countertop” in Fig 1). As some CS nouns appear multiple times in one corpus, resulting in multiple matching non-CS nouns, 5 samples were randomly selected for the test to make sure that the result is replicable. For both corpora, between the CS nouns and their closest Chinese word neighbor, the mean distance is significantly larger than that of non-CS nouns; the mean cosine similarity is significantly smaller (Table 1).

**Conclusion** In this paper, we hypothesize that people CS when it is hard to express their intended meaning accurately in the source language. Our comparison between the CS and non-CS words shows evidence supporting the hypothesis: the CS nouns in English are significantly farther away from their closest Chinese neighbors, with a significantly smaller cosine similarity. This suggests that it is harder to express the meaning of the English CS word using any Chinese words. CS is therefore providing a higher utility for achieving the communication goal.

code-switch sentence:
客厅 还 有 一 个 小 的 balcony。
The living room has a small balcony.

matching sentence:
厨房 面积 大 ， 还 有 一 个 小 的 吧台。
The kitchen is big, and has a small countertop.

word(s) going to word pool 1:
balcony

word(s) going to word pool 2:
吧台

word(s) going to word pool 3:
厨房, 面积, 大, 还, 有, 一, 个, 小, 的, 吧台

**Figure 1:** Left: an example of the sentence pairs in the corpora. Each CS sentence is paired with a monolingual Chinese sentence that has a similar syntactic structure [3]: the matching sentence has a word with the same POS tag as the CS word (in orange), but is not CSed. Right: an illustration of words included into word pool 1, 2, and 3. Note that only single-worded CS nouns are included into word pool 1. For word pool 2 and 3, only single-worded English translations of the non-CS words are kept.

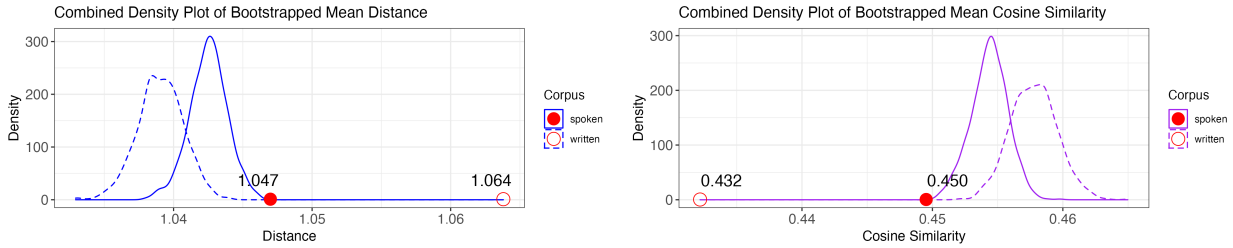


**Figure 2:** Density plots of the mean distance (left) and mean cosine similarity (right) bootstrapped from the non-CS words from the two corpora. The solid line indicates data from the spoken corpus (n=2181); the dashed line indicates data from the written corpus (n=1425). The red dots represent the mean values of the CS nouns, with solid dots for the spoken corpus and the hallow dots for the written corpus.

| 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 | 1.849e-05 | 0.468 | -4.422 | 1.715e-05 |
| | non-CS 2 | 1.030 | 4.335 | 2.456e-05 | 0.468 | -4.368 | 2.147e-05 |
| | non-CS 3 | 1.028 | 4.698 | 5.307e-06 | 0.469 | -4.526 | 1.109e-05 |
| | non-CS 4 | 1.029 | 4.559 | 9.614e-06 | 0.468 | -4.556 | 9.753e-06 |
| | non-CS 5 | 1.030 | 4.443 | 1.567e-05 | 0.468 | -4.454 | 1.500e-05 |
| 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 df=476 for the spoken corpus and df=175 for the written corpus.

**References.** [1] Solorio et al. (2014) *Proc. First Workshop on Computational Approaches to Code Switching*; [2] Myslín & Levy (2015) *Language*; [3] Calvillo et al. (2020) *EMNLP*; [4] Bhattacharya & van Schijndel (2023) *HSP*; [5] Smith et al. (2017) *ICLR*; [6] Conneau et al. (2018) *ICLR*; [7] Joulin et al. (2018) *EMNLP*; [8] Bojanowski et al. (2017) *TACL*; [9] Lovenia et al. (2022) *LREC*; [10] Han, S. (2020). *python package*