# Word Predictability on Points of Code-switching

Billy Gao*, Ariel Chan*, Yanting Li#
*Stanford University #UC Irvine
billygao@stanford.edu arielchan@stanford.edu yantil5@uci.edu

This study examines the key role word predictability plays in driving code-switching (CS) in spoken discourse. Words less predictable in the matrix than the embedded language[1] are more likely to be CSed. This aligns with earlier findings that CSing helps convey meaning more efficiently, reducing cognitive load by favoring more predictable words given context[2-5].

**Introduction** CS occurs when a speaker shifts from the matrix language to the embedded language during communication[6]. Previous research highlighted several linguistic factors influencing CS points, including word length[2-3], predictability[2-3], intended meanings[4], and more. In this study, we examine CSing in Cantonese-English conversational data through computational modeling. We hypothesized that speakers are more likely to CS words that are less predictable in the matrix language than in the embedded language.

**Method** The spoken data was taken from thirteen 30-minute interviews with Cantonese-English bilinguals from homeland (Hong Kong)[8]. Sentences were extracted from the raw recordings and processed[9] Table 1 shows our precedure of data processing: for each sentence, the matrix language was determined to be the language with more words. The full sentences, along with the list of CS words, were machine translated into the matrix and embedded languages[10]. These translations were further processed into context-token pairs, from which the predictabilities of the CS words were calculated using pretrained language transformer models. Specifically, we used a multilingual fill-mask model trained on 2.5TB of filtered CommonCrawl data spanning 100 languages[7]. This model supports multilingual computation, including both English and Cantonese. We processed existing speech data, converting it into contextual texts with the CSed words masked. The predictability of these words was determined by the probability of the masked token being the CSed word. Higher predictability suggests more frequent usage of that word in the given language context.

Analysis was then conducted to compare the CS word's predictability given context in the embedded language versus the translated CS word's predictability given context in the matrix language to verify our hypothesis.

**Result** **Low Predictability CSing:** The results reveal a clear trend that the predictability of CS words in the embedded language within an embedded language context is significantly higher than in the matrix language within its original context (see Fig 1). This is confirmed by a paired t-test ($t = 4.12$, $p < .001$). Additionally, the higher predictability of the CS word in the embedded language context implies more frequent usage in the specific context, which may lead to lower retrieval costs and reduced cognitive load, making it easier for production. **Language Symmetry:** We did not find evidence showing that the switch direction changes the effect of word predictability on CS. Instead, the effect is symmetrical regardless of whether English or Cantonese serves as the matrix or embedded language.

**Conclusion** This paper demonstrates that CS occurs between English and Cantonese when the word has lower predictability in the matrix language compared to the embedded language. This finding adds one more piece of evidence showing that CS behavior is affected by predictability. Bilinguals may be using CS as a tool to reduce cognitive effort and facilitate more efficient communication during production.

| | |
|---|---|
| (1) Original sentence | 但係如果係international school好似我咁就未必得囉 |
| (2) Matrix language<br>Embedded language | Cantonese<br>English |
| (3) Monolingual sentence in the matrix language | 但係如果係國際學校好似我咁就未必得 |
| (4) Monolingual sentence in the embedded language | But if it's an international school like mine it might not necessarily be possible |
| (5) Predictability of CS words in the embedded language | $p(\text{international} \mid \text{But if it's an}) = 0.000578$,<br>$p(\text{school} \mid \text{But if it's an international}) = 0.001126$ |
| (6) Predictability of translation of CS words in the matrix language | $p(\text{國際} \mid \text{但係如果係}) = 5.873744e^{-7}$,<br>$p(\text{學校} \mid \text{但係如果係國際}) = 4.795434e^{-6}$ |

**Table 1:** Procedure of data processing. (2) is determined by the length of each language in (1). (3) and (4) are machine translated from (1) by GPT-2. (5) and (6) are calculated with [7].

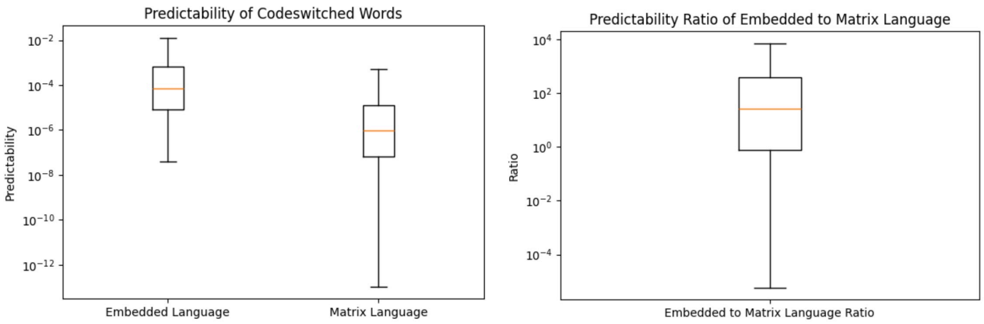

**Figure 1:** Predictability of CS words in both the embedded and matrix languages (left). Predictability ratio of embedded over matrix language (right). The average ratio is $27.18$.

**References.** [1] Myers-Scotton (1993) *Language Arts & Disciplines*; [2] Myslín & Levy (2015) *Language*; [3] Calvillo et al. (2020) *EMNLP*; [4] Li et al. (2024) *SCiL*; [5] Muthusamy et al. (2020) *IJHE*; [6] Solorio et al. (2014) *Proc. First Workshop on Computational Approaches to Code Switching*; [7] Conneau et al. (2019) *CoRR*; [8] Chan (2023) *UCLA*; [9] Lubbers et al. (2021) *python package*; [10] OpenAI (2023) *arXiv*; [11] Bhattacharya & van Schijndel (2024) *arXiv*;