Word Predictability on Code-switching Points in Cantonese–English Discourse

Ariel Chan
University of California, Santa Cruz
arielchan@ucsc.edu

Yanting Li
University of California, Irvine
yantil5@uci.edu

Jacob Poschl
University of California, Santa Cruz
jposchl@ucsc.edu

Abstract

This paper investigates how word predictability influences code-switching probability. We analyze 1,010 code-switched instances drawn from naturalistic sociolinguistic interviews with 41 Cantonese–English bilinguals across three bilingual groups (homeland, immersed, and heritage). In particular, we examine whether the predictability of switch points, operationalized as surprisal, influences the likelihood of code-switching. Using pretrained transformer-based language models, we estimate surprisal at the switch point under different modeling conditions, including autoregressive and masked models and varying amounts of contextual information. Mixed-effects logistic regression analyses show that less predictable words are more likely to be code-switched. These effects are largely consistent across model types and bilingual groups. Overall, these findings highlight the role of predictability in bilingual speech production and provide new insights into code-switching among bilingual speakers with diverse language experiences.

1 Introduction

Code-switching, the alternation between two or more languages within a conversation, is a prevalent phenomenon among bi/multilinguals (Milroy and Muysken, 1995; Poplack, 1988). It offers unique insights into the interaction between sociolinguistics, cognition, and language processing (Kroff and Dussias, 2023; Stell and Yakpo, 2015). Because of this, code-switching has attracted sustained attention from psycholinguistics, sociolinguistics, cognitive scientists, computational linguistics and more.

This paper zooms in onto intra-sentential code-switches, that is, switches that occur within a sentence, and investigates the factors shaping the choice of the switching point. Specifically, we examine the role of word predictability given context, which has been found in previous corpus studies and behavioral experiments to influence code-switching behaviors (Myslín and Levy, 2015; Calvillo et al., 2020; Bhattacharya and van Schijndel, 2026; de Bruin and Shiron, 2024). However, in prior work, predictability has typically been estimated based solely on the *preceding* context of the code-switched element, which does not fully capture the information available to the *speaker*. Given a communicative goal, speakers are likely to possess at least a partially specified representation of the upcoming message even if it is not fully lexicalized, which may also shape expectations about the current word (Bell et al., 2009; Pluymaekers et al., 2005; Upadhye and Futrell, 2025).

Another dimension missing from previous analyses is the language background of the speakers, which can strongly influence code-switching behaviors, including the frequency and type of switches (Poplack, 1980; Backus, 2003; Gollan and Ferreira, 2009; Kootstra et al., 2012; Blanco-Elorrieta and Caramazza, 2021), and potentially the switching point as well. For instance, less balanced bilinguals may have faster or more reliable lexical access in their dominant language, making words in that language more predictable overall. In contrast, balanced speakers with similar proficiency in both languages may experience greater uncertainty in lexical selection, which could influence the way predictability interacts with code-switching, although the direction and magnitude of this effect may vary. Despite these insights, few studies have systematically examined how word-level predictability interacts with speaker language background to shape intra-sentential code-switching.

Drawing on a naturalistic corpus of Cantonese–English code-switching, this study investigates the effect of word predictability on bilingual code-switching probability. The analysis focuses on single-word switches to English within Cantonese-majority sentences, comparing code-switched and

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 230–243
July 3-4, 2026

non-code-switched words occurring in similar syntactic environments. Word probability is estimated using both autoregressive and bidirectional masked language models and with different context lengths to test whether code-switching is better predicted by models that incorporate information beyond the immediately preceding context. Additionally, to capture variation in bilingual language experience and sociocultural context, we examine three groups of bilingual speakers with distinct acquisition contexts and interactional contexts: homeland bilinguals, immersed bilinguals and heritage bilinguals. Specifically, this study addresses the following three research questions:

- Are code-switched words in Cantonese–English speech less predictable given context?
- Does the effect of predictability on code-switching change depending on how context is defined—for example, when using longer preceding context or when incorporating both preceding and following contexts?
- Does the effect of predictability on code-switching differ across bilingual groups with distinct acquisition and interactional contexts (homeland, immersed, and heritage)?

Because most studies examine a single bilingual population, it remains unclear whether predictability effects vary across bilingual groups with different language experiences. We do not expect a main effect of bilingual group; however, we predict an interaction between bilingual group and surprisal, such that the surprisal effect may differ across groups: stronger for homeland and immersed bilinguals, whose language use more closely aligns with the texts on which the LMs were trained, and smaller and potentially not significant for heritage bilinguals.

2 Background

Cantonese–English bilinguals provide a unique opportunity to examine code-switching, as their conversations often exhibit dynamic and complex switching patterns (Gibbons, 1987; Luke, 1998; Li, 2000). This is partially due to the historical development of Cantonese diasporic communities: Hong Kong has a history of being a British colony (Bolton, 2000; Sung, 2015), and there has been a significant migration of Cantonese speakers to the United States (Chang, 2004). Fig 1 shows an

咁 都 清楚 因為 譬如 藏族人 同 滿族人 就 assimilate 咗
That's clear because, for example, Tibetans and Manchus have assimilated.

Figure 1: Example of intra-sentential Cantonese–English code-switch.

example of a code-switched sentence where the majority part of the sentence is in Cantonese, with a single-word switch to English.

Why bilinguals code-switch has always been an interesting question. Various factors have been identified in the literature, including sociolinguistic considerations such as the identity of the speaker and listener (Chen, 2008; Lo, 1999), register (Halmari and Smith, 1994), indexicality (Myers-Scotton, 2020), discourse-functional purposes such as to emphasize important information (Gumperz, 1982; Myers-Scotton, 1993) and to strengthen discourse cohesiveness by maximizing the salience of discourse markers (De Rooij, 2000), as well as linguistic factors such as word length (Bhattacharya and van Schijndel, 2026; Calvillo et al., 2020; Myslín and Levy, 2015), syntactic roles (Myslín and Levy, 2015; Poplack, 1980), cognate priming (Broersma and De Bot, 2006; Kootstra et al., 2012), lexical gap filling (Grosjean, 1982; Li et al., 2024; Otheguy and García, 1993), and more.

Among these linguistic factors, word predictability has attracted significant attention in recent years, and it is also the main focus of this study. It has been found across languages that code-switched words tend to be less predictable given context. Using a corpus of naturalistic Czech–English bilingual conversation, Myslín and Levy (2015) designed a cloze test for sentence-final words that were either code-switched or not. Bilinguals listened to each sentence up to the final word and guessed the final word in either language. The results revealed that code-switched words were harder to guess, suggesting a lower predictability.

In another study, Calvillo et al. (2020) examined the written communication of Mandarin–English bilinguals and found the same effect. They estimated word probability by calculating surprisal, defined as the negative logarithmic probability of a word given its context (Hale, 2001, 2016; Levy, 2008), using a 5-gram model. Bhattacharya and van Schijndel (2026) replicated these findings using large language models (LLMs) and further extended them from written data to a speech corpus of the same language pair, providing cross-model evidence of robustness of the surprisal effect.

In addition to corpora studies, another piece of evidence comes from de Bruin and Shiron (2024)'s Bulgarian–English production experiment. In the experiment, participants were presented with a carrier sentence followed by a picture, and were instructed to read the sentence out loud and complete it by naming the picture in either language. The carrier sentences either made the pictures more predictable (e.g. *She packs the* + bag) or not (*They invite the* + singer). The results showed that code-switching was more likely to happen when the pictures were less predictable given the contexts.

Evidence from both experimental and computational studies have suggested a robust effect of predictability on code-switching. To account for this effect, Myslín and Levy (2015) proposed that this is to help alert the audience, which is supported by Bhattacharya and van Schijndel (2026). Knowing that an upcoming word is unpredictable, the speaker switches to another language, and such a salient language change will invite the listener to pay more attention to the content, making comprehension easier. Calvillo et al. (2020) on the other hand, offered explanations from a speaker-oriented angle: code-switching occurs at less predictable words because the increased cognitive demand to retrieve such words leaves fewer resources for inhibiting the non-target language, resulting in code-switching. de Bruin and Shiron (2024)'s explanation is also more along the line of this speaker-oriented proposal, as they found that predictive carrier sentences lowered code-switching frequency compared to no-context picture naming, but unpredictive carrier sentences did not necessarily increase code-switching frequency. The authors suggested that when contextual cues make an upcoming word more predictable, the co-activation of its translation equivalents in the other language is reduced, leading to fewer code-switches.

In all studies above, predictability is estimated using preceding context.[1] However, to understand why *speakers* choose to code-switch at a certain point, it is crucial to consider the future context in addition to the preceding context. Unlike comprehenders who process sentences incrementally, speakers have access to at least partial future message even if it is not fully planned, and the upcoming content may influence production at the current time point, from acoustic realization to word choice (Bell et al., 2009; Ranjan et al., 2022; Pluymaekers et al., 2005; Zhan and Levy, 2019; Upadhye and Futrell, 2025). In the context of code-switching, this forward-looking planning is particularly important: a speaker may anticipate that an upcoming word or syntactic structure is more easily accessed in one language than the other, or that a switch aligns better with the overall sentence plan. As a result, the decision to switch languages at a certain word can be shaped not only by the preceding context but also by the planned upcoming elements. Incorporating future context may therefore help better capture the planning processes underlying code-switching behavior.

3 Method

To answer our research questions, we need words that appear in comparable syntactic environments, with some instances involving a code-switch and others remaining unilingual. We then estimate and compare their predictabilities using pre-trained LMs. The procedures are described below.

3.1 Corpus

We use the CANGLISH Bilingual Corpus (Chan, 2023), which includes conversational speech from 41 Cantonese–English bilinguals across three bilingual groups: 13 homeland bilinguals raised in a Cantonese-dominant environment and educated in English in Hong Kong, 14 heritage bilinguals raised in Cantonese-speaking households in the U.S., and 14 immersed bilinguals from Hong Kong who later moved to the U.S. All participants acquired Cantonese as their first language and English as their second.

Data were collected online via Zoom between September and December, 2022, containing 36 hours and 17 minutes of audio and video recordings. Each participant completed two tasks: (1) a map task in which the participant and experimenter took turns describing slides to match each other's and (2) a sociolinguistic interview exploring language use and identity between the participant and experimenter. Code-switching was naturally introduced by the experimenter with language switches approximately every 10 minutes. Participants were informed that they could use either Cantonese, English, or both during the conversation.

[1] In de Bruin and Shiron (2024), the code-switching point always occurred at the end of a trial, so there was no following context. For the other studies (Myslín and Levy, 2015; Calvillo et al., 2020; Bhattacharya and van Schijndel, 2026), code-switches could be followed by additional content, either within the same sentence or within the same conversation.

3.2 Preprocessing

The recorded interviews were transcribed and annotated using the ELAN software (Version 6.8) (Wittenburg et al., 2006) manually by highly proficient Cantonese–English bilinguals, with sensitive or identifying information masked. Transcriptions were extracted from 41 ELAN `.eaf` files using the `pympi` (Lubbers and Torreira, 2013-2025) library. Sentence boundaries were taken directly from human annotations, and each annotated span was treated as a sentence without further automatic segmentation. After sentence segmentation, a cleaning pipeilne was applied that removed anonymization/annotation markers, normalized whitespace, and removed punctuation tokens and filler words (e.g., eh, oh, 呃, 嗯, see Appendix A for a full list).

Sentences containing fewer than two words after filtering were excluded. Cantonese and English tokens were identified using Unicode script detection (CJK vs. ASCII alphabetic characters). Cantonese text was tokenized with `pycantonese` (Lee et al., 2022) while English text was tokenized using whitespace segmentation. Full preprocessing/tokenization details are provided in Appendix A.

Sentences containing both Cantonese and English tokens were labeled as code-switched. A code-switched sentence's majority language was determined by word counts. Sentences with an equal word count for English and Cantonese were labeled "Equal". For the present analysis, we restricted our dataset to code-switched sentences with Cantonese as the majority language because switches from Cantonese to English represent the unmarked and most common switch direction both in our corpus and among Cantonese–English bilingual communities (Chan et al., 1983). We also made sure that at least two Cantonese words preceded the switching point to avoid sentence-initial switches and short fragments where predictability estimates were unreliable. One heritage participant did not produce any code-switched sentences that met these criteria and was therefore not included in the analysis.

3.3 Identifying Syntactically Matched Sentence Pairs

For each code-switched sentence, we translated the English span(s) into Cantonese using Meta's NLLB model (Costa-Jussà et al., 2022), and replaced the English span(s) with their Cantonese translations, producing a Cantonese-only sentence (see Appendix Table 3 for more deteails). The resulting Cantonese sentences were then POS-tagged using PyCantonese (Lee et al., 2022).

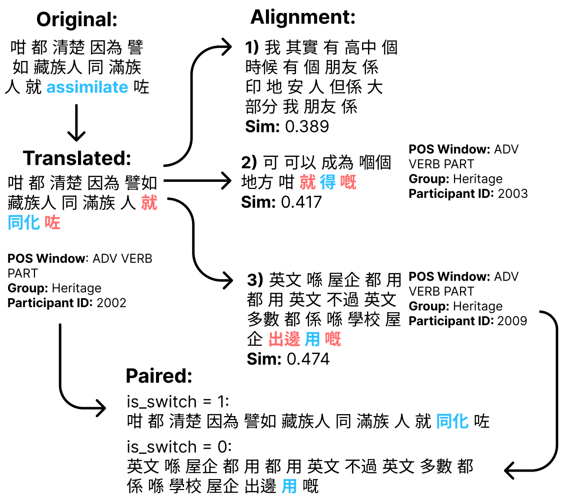

Figure 2: Demonstration of the matching procedure for a code-switched instance. The original sentence is translated into Cantonese and POS-tagged. Candidate 1 fails the similarity criterion; Candidates 2 and 3 pass and are ranked by match preference, yielding Candidate 3 as the selected match. The switch point (blue) is the index for surprisal calculation.

Figure 2 illustrates the procedure for identifying a syntactically similar unilingual Cantonese sentence. Following Calvillo et al. (2020), candidate sentences were required to satisfy two criteria: (i) the unilingual sentence achieved a Levenshtein similarity ≥ 0.4 with the translated code-switched sentence and (ii) the unilingual sentence contained the same POS-tag window of size three centered on the switch point (truncated at sentence boundaries). When multiple candidates satisfied these constraints, we selected a match using a deterministic ranking procedure. Same-speaker matches were preferred over matches from other speakers within the same bilingual group to control for speaker-specific lexical preferences and habitual code-switching patterns. When multiple same-speaker candidates were available, we selected the sentence closest in time to the code-switched sentence, under the assumption that temporally proximate utterances are more likely to share discourse topic and contextual constraints. Remaining ties were resolved by selecting the candidate with the highest similarity score. Using this procedure, 1,010 of the 1,250 code-switched sentences were successfully paired with a unilingual Cantonese

	Autoregressive LM	**Masked LM**
Code-switched sentence	p(同化 | c_{pre})	p(同化 | c_{pre}, [MASK], c_{post})
Matching sentence	p(用 | c_{pre})	p(用 | c_{pre}, [MASK], c_{post})

Table 1: Target words and contexts used for surprisal estimation without discourse context, illustrated using the example in Fig 2. c_{pre} and c_{post} denote the preceding and following context relative to the target word. The autoregressive model conditions only on c_{pre}; the masked model additionally incorporates c_{post} via masked prediction.

sentence, with a mean Levenshtein similarity of 0.464 between pairs (SD = 0.077).

3.4 Language Model Surprisal Estimation

Using our example in Fig 2, in this step, we calculate $p(\text{同化}|context)$ for the code-switched sentence, and $p(\text{用}|context)$ for the matching sentence for predicting whether a switch would happen. Next, we introduce the LMs used, as well as how contexts are defined for each model.

The two pretrained transformer LMs used are:

- `hon9kon9ize/bert-large-cantonese` (hon9kon9ize Group, 2024a): a masked LM
- `hon9kon9ize/CantoneseLLMChat-v1.0-7B` (hon9kon9ize Group, 2024b): an autoregressive LM

Table 1 illustrates the target word and contexts used for each model. For each sentence, both models were conditioned on the discourse context preceding the switch point. The masked model additionally received available post-switch content (i.e., words following the switch point) to allow bidirectional conditioning, whereas the autoregressive model only used the preceding context by design. Token-level probabilities at the first code-switched word were used to compute surprisal.

Let $w = w_{1:K}$ denote the target word and c the discourse context. Token surprisal is defined as:

$$S_k = -\log_2 P(w_k \mid c, k) \tag{1}$$

Word-level surprisal is the sum of token surprisals:

$$S(w \mid c) = \sum_{k=1}^{K} S_k \tag{2}$$

Token probabilities $P(\cdot \mid c, k)$ were obtained using left-to-right conditionals for the autoregressive model and a masking-based pseudo-log-likelihood procedure for the masked model (see Appendix C).

We also examined two context conditions: no preceding discourse context ($n = 0$, sentence-internal context only) and three preceding sentences ($n = 3$). In the $n = 3$ condition, any English spans in the preceding sentences were translated to Cantonese using the same NLLB procedure described in Section 3.3, ensuring that discourse context remained in-distribution for the Cantonese LMs. Instances where translation quality was insufficient were excluded (see Appendix C).

3.5 Regression

We tested whether LM-estimated surprisal at the switch point predicted code-switching using a generalized linear mixed-effects logistic regression model. All analyses were conducted in R (R Core Team, 2024) using the `lme4` package (Bates et al., 2015) within RStudio (Posit team, 2025). Each observation corresponded to the word at the potential switch point, which is either the translation of the English word in the code-switched sentence or the matched word in the unilingual matching sentence (e.g. 同化 or 用 in the example in Fig 2).

The dependent variable for our mixed-effect logistic regression models was binary: `is_switch` $= 1$ for code-switched instances and `is_switch` $= 0$ for matched unilingual instances. Predictor variables included word length, part of speech (POS), surprisal, bilingual group (homeland, immersed, heritage), and the interaction between surprisal and bilingual group. Word length was calculated as the number of Chinese characters in the segmented Cantonese target word.

$$\begin{aligned}\eta_{ij} &= \beta_0 + \beta_1 \text{WordLength}_{ij} + \beta_2 \text{POS}_{ij} \\ &\quad + \beta_3 \text{BilingualGroup}_{ij} + \beta_4 \text{Surprisal}_{ij} \\ &\quad + \beta_5 \text{BilingualGroup}_{ij} \text{Surprisal}_{ij} + u_{0j}\end{aligned} \tag{3}$$

where $\eta_{ij} = \text{logit}\big(P(\text{is\_switch}_{ij} = 1)\big)$ and $u_{0j} \sim \mathcal{N}(0, \sigma_u^2)$.

We fit four models: (1) two models using sur-

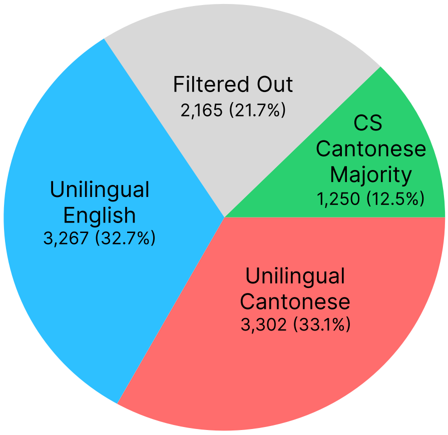

Figure 3: Distribution of sentences by language composition from a total of 9,984 initially processed sentences. The filtered-out section (grey) includes filler-only sentences (n=616), sentences below the minimum word threshold (n=990), code-switched sentences where English or neither language was the majority (n=230), and code-switched Cantonese-majority sentences excluded by the pattern filter or failed translation (n=329).

prisal estimated by the autoregressive LM with (1a) no preceding discourse context and (1b) three preceding sentences in the same discourse, and (2) two models using surprisal estimated by the masked LM with (2a) no preceding discourse context and (2b) three preceding sentences in the same discourse.

POS tags were collapsed into three categories: noun, verb, other.[2] Proper nouns were excluded, and analyses were restricted to single-worded switches. Categorical predictors (bilingual group and POS) were sum-coded using sum-to-zero contrasts, so coefficients represent deviations from the grand mean. Models also included a random intercept for participant to allow baseline switching probabilities to vary across speakers. Full model formulas are provided in Eq. 3.5.

4 Results

4.1 Descriptive Summary

The final dataset comprised 1,010 matched sentence pairs drawn from 1,250 code-switched sentences (Figure 3). Among these, 40.2% were contributed by the homeland group, 31.5% by the immersed group, and 28.3% by the heritage group. The two regression models without preceding discourse context (1a, 2a) each included 1,348 observations, whereas the two with three sentences of preceding context (1b, 2b) each included 1,346 observations, since two observations were at the beginning of the interview (i.e. without any preceiding discourse contexts).

[2]Tokens with unavailable or unclassifiable POS tags (e.g., returned as UNK or missing by the tagger) were excluded prior to regression analyses.

4.2 Overall Effect of Predictability on Code-Switching

As specified above, we fit four models that varied in the type of language model (autoregressive vs. masked) and the amount of preceding discourse context (no prior discourse vs. three preceding sentences). Table 2 summarizes the results of all four models, including the direction of effect and significance level of each predictor. Detailed results for each model are reported in Appendix D. Across Models (1a), (2a), and (2b), surprisal showed a significant positive main effect, such that higher surprisal values were associated with a greater likelihood of code-switching. However in Model (1b), surprisal showed an opposite effect. Overall, the positive relationship between surprisal and code-switching is robust across most models, suggesting that lower word predictability generally increases the likelihood of switching.

To directly compare the predictive contributions of autoregressive and masked surprisal, we fit models containing autoregressive surprisal only, masked surprisal only, and both predictors simultaneously using three preceding sentences as discourse context. Likelihood ratio tests showed that adding masked surprisal to the autoregressive model significantly improved model fit, $\chi^2(3) = 33.62$, $p < .001$. In contrast, adding autoregressive surprisal to the masked model resulted in only a marginal improvement in fit, $\chi^2(3) = 7.43$, $p = .059$. Comparisons of model fit indices further supported this pattern: the masked-only model (AIC = 1703.27, BIC = 1755.32) substantially outperformed the autoregressive-only model (AIC = 1729.46, BIC = 1781.50), whereas the full model showed only a minimal improvement in AIC (1701.84) and a higher BIC (1769.50) due to increased model complexity. Together, these findings suggest that masked surprisal captures substantially more variance than autoregressive surprisal, although autoregressive surprisal may still contribute limited additional information when both measures are considered simultaneously.

Model Type	**Autoregressive** (1a)	**Autoregressive** (1b)	**Masked** (2a)	**Masked** (2b)
Model Context	Sentence-internal	Sentence-internal + 3 Preceding Sentences	Sentence-internal	Sentence-internal + 3 Preceding Sentences
Word Length	$+^{***}$	$+^{***}$	$+^{***}$	$+^{***}$
POS: Noun	$-^{**}$	$-^{*}$	$-^{**}$	$-^{*}$
POS: Verb	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Surprisal	$+^{***}$	$-^{**}$	$+^{***}$	$+^{***}$
Bilingual Group: Heritage	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Bilingual Group: Immersed	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Surprisal × Heritage	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Surprisal × Immersed	*n.s.*	*n.s.*	*n.s.*	*n.s.*

Table 2: Summary of predictor significance across the four mixed-effects logistic regression models. Signs ($+/-$) indicate the direction of the log-odds estimate. Significance levels are marked as $*p < .05$, $**p < .01$, and $***p < .001$. Not-significant is labeled as *n.s.*.

4.3 Effects of Word Length, POS, and Bilingual Group

The main effects of word length, POS, and bilingual group were consistent across all four models. For ease of discussion, we report results from Model (2b). The interaction between surprisal and the heritage bilingual group showed a marginal trend (β = 0.03, SE = 0.02, z = 1.75, p = .080), suggesting a potential difference in the strength of the surprisal effect across groups, although this effect did not reach statistical significance. No significant interaction was observed for the immersed bilingual group (β = -0.01, SE = 0.01, z = -0.55, p = .583). There was also no significant main effect of bilingual group: neither heritage speakers (β = -0.19, SE = 0.20, z =-0.96, p = .339) nor immersed speakers (β = -0.02, SE = 0.19, z = 0.08, p = .937) differed significantly from the grand mean switching probability.

Word length had a significant positive effect on switching (β = 0.98, SE = 0.10, z = 9.58, p < .001), indicating that longer words were more likely to be switched. POS also influenced switching: the contrast for nouns was negative and significant (β = -0.21, SE = 0.08, z =-2.47, p = .013), indicating that nouns were less likely than the overall POS mean to be switched. The contrast for verbs was not significant (β = 0.06, SE = 0.10, z =0.64, p = .520).

5 Discussion

Previous work has shown that less predictable words are more likely to be code-switched, and this study is, to our knowledge, the first to extend this line of investigation to Cantonese–English bilingual speech, a relatively under-resourced language pair. In addition, we explored whether the predictability effect remained consistent across different types of LMs, different amounts of contextual information, and different bilingual groups. This comparison is novel and exploratory, as many previous studies focus on a single bilingual population from a relatively homogenous language context.

5.1 Evidence for a Predictability Effect in Cantonese–English Code-Switching

We first replicated the analysis of Calvillo et al. (2020) with the Cantonese–English dataset. Sentences containing a single-word switch from Cantonese to English were matched with unilingual Cantonese sentences with similar syntactic structure. After translating the switched word into Cantonese, we used an autoregressive model to estimate the word predictability at the switch (or matching) point given sentence-internal preceding context, and tested whether it predicted code-switching (Modal (1a)). The result was consistent with previous findings from other language pairs: less predictable words were more likely to be code-switched (Myslín and Levy, 2015; Calvillo et al., 2020; Bhattacharya and van Schijndel, 2026; de Bruin and Shiron, 2024). Additionally, we observed a word length effect consistent with the literature, with longer words being more likely to be code-switched (Myslín and Levy, 2015; Calvillo et al., 2020; Bhattacharya and van Schijndel, 2026). Meanwhile, nouns were less likely to be switched, which is a pattern that contrasts with the common prediction in the literature (Kroff, 2016; Chan, 1999), but it is consistent with the findings of Calvillo et al.. One possible explanation is that code-switched nouns can introduce new topics or discourse-salient referents (Barredo, 1997). Such nouns are therefore typically less predictable. Because predictability was included as a predictor in our model, it may have captured some of the variance that would otherwise be attributed to part of speech, thereby reducing the independent effect of POS.

5.2 Robustness of the Predictability Effect across Language Models and Context Windows

To test the robustness of the predictability effect, we estimated surprisals using two types of models: an autoregressive model, which relies solely on preceding context, and a masked model, which allows bidirectional prediction on both preceding and following context. We also varied the amount of discourse information included for each model. When only the current sentence was used for surprisal estimation, the predictability effect was consistent across both models. However, when the preceding context was expanded to include the preceding three sentences in the same conversation, the effect reversed for the autoregressive model while remaining unchanged for the masked model.

Examining the correlation between surprisal estimates from no prior vs. extended contexts, we noticed a Pearson correlation of 0.78 for the masked model but only 0.11 for the autoregressive model, which may help explain the reversal in the surprisal effect. The low correlation for the autoregressive model may be due to the fact that all English words in the preceding context had to be translated to Cantonese, as the model was trained exclusively on Cantonese text. These translations may have introduced some unnaturalness that affected the surprisal estimates. In contrast, the masked model had access to both preceding and following context, which likely stabilized its predictions.

When autoregressive and masked surprisal were entered into the same model, masked surprisal remained a significant positive predictor of code-switching, while autoregressive surprisal exhibited a weaker negative association after controlling for masked surprisal. Likelihood ratio tests showed a consistent result: adding masked surprisal to the autoregressive model significantly improved model fit, while adding autoregressive surprisal to the masked model resulted only in a marginal improvement. Comparisons of AIC and BIC values similarly indicated that the masked model estimation substantially outperformed the autoregressive model estimation. Together, these findings suggest that the bidirectional masked language model provided a better prediction for code-switching than the autoregessive model that rely only on the preceding context. This result is consistent with the idea that bilingual speakers may make word choice decisions not only based on incremental context information, but also on expectations about the future, which are better captured by the bidirectional model. At the same time, the marginal contribution of autoregressive surprisal suggests that sequential prediction may still capture partially distinct aspects of bilingual predictive processing during code-switching.

5.3 Absence of a Bilingual Group Effect on Code-Switching Probability

Consistent with our expectation, we found no significant main effect of bilingual group. Contrary to our prediction, however, we also found no interaction between bilingual group and surprisal. This suggests the effect of predictability on switching probability were similar across heritage, immersed, and homeland bilinguals in our dataset. Although somewhat surprising, this pattern is consistent with findings from another study on comprehension of code-switched words in sentences (Chan, 2023). Together, these results suggest that bilingual speakers with diverse language acquisition backgrounds and interactional contexts may nevertheless exhibit similar patterns of code-switching. This provides a different perspective from studies that emphasize the role of bilinguals' language dominance or proficiency in shaping code-switching behavior (for a review, see Declerck and Philipp 2015).

One possible explanation is that cognitive mechanisms underlying lexical predictability and language selection operate similarly across bilingual speakers despite differences in interactional contexts. From a language control perspective (Green and Wei, 2014), speakers across groups may be similarly sensitive to predictability pressures when deciding whether to perform insertional lexical switches during language production. Alternatively, bilingual experience may primarily influence other aspects of code-switching (Gollan and Goldrick, 2018), such as language preference or discourse functions, rather than probability of switching at a particular lexical position. Our analysis also focused only on unidirectional lexical switches from Cantonese to English, the unmarked and default direction of switching in our corpus, which may further reduce the likelihood of observing an interaction between bilingual group and surprisal. Finally, it is also possible that the absence of group effect reflects a selection bias: heritage bilinguals who agreed to participate in the study may represent a subgroup with relatively high self-rated Cantonese proficiency. It is therefore not surprising that they exhibited code-switching patterns similar to those of the other two groups.

6 Conclusion and Future Directions

This study provides evidence that word predictability plays a role in shaping intra-sentential single-word code-switches. Our findings indicated that less predictable words were more likely to be switched. This effect was largely consistent across different LMs, context windows, and bilingual groups, suggesting that predictability constraints operate similarly across speakers with diverse interactional contexts. However, several limitations should be noted. Our analysis focused on unidirectional lexical switches from Cantonese to English and did not consider additional switches in the surrounding discourse. Future work could examine English to Cantonese switches to see whether predictability effects generalize across switch directions.

Acknowledgments

We thank the research assistants in the Bilingualism, Mind, and Brain Lab at the University of California, Irvine, and the Bilingualism, Sociolinguistics, and Cognition (BiSoN) Lab at Stanford University and University of California, Santa Cruz for their assistance with this project. We also thank the reviewers and attendees of the 2024 California Meeting on Psycholinguistics (CAMP), the 38th Annual Conference on Human Sentence Processing (HSP), and Marten van Schijndel for their feedback and discussions. Data collection for this project was supported by a UCLA Dissertation Year Fellowship, a National Science Foundation Linguistics DDRI Grant (Award No. BCS-20424), and a Mellon/ACLS Dissertation Completion Fellowship awarded to Ariel Chan. This research was additionally supported by start-up funding from the Humanities Division at the University of California, Santa Cruz awarded to Ariel Chan, Assistant Professor of Languages and Applied Linguistics and Principal Investigator of the BiSoN Lab.

References

Ad Backus. 2003. Units in code switching: Evidence for multimorphemic elements in the lexicon. *Linguistics*, 41(1).

Inma Munoa Barredo. 1997. Pragmatic functions of code-switching among Basque–Spanish bilinguals. In *Proceedings of Actas do i Simposio Internacional sobre o Bilingismo*, pages 528–541.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48.

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Debasmita Bhattacharya and Marten van Schijndel. 2026. Code-switching in text and speech challenges information-theoretic speaker design. *Language and Cognition*, 18:e31.

Esti Blanco-Elorrieta and Alfonso Caramazza. 2021. A common selection mechanism at each linguistic level in bilingual and monolingual language production. *Cognition*, 213:104625.

Kingsley Bolton. 2000. The sociolinguistics of Hong Kong and the space for Hong Kong English. *World Englishes*, 19(3):265–285.

Mirjam Broersma and Kees De Bot. 2006. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and Cognition*, 9(1):1–13.

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.

Ariel Shuk Ling Chan. 2023. *The Diaspora of Bilinguals: Code-Switching in Three Groups of Cantonese-English Bilinguals*. University of California, Los Angeles.

Brian Hok-Shing Chan. 1999. *Aspects of the syntax, production and pragmatics of code-switching-with special reference to Cantonese–English*. University of London, University College London (United Kingdom).

Mun-Chee Chan, Helen LH Chau, and Rumjahn Hoosain. 1983. Input/output switch in bilingual code switching. *Journal of Psycholinguistic Research*, 12(4):407–416.

Iris Chang. 2004. *The Chinese in America: A narrative history*. Penguin.

Katherine Chen. 2008. Positioning and repositioning: Linguistic practices and identity negotiation of overseas returning bilinguals in Hong Kong. *Multilingua*, 27(1/2):57.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Angela de Bruin and Veniamin Shiron. 2024. Putting language switching in context: Effects of sentence context and interlocutors on bilingual switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 50(7):1112.

Vincent A De Rooij. 2000. French discourse markers in Shaba Swahili conversations. *International Journal of Bilingualism*, 4(4):447–466.

Mathieu Declerck and Andrea M Philipp. 2015. A review of control processes and their locus in language switching. *Psychonomic Bulletin & Review*, 22(6):1630–1645.

John Gibbons. 1987. *Code-mixing and code choice: A Hong Kong case study*. Multilingual Matters.

Tamar H Gollan and Victor S Ferreira. 2009. Should I stay or should I switch? A cost–benefit analysis of voluntary language switching in young and aging bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):640.

Tamar H Gollan and Matthew Goldrick. 2018. A switch is not a switch: Syntactically-driven bilingual language control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(1):143.

David W Green and Li Wei. 2014. A control process model of code-switching. *Language, Cognition and Neuroscience*, 29(4):499–511.

François Grosjean. 1982. Code-switching and borrowing. In *Bilingual: Life and Reality*, pages 51–62. Harvard University Press.

John J. Gumperz. 1982. *Discourse Strategies*. Cambridge University Press, Cambridge, UK.

J. 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.

Helena Halmari and Wendy Smith. 1994. Code-switching and register shift: Evidence from Finnish-English child bilingual conversation. *Journal of Pragmatics*, 21(4):427–445.

hon9kon9ize Group. 2024a. bert-large-cantonese. `https://huggingface.co/hon9kon9ize/bert-large-cantonese`. Accessed: 2026-03-02.

hon9kon9ize Group. 2024b. Cantonesellmchat-v1.0-7b. `https://huggingface.co/hon9kon9ize/CantoneseLLMChat-v1.0-7B`. Accessed: 2026-03-02.

Gerrit Jan Kootstra, Janet G Van Hell, and Ton Dijkstra. 2012. Priming of code-switches in sentences: The role of lexical repetition, cognates, and language proficiency. *Bilingualism: Language and Cognition*, 15(4):797–819.

Jorge R Valdés Kroff. 2016. Mixed NPs in Spanish-English bilingual speech: Using a corpus-based approach to inform models of sentence processing. *Spanish-English Codeswitching in the Caribbean and the US*, 11:281.

Jorge R Valdés Kroff and Paola E Dussias. 2023. Production, processing, and prediction in bilingual codeswitching. In *Psychology of Learning and Motivation*, volume 78, pages 195–237. Elsevier.

Jackson Lee, Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022. PyCantonese: Cantonese linguistics and NLP in python. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6607–6611.

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

David CS Li. 2000. Cantonese-English code-switching research in Hong Kong: A Y2K review. *World Englishes*, 19(3):305–322.

Yanting Li, Gregory Scontras, and Richard Rutrell. 2024. On the communicative utility of code-switching. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 343–349.

Adrienne Lo. 1999. Codeswitching, speech community membership, and the construction of ethnic identity. *Journal of Sociolinguistics*, 3(4):461–479.

Mart Lubbers and Francisco Torreira. 2013-2025. pympi-ling: a Python module for processing ELANs EAF and Praats TextGrid annotation files. `https://pypi.python.org/pypi/pympi-ling`. Version 1.71.

Kang-Kwong Luke. 1998. Why two languages might be better than one: Motivations of language mixing in Hong Kong. In Martha C. Pennington, editor, *Language in Hong Kong at Century's End*, pages 145–159. Hong Kong University Press Hong Kong.

Lesley Milroy and Pieter Muysken. 1995. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press.

Carol Myers-Scotton. 1993. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

Carol Myers-Scotton. 2020. Code-switching as indexical of social negotiations. In *The Bilingualism Reader*, pages 97–122. Routledge.

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

Ricardo Otheguy and Ofelia García. 1993. Convergent conceptualizations as predictors of degree of contact in us spanish. *Spanish in the United States: Linguistic contact and diversity*, pages 135–154.

Mark Pluymaekers, Mirjam Ernestus, and R Baayen. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, 62(2-4):146–159.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. *Linguistics*, 18:581–618.

Shana Poplack. 1988. Contrasting patterns of code-switching in two communities. In Monica Heller, editor, *Codeswitching : Anthropological and Sociolinguistic Perspectives*, pages 215–244. De Gruyter, Inc.

Posit team. 2025. *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2022. Linguistic complexity and planning effects on word duration in hindi read aloud speech. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 119–132.

Gerald Stell and Kofi Yakpo, editors. 2015. *Code-switching between structural and sociolinguistic perspectives*. De Gruyter.

Chit Cheung Matthew Sung. 2015. Hong Kong English: linguistic and sociolinguistic perspectives. *Language and Linguistics Compass*, 9(6):256–270.

Shiva Upadhye and Richard Futrell. 2025. Back to the future: The role of past and future context predictability in incremental language production. *arXiv preprint arXiv:2511.07752*.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Meilin Zhan and Roger Levy. 2019. Availability-based production predicts speakers' real-time choices of mandarin classifiers. *arXiv preprint arXiv:1905.07321*.

A Preprocessing Details

A.1 Implementation:

Raw transcriptions were stored in ELAN Annotation Format (EAF) files, one per participant. Each file was loaded using the pympi library (Lubbers and Torreira, 2013-2025), in which we can extract the relevant per-tier annotation data as `(start_time, end_time, text)` tuples. The participant tier was identified by its name prefix (ACHE, ACI, or ACH), and the interviewer tier (included for discourse context) was identified by the tier title: IR. Speaker group membership was assigned based on tier name prefix:

- ACHE → Heritage
- ACI → Immersed
- ACH → Homeland

A.2 Tokenization

Cantonese and English tokens were identified using Unicode script detection. Each annotation was first segmented into contiguous script runs: characters in CJK Unicode ranges were labeled Cantonese and ASCII alphabetic characters were labeled English. Cantonese segments were then tokenized using `pycantonese` (Lee et al., 2022) and English segments were split on whitespace. Some examples of segmentation:

- 我去*school*啦→ 我去(C), school (E), 啦(C)
- 我*local*人→ 我(C), local (E), 人(C)

Tokens were then cleaned by: removing anonymization markers (strings matching `^x+$`), stripping leading and trailing punctuation, normalizing Unicode dashes to ASCII hyphens and splitting on them, splitting on commas and ellipses, and discarding any tokens that were punctuation-only or empty after cleaning.

Filler and hesitation tokens were excluded prior to all analyses. Because matrix language determination relies on word counts in each language, retaining fillers would introduce noise due to the complexity of assigning such fillers to a particular language (e.g., *eh*). Filler detection was applied cross-lingually, so that a hesitation token was excluded regardless of which span it appeared in. The list of excluded tokens are as displayed in Figure 4.

Language	Fillers
English	uh, um, uhm, er, err, ah, eh, mm, hm, hmm, mhm, ehm, uh-huh, mm-hmm, uh-uh, mm-mm, em, emm, ehh, umm, ummm, uhh, mmm, huh, oh, ohh, ohhh, ummmm, uhmm
Cantonese	呃, 嗯, 啊, 哦, 唔, 吖, 哎, 咦

Figure 4: Filler and hesitation tokens excluded prior to all analyses, listed by language.

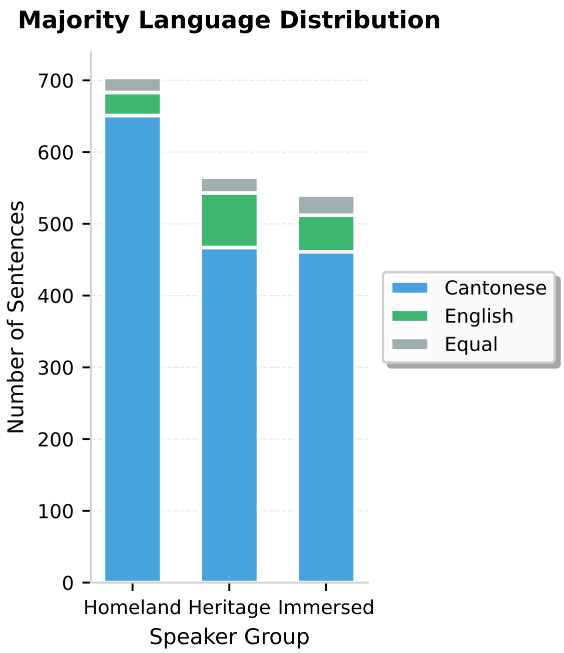

Figure 5: Distribution of sentences by majority language across the three bilingual speaker groups (homeland, heritage, and immersed).

B Sentence Matching and Alignment Examples

B.1 Translation and Truncation

Table 3 provides a concrete example of the transformation from a CS sentence to the counterfactual unilingual baseline used for surprisal scoring. NLLB occasionally fails to translate embedded English segments fully, leaving English tokens or malformed outputs. Rather than discarding these sentences entirely, we applied a targeted truncation procedure. After translation, each output was checked for residual English tokens. If any appeared before the switch point, the sentence was rejected outright. If they appeared only after the switch point (in or beyond the originally embedded English segment) the translation was truncated at the first offending token, provided the switch point itself was preserved. Sentences where truncation would have removed the switch point were also rejected.

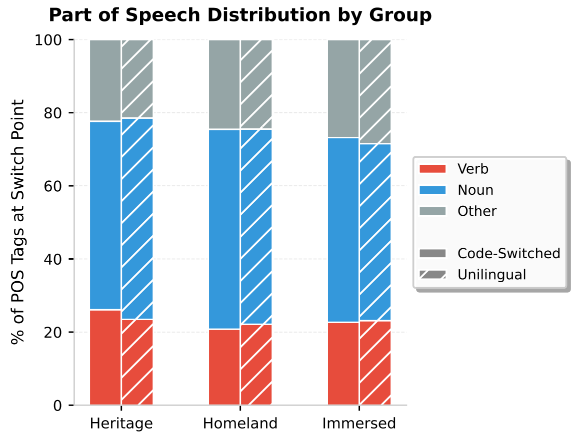

Figure 6: Distribution of POS tags at the switch point by bilingual group, drawn from the full matched dataset prior to model-specific filtering (e.g., exclusion of proper nouns and multi-word switches). POS distributions are broadly comparable across groups.

B.2 Sentence Matching Diagnostics

Table 4 reports diagnostics for the deterministic ranking procedure used to select a unilingual match when multiple candidates satisfied the filtering criteria. Lower ranks indicate higher-priority candidates in the selection process.

C Computational Methodology Details

C.1 Surprisal Formulations

For the masked model, each subword token w_k of the target word was masked in turn and the probability assigned to the original token at the masked position was extracted from the softmax distribution over the full vocabulary.

Stage	Sentence
Original (CS)	但係如果**international school** 好
Invalid Translation	但係如果國際學校**school** 好
Truncated	但係如果國際學校
POS Tags	`CCONJ CCONJ NOUN`

Table 3: The truncation pipeline: The original code-switched sentence contains at least two Cantonese words before the switch point. To satisfy the lexical constraint of no English words post-translation, the sentence is truncated after the last valid translation, in this case occurring at the switch point. This removes the untranslated English segments.

Table 4: Diagnostics for the deterministic sentence-matching procedure.

Matching Rank Statistic	Mean	Median	SD
Same-speaker match rank	0.893	0	2.569
Same-group match rank	5.790	3	8.401

Token log-probabilities were computed in log space for numerical stability. Any non-finite log-probability (i.e., underflow to $-\infty$) was treated as a precision-floor event, with the corresponding token surprisal capped at $-\log_2(\epsilon_{\text{float32}}) = 126.9$ bits rather than discarding the instance. Only instances with undefined values due to token–word alignment failures (NaNs) were excluded, along with their matched unilingual pairs.

C.2 Discourse Context Translation Quality

Translation can occasionally yield unusable outputs in multi-sentence context windows (e.g., empty or malformed generations, or outputs that still contain English tokens), which would reintroduce out-of-distribution tokens into the Cantonese context. To guard against this, we computed a word-level translation coverage ratio for each context window, defined as the proportion of words in the translated output that were neither untranslatable tokens (`UNKNOWN`, `UNK`) nor residual English words. Context windows consisting entirely of Cantonese sentences require no translation and receive a ratio of 1 by construction. Context windows with a coverage ratio < 0.3 were flagged as invalid; in practice, no context windows in the final dataset fell below this threshold.

D Generalized Mixed-Effects Logistic Regression Models

The following tables are the other four mixed-effects logistic regression models: (1) Table 5 is a model using surprisal from an autoregressive language model with no prior context, (2) Table 6 is a model using surprisal from the autoregressive model with three sentences of preceding context, (3) Table 7 is a model using surprisal from a masked language model with no prior context, and (4) Table 8 is a model using surprisal from a masked language model with three sentences of preceding context.

	Log-Odds Estimate	Std. Error	z score	p-value
Intercept	-2.05	0.23	-9.07	< 0.001
Word Length	0.83	0.11	7.89	< 0.001
POS: Noun	-0.22	0.08	-2.62	0.009
POS: Verb	0.10	0.10	1.08	0.279
Surprisal	0.03	0.01	4.20	< 0.001
Bilingual Group (Heritage)	0.10	0.24	0.41	0.683
Bilingual Group (Immersed)	-0.36	0.25	-1.47	0.143
Surprisal × Heritage	-0.01	0.01	-0.49	0.623
Surprisal × Immersed	0.02	0.01	1.48	0.138

Table 5: Mixed-effects logistic regression results from Model (1a) (autoregressive language model with no prior context) predicting CS probability. Higher surprisal was associated with a greater likelihood of CS.

	Log-Odds Estimate	Std. Error	z score	p-value
Intercept	-1.58	0.22	-7.14	< 0.001
Word Length	1.06	0.10	10.08	< 0.001
POS: Noun	-0.19	0.08	-2.28	0.022
POS: Verb	0.11	0.09	1.18	0.239
Surprisal	-0.02	0.01	-2.92	0.004
Bilingual Group (Heritage)	0.05	0.23	0.23	0.814
Bilingual Group (Immersed)	-0.11	0.23	-0.47	0.635
Surprisal × Heritage	-0.00	0.01	-0.43	0.665
Surprisal × Immersed	0.00	0.01	0.30	0.766

Table 6: Mixed-effects logistic regression results from Model (1b) (autoregressive language model with three sentences of preceding context) predicting CS probability. Higher surprisal was associated with a lower likelihood of CS.

	Log-Odds Estimate	Std. Error	z score	p-value
Intercept	-2.12	0.23	-9.26	< 0.001
Word Length	0.87	0.10	8.53	< 0.001
POS: Noun	-0.24	0.08	-2.89	0.004
POS: Verb	0.07	0.10	0.71	0.478
Surprisal	0.06	0.01	6.46	< 0.001
Bilingual Group (Heritage)	-0.07	0.21	-0.35	0.729
Bilingual Group (Immersed)	-0.24	0.21	-1.13	0.259
Surprisal × Heritage	0.01	0.01	0.45	0.651
Surprisal × Immersed	0.02	0.01	1.29	0.197

Table 7: Mixed-effects logistic regression results from Model (2a) (masked language model with no prior context) predicting CS probability. Higher surprisal was associated with a greater likelihood of CS.

	Log-Odds Estimate	Std. Error	z score	p-value
Intercept	-2.15	0.23	-9.26	< 0.001
Word Length	0.98	0.10	9.58	< 0.001
POS: Noun	-0.21	0.08	-2.47	0.013
POS: Verb	0.06	0.10	0.64	0.520
Surprisal	0.06	0.01	5.59	< 0.001
Bilingual Group (Heritage)	-0.19	0.20	-0.96	0.339
Bilingual Group (Immersed)	-0.02	0.19	-0.08	0.937
Surprisal × Heritage	0.03	0.02	1.75	0.080
Surprisal × Immersed	-0.01	0.01	-0.55	0.583

Table 8: Mixed-effects logistic regression results from Model (2b) (masked LM with three sentences of preceding context) predicting code-switching. Higher surprisal was associated with a greater likelihood of code-switching.