Surprisal Predicts Code-Switching in Chinese-English Bilingual Text

Jesús Calvillo [1]
jzc1104@psu.edu

Le Fang [1]
fredfang1203@gmail.com

Jeremy Cole [2,1]
jrcole@google.com

David Reitter [2,1]
reitter@google.com

[1]Pennsylvania State University
[2]Google Research

Abstract

Why do bilinguals switch languages within a sentence? The present observational study asks whether word surprisal and word entropy predict code-switching in bilingual written conversation. We describe and model a new dataset of Chinese-English text with 1476 clean code-switched sentences, translated back into Chinese. The model includes known control variables together with word surprisal and word entropy. We found that word surprisal, but not entropy, is a significant predictor that explains code-switching above and beyond other well-known predictors. We also found sentence length to be a significant predictor, which has been related to sentence complexity. We propose high cognitive effort as a reason for code-switching, as it leaves fewer resources for inhibition of the alternative language. We also corroborate previous findings, but this time using a computational model of surprisal, a new language pair, and doing so for written language.

1 Introduction

Code-Switching (CS) occurs when a speaker alternates from one language to another during linguistic communication (e.g., Poplack, 1980). For example, in: "洗衣房在basement。" ("The laundry room is in the basement."), the speaker alternates from Chinese to English by introducing the word "basement", replacing the Chinese word "地下室". This behavior is very common among bilinguals.

Many factors have been shown to affect the propensity of a bilingual to code-switch. Among others, there are variables related to the participants in the conversation (e.g., Blom and Gumperz, 1972), the ease of production of the relevant words (e.g., Gollan and Ferreira, 2009), the linguistic context (e.g., Clyne, 1991), memory limitations of the speaker (Eppler, 2011), cognitive load and emotional state of the speaker (e.g., Grosjean, 1982; Dornic, 1978), and the type of information to be conveyed (e.g., Karrebæk, 2003; Myslín and Levy, 2015). Among the latter, predictability, as measured by word completion, has been correlated with code-switching (Myslín and Levy, 2015). In this paper, we model predictability using word surprisal calculated with a language model.

We ask whether word surprisal (Hale, 2001) and word entropy (Roark et al., 2009) affect the probability of CS within a sentence (intra-sentential CS), while controlling for other known psycholinguistic factors. Word surprisal measures how unpredictable a word is in its context, typically operationalized as the negative log-probability of a word w_i conditioned on a window of t previous words:

$$surp(w_i) = -logP(w_i|w_{i-1}, ..., w_{i-t}) \quad (1)$$

Word entropy *before* w_i measures the uncertainty when w_i is still unknown, operationalized as the expectation over the vocabulary of word surprisal:

$$H_{i-1} = \sum_{w \in \text{vocab}} -logP(w|w_{i-1}, ..., w_{i-t}) * P(w|w_{i-1}, ..., w_{i-t}) \quad (2)$$

Thus, given a context, word surprisal measures how unpredictable a *specific* word is, while word entropy measures how unpredictable *all* words are in average. These variables have been related to a very wide range of psycholinguistic phenomena (e.g, Hale, 2001, 2006; Smith and Levy, 2013; Demberg and Keller, 2008; Calvillo and Crocker, 2015; Henderson et al., 2016; Frank and Willems, 2017; van Schijndel and Linzen, 2018; Brennan and Hale, 2019).

We collected a corpus of Chinese-English text from online forum conversations where the majority of sentences are in Chinese but some sentences contain segments in English. These code-switched

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4029–4039,
November 16–20, 2020. ©2020 Association for Computational Linguistics

sentences were translated into Chinese and compared to sentences with similar syntactic structure but without any code-switch, in order to see what factors affected the propensity of CS. With this paper, we make a curated version of this dataset publicly available, together with the code that was used for its extraction and processing.[1] Then, we fitted a logistic regression model to predict CS in a sentence, testing whether the addition of surprisal and entropy improves a model that only contains control factors.

The results show that word surprisal improves the quality of the model. Since surprisal has been related to cognitive effort of language production (e.g., Kello and Plaut, 2000) and comprehension (Hale, 2001), we can relate CS to states in which the speaker faces difficulties, and/or, similar to Myslín and Levy (2015), as a strategy to signal highly informative content.

Conversely, we found no evidence of word entropy improving the model. Furthermore, word entropy does not reach significance even when used as the only predictor. While further testing is needed, we attribute this result to the fact that during language production, speakers are completely aware of the semantics they try to convey, radically reducing the number of possible word continuations, thus reducing the effort of selecting a word among multiple possibilities. In the case of surprisal, we interpret the effect observed here as the facilitation that the previous words could have on the production of the next word, irrespective of the semantics' effect.

The rest of this document is organized as follows: Section 2 presents a selection of factors that have been known to affect CS. Section 3 explains the method that we used to obtain the Chinese-English corpus and analyze it. Section 4 shows the results of the analysis. Finally, sections 5 and 6 present the Discussion and Conclusion respectively.

2 Factors that predict CS

We can arrange some of the factors that have been shown to affect CS according to their source:

Sociocultural: CS can be used to construct identity and modulate social distance and affiliation (Beebe and Giles, 1984). Moreover, CS can be affected by the kind of participants in a conversation. For example, Blom and Gumperz (1972) observed that Norwegian locals tended to switch from a dialect form to a standard form of Norwegian as soon as they felt the presence of non-locals.

CS can also be affected by the type of content that is conveyed. For example, speakers can use CS to try to distance themselves while talking about embarrassing (Bond and Lai, 1986) or emotional (Altarriba and Santiago-Rivera, 1994) topics.

Linguistic: CS seems to obey certain linguistic rules. E.g., at the morphological level, CS has been proposed to occur only if the switched morpheme is not bound (Poplack, 1980), and if it does not violate any syntactic rule of the languages involved (Poplack, 1980; Lederberg and Morales, 1985).

Speaker-related: Factors related to the difficulties that speakers encounter during language production. Indeed, one view of CS is that it occurs to compensate for a lack of language proficiency (Heredia and Altarriba, 2001).

Independent of the proficiency level, some words are inherently more difficult to access, in which case a speaker might choose to produce a word in a different language if it is more accessible (e.g., Gollan and Ferreira, 2009), in line with the idea of an integrated representation of a bilingual's linguistic knowledge (Putnam et al., 2018). For instance, words with higher frequency and shorter length are more accessible (D'Amico et al., 2001; Forster and Chambers, 1973). Moreover, words referring to concrete and highly imageable concepts are suggested to be more integrated in the bilingual lexicon than abstract words, predicting a greater probability of CS (Marian, 2009). Similarly, nouns are suggested to be stored in a common semantic system shared across languages, while other words are stored in language-specific areas, since the latter elicit slower and less consistent associations across languages (Marian, 2009; G. van Hell and De Groot, 1998). Thus, nouns are the class of words that is most frequently code-switched (e.g., Myers-Scotton, 1993) and borrowed (Muysken, 2000), followed by verbs and other parts of speech.

Following *Dependency Locality Theory* (Gibson, 1998, 2000), Eppler (2011) shows that the more intervening words between a potentially code-switched word and its dependency governor, the more difficult it is for the speaker to track the language of the governor, due to memory limitations, and therefore the more likely to code-switch.

Comprehender-related: CS has also been proposed as a strategy to facilitate comprehension

[1] https://github.com/lfang1/CodeSwitchingResearch

by marking portions of discourse (Auer, 1995; Gumperz, 1982; Zentella, 1997). From this view, the act of CS carries information, similar to how prosody helps comprehenders to recognize the focus of an utterance. Thus, CS has been reported to be used to increase the salience of discourse markers (De Rooij, 2000), to signal new discourse topics (Barredo, 1997; Zentella, 1997), to contrast topic and focus elements (Romaine, 1995), and to mark important discourse information (Karrebæk, 2003).

Similar to Karrebæk (2003), Myslín and Levy (2015) show evidence suggesting that speakers code-switch to mark important information in Czech-English bilingual speech, where importance is measured by the amount of semantic information that they convey. Then, more informative meanings receive more distinct encodings, reducing the risk of miscommunication.

3 Methods

Our initial hypothesis was that bilinguals are more prone to CS when the words to be produced have high surprisal; and at states of high uncertainty, as measured by word entropy. In order to test this, we used binary logistic regression to assess the effect of surprisal and entropy for predicting CS, while controlling for other well-known factors.

First, we collected a corpus of bilingual Chinese-English text that contains code-switched sentences. Then, the code-switched sentences were translated into Chinese, obtaining fully Chinese sentences. The correctness and fluency of these translations were verified using a survey in Amazon Mechanical Turk. We will refer to each of the translated versions of the code-switched sentences as a *CS-sent*; note that *these* are the sentences that are used for the analysis. Afterwards, for each CS-sent, we selected a sentence from those that did not originally contain a code-switch and that had a similar syntactic structure to the corresponding CS-sent, as described in the Alignment section below. We use *nonCS-sent* to refer to these sentences. Finally we trained a binary logistic regression model to predict whether a sentence was a CS-sent or a nonCS-sent.

To investigate the effect of surprisal and entropy more directly, the logistic regression model used several control factors reflecting some of the findings in Section 2. Then, a genetic algorithm was used to select, among all the possible models that can be obtained by combining all control factors and their two-way interactions, the model that would minimize the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which are measures of model quality that are based on the log likelihood of the model, the number of parameters of the model, and the size of the dataset. This selected model is the control model of our experiments.

Finally, the control model was compared to models that included word surprisal and entropy, respectively, to examine their relative contributions in explaining CS beyond the known correlates.

The next subsections explain more in detail the procedure that we followed to collect the corpus, obtain the measurements related to the control factors, and train the final logistic regression model.

3.1 Chinese-English Text Corpus

Previous CS research has focused on speech (e.g., Poplack, 1980; Myslín and Levy, 2015; Beebe and Giles, 1984; Karrebæk, 2003; Zentella, 1997). In contrast, we examine written language produced by Chinese-English bilinguals to generalize previous findings. Speech differs from text in that, during written language production, speakers have relatively more time to think and modify their utterances, making them less spontaneous and more complex. For example, spoken utterances tend to have fewer words, their words are shorter, and their vocabulary is less diverse than in text (Drieman, 1962; Gibson et al., 1966). Consequently, we expect CS to be a more conscious and less spontaneous act in text, possibly increasing its strategic use, and decreasing the influence of other factors that could be related to the spontaneity of speech.

To our knowledge, there is currently no Chinese-English text corpus that contains translations of the code-switched parts. Since we need the translations in order to properly estimate word surprisal, we built a corpus that includes them. The next paragraphs explain the procedure that we performed.

Data source: Data were acquired from the publicly available Chinese Students and Scholars Association Bulletin Board Systems (CSSA BBS) of the Pennsylvania State University, the Carnegie Mellon University, and the University of Pittsburgh. The users of CSSA BBS are Chinese-English bilinguals who have studied in the USA for several years.

The Stanford Chinese word segmenter (The Stanford NLP Group, 2018) was used to segment the sentences into words with the Chinese Penn Treebank standard (Xue et al., 2005). Any personal information present in the sentences such as people's proper names, telephones or addresses was removed. The sentences are related to four main topics: housing, secondhand goods, experience sharing, and ride sharing.

CS identification: To identify the sentences that contained a code-switch, we used the Google English 1-gram corpus, such that if an English word (a word contained in the English 1-gram corpus) was identified in a sentence, that sentence was considered as code-switched. By this simple definition, 4740 code-switched sentences and 14956 non-code-switched sentences were identified and extracted.

Translation: Five Chinese-English bilinguals translated the code-switched sentences into Chinese, obtaining one translation per sentence. While multiple translations could be possible for some sentences, more than half of the code-switches corresponded to single words, suggesting that only few alternatives were available in most cases. The translators were all international Chinese undergraduate students who have similar language proficiency and cultural background to the original posters in the CSSA BSS corpus.

Cleaning: During the translation, it became clear that a large amount of code-switches corresponded to proper nouns and words that had no clear translation to Chinese. These types of code-switches might occur for completely different reasons: for instance, there might be no way in Chinese to refer to a particular bar in Pittsburgh. In order to distinguish the sentences that clearly contained an interesting code-switch, we manually classified the sentences into four categories: *clean_cs*, *proper_nouns*, *internet_slang*, and *other*. The *other* category includes incomplete sentences and unidentified words. After this point, we only considered the sentences in *clean_cs* because in those cases a clear equivalent in Chinese existed. This group had 1690 sentences with code-switches mostly related to common nouns and adjectives (e.g., "neighborhood", "basement", and "available").

Alignment: The predictors are defined with respect to the first word that was code-switched in each CS-sent, which we call the *CS-point*. Since CS normally occurs if it does not violate any syntactic rule of the involved languages (Poplack, 1980), we compared the CS-sents to the nonCS-sents only at points of the nonCS-sents where CS would be plausible. Thus, we paired each CS-sent to the nonCS-sent that had the most similar syntactic structure among the available nonCS-sents. Through this process, we obtain a single CS-point for each nonCS-sent, at which CS is plausible to happen according to the syntactic structure. We expect this balancing also to reduce unforeseen confounds that could make difficult the results' interpretation, allowing us to analyze CS at any location of the sentence and with any syntactic structure.

original	整个	**house**	家具		齐全
CS-sent	整个	房子	家具		齐全
POS	DT	**NN**	NN		VA
	whole	**house**	furniture		complete
nonCS-sent	全部	木头	地板	,	干净
POS	DT	**NN**	NN	PU	VA
	all	**wood**	floor		clean

Table 1: Example of CS-sent / nonCS-sent alignment.

First, we used the Stanford Parser to obtain part-of-speech tags (*POS*) and dependency trees of all sentences. Then, for each CS-sent, we selected the most similar nonCS-sent according to the Levenshtein similarity of their POS sequences. We only considered alignments that had at least a 40% Levenshtein similarity, discarding all CS-sents for which no nonCS-sent fulfilled that requirement. In addition, the selected nonCS-sent had to contain the same ngram of POS corresponding to the POS of the words at CS-point-1 (if the switch is not sentence-initial), CS-point, and CS-point$+1$ (if the switch is not sentence-final). Table 1 shows an example, where DT NN NN is said ngram. Finally, when there were available candidates (which happened for 92.2% of the CS-sents), the dependency relation of the word at the CS-point to its governor had to be the same in the nonCS-sent (compound:nn in the example of Table 1).

In the end, there were 1476 pairs CS-sent/nonCS-sent. Regarding sentence length, μ=11.1, mode=6, min=2 and max=43. Concerning the index of the CS-point, μ=5.15, mode=2, and max=30. For the words at the CS-point, 62% are nouns, 23% are verbs, and 15% have other POS, replicating previous findings (Myers-Scotton, 1993).

Verification of the Translations: The translations were verified in their fluency (whether they resemble native Chinese utterances) and correctness (whether they reflect the semantics of the original sentences). Using a sample of 500 CS-sents of our dataset, and Amazon's Mechanical Turk, we recruited 33 native Chinese speakers with high proficiency in English. Each participant was shown

60 pairs of the original code-switched sentences and their translations, such that each pair was verified by 3 participants. Each participant was asked to rate the correctness and fluency of the translations. Additionally, the survey contained low/high quality control items, in order to verify the participants' engagement. We removed the participants who showed no reaction to the manipulations of the control items. The participants rated in average the quality of the translations as 4.04 ($\sigma = 1.3$) out of 5 in correctness, and 4.08 ($\sigma = 1.3$) out of 5 in fluency, indicating that the translations are fluent Chinese sentences and that they adequately reflect the original meaning of the code-switched sentences.

3.2 Control Variables

We introduced several controls in order to account for findings documented in the CS literature described in Section 2, and to see whether our variables of interest can explain the data beyond the controls.

Considering that all sentences were produced by speakers of the same community, with similar age and educational background, we can assume that sociocultural factors are homogeneous in the analyzed sentences. Moreover, with respect to linguistic factors, we selected the nonCS-sents and their CS-points to be similar to the CS-sents.

We considered the following variables as controls in our experimental setup. These are measured in the CS-sents and the nonCS-sents at their CS-points, and introduced in the logistic regression model as predictors. Note that many code-switched segments contain more than a single word, however we focus only on the first word, as it is where the code-switch actually occurs.

Word Frequency: Words with lower frequency are considered less accessible, so bilinguals are more likely to code-switch when the intended words are infrequent. The relative frequency of the word at the CS-point was calculated from the Google Chinese 1-gram corpus. These frequencies were converted to negative logs before introducing them to the logistic regression model.

Word length: The number of Chinese characters forming the word at the CS-point. Longer words are considered less accessible, and therefore more prone to CS.

Sentence length: The number of words in the sentence. To our knowledge, there has been no studies analyzing the relation between CS and sentence length, however, this measure has been used to assess sentence complexity (e.g., Howcroft and Demberg, 2017; Petersen, 2007). We hypothesize that bilinguals are more likely to CS when a sentence is longer, as it implies a higher effort to retrieve and produce the relevant structures and words. In these more demanding occurrences, there would be less available resources left to inhibit the alternative language, thereby increasing the propensity to CS.

Part-of-speech tag: The POS of the word at the CS-point. "NR" (proper noun), "NN" (common noun), "NT" (temporal noun) were all converted to "noun". "VE" (e.g, "be" and "have") and "VV" (other verb) were all converted to "verb". For the regression model, we only considered 3 classes: "noun", "verb", and "other"; the latter referring to the POS-tags that are not related to nouns or verbs.

Dependency relation: Bilinguals might be more likely to CS when a word holds a specific dependency relation to its governor in a dependency tree. Hence, the word at the CS-point was annotated with the dependency relation that connects it to its governor. We only considered and introduced as categorical predictors the relations that occurred more than 100 times: "compound:nn", "nsubj", "dobj", "root", "dep", "amod". Then, every relation that did not occur at least 100 times was grouped into "other".

Dependency distance: Bilinguals are more prone to CS when the distance between a word and its dependency governor is longer (Eppler, 2011). We measured dependency distance as the difference between the index of the word at the CS-point and the index of its dependency governor (i.e., two adjacent words have a dependency distance of 1), only when the governor is to the left of the CS-point (otherwise we assign a distance of 0).

Location: We use a discrete variable with 3 levels to encode whether the CS-point is located at the beginning, middle or end of the sentence (10-80-10 percent of the sentence).

Regarding POS, dependency relations and location, we did not expect a strong effect on CS because CS-sents and nonCS-sents are matched to be syntactically similar. Nonetheless, we include these predictors in order to capture any remaining effect of syntactic structure on CS.

3.3 Variables of Interest

These are the variables we wanted to test in order to see whether they are relevant predictors of CS.

Word Surprisal: The negative log probability of the word at the CS-point given the 4 previous words in the sentence calculated using a 5-gram language model trained on the Chinese Wikipedia with the SRILM framework $-logP(w_i|w_{i-1},...,w_{i-4})$. Out-of-vocabulary words were assigned the highest surprisal value found in the corpus (i.e., 10.166).

Word Entropy: Measured *before* the word at the CS-point w_i, and given the previous 4 words ($t = 4$ in equation 2), calculated with the same language model that we used for surprisal.

3.4 Modeling Procedure

We used logistic regression to predict whether a sentence belongs to the set of CS-sents or to the set of nonCS-sents. The model's predictors are a combination of the control factors mentioned above, as well as word surprisal and entropy. Before training, all numerical predictors were standardized such that their mean is 0 and standard deviation is 0.5.

Considering that we are mainly interested in the effect of surprisal and entropy, we first obtained a parsimonious control model using all the control factors and their two-way interactions. In order to select the best combination of control factors and interactions, we followed Myslín and Levy (2015) using a genetic algorithm (Calcagno et al., 2010) to find the model that minimizes the AIC and BIC. Afterwards, we introduced word surprisal and entropy in order to see whether they improve the quality of the model.

4 Results

4.1 Selection of the Control Model

The models selected using AIC and BIC differ slightly: they have the same main effects, but the AIC model has some additional interactions. BIC penalizes the number of parameters more heavily than AIC when the number of data points is relatively large. In contrast, the value of AIC does not depend on the number of data points. Since our dataset is relatively large ($n = 2952$), we chose the model selected using BIC:

$$CS \sim postag + freq + w\_length + s\_length \quad (3)$$

where postag, freq and w_length refer respectively to the POS, frequency and length of the word at the CS-point; and s_length is the sentence length. This model has an AIC of 3967.3 and a BIC of 4003.2. As expected, most factors related to syntax (dependency relation, dependency distance and location) were not selected, as they were mostly counterbalanced during the alignment.

4.2 Variables of Interest

After obtaining the control model, we added word surprisal and entropy to test whether they improve the quality of the model. If they do, it would mean that they are relevant predictors of CS above and beyond the control factors.

Word Surprisal: Adding word surprisal indeed improves the quality of the control model, reducing the AIC from 3967.3 to 3954.5 and the BIC from 4003.2 to 3996.4. This finding is confirmed using a likelihood ratio test ($\chi^2(1) = 14.81, p < 0.001$). Adding surprisal to a model that also includes entropy gives similar results.

The direction of this effect was as expected: words with higher surprisal are more likely to be code-switched ($\beta = 0.37, z = 3.82, p < .001$).

Word Entropy: Adding word entropy did not improve the quality of the model, increasing the AIC from 3967.3 to 3967.9 and the BIC from 4003.2 to 4009.8. Furthermore, in the resulting model word entropy does not reach significance as predictor. Adding word entropy to a model that also includes surprisal gives similar results. Finally, word entropy does not reach significance even when it is used as the only predictor.

This result was unexpected considering the relation of word entropy to word surprisal: entropy is the average over the vocabulary of the surprisal values at the CS-point. Intuitively, entropy is related to the effort of selecting the correct word at a given time, which would be related to the number of plausible words continuations at that point. However, the number of alternatives would be also limited by the semantics the speaker tries to convey. So, it is likely that the semantics reduce drastically the amount of plausible word continuations. In that case, a better measure could be the entropy over the probability distribution $P(w_i|w_{i-1},...,w_0, semantics)$, which is a direction that can be explored in future work.

4.3 Control Factors

Since word entropy did not improve the control model, we only report the model that adds word surprisal, whose parameters are shown in Table 2.

	Parameter Estimates		Wald's Test		Likelihood Ratio Test	
Predictor	Coef.β	SE(β)	Z	p_z	χ^2	p
(intercept)	-0.13	0.05	-2.75	$< .01$		
surprisal	0.37	0.09	3.82	$< .001$	14.81	$< .001$
frequency	0.21	0.11	1.91	.055	3.67	.055
word length	0.47	0.09	4.79	$< .001$	23.34	$< .001$
sentence length	0.58	0.08	7.32	$< .001$	56.10	$< .001$
POS=verb	0.46	0.10	4.42	$< .001$	20.52	$< .001$
=other	0.23	0.11	2.08	$< .05$		

Table 2: Summary of the logistic regression model after including word surprisal: coefficient estimates β, Wald's z-scores and their significance level, contribution to likelihood χ^2 and its significance level. The response variable was coded as nonCS-sent = 0 and CS-sent = 1. The baseline of the categorical variable POS is "noun". AIC/BIC before introducing surprisal: 3967.3/4003.2; after introducing surprisal: 3954.5/3996.3.

Some predictors correlate with each other. For example, infrequent words tend to be longer (Zipf, 1935), and have higher surprisal. In our dataset, the negative log of word frequency and surprisal have a Spearman's $\rho = 0.62 (p < 0.001)$. Similarly, the negative log of word frequency and word length have a Spearman's $\rho = 0.57 (p < 0.001)$. In order to asses whether collinearity would impact the quality of the model, we used the Generalized Variance Inflation Factor (GVIF). In our case, all values were below 2, meaning that although some collinearity exists, it should not be problematic for the model's results.

As one can see, most factors were significant predictors of CS, replicating previous findings:

Frequency: Words with lower frequency (higher negative log frequency) show a slight tendency to be code-switched, even while having word surprisal as a predictor ($\beta = 0.21, z = 1.91, p = 0.055$), corroborating previous findings showing that, independent of context, frequent words are more accessible and consequently less prone to CS. The relatively high p-value is likely due to the correlation of word frequency with word surprisal.

Word length: Longer words are more likely to be code-switched ($\beta = 0.47, z = 4.79, p < .001$). Similar to frequency, word length has been related to accessibility, such that longer words are less accessible and therefore more prone to CS.

Sentence length: Speakers are more likely to switch in longer sentences ($\beta = 0.58, z = 7.32, p < .001$). We explain this as a side effect of the higher production effort that longer sentences imply, as longer sentences require more tokens and structures to be retrieved and produced. Under these circumstances, people may have less resources to control/inhibit the production of words in the alternative language, thereby increasing the probability of CS.

Part-of-speech tag: Compared to the baseline (nouns), verbs are more likely to be classified as code-switched ($\beta = 0.46, z = 4.42, p < .001$), followed by other POS ($\beta = 0.23, z = 2.08, p < .05$). Since *noun* is the most common code-switched POS in the corpus, we interpret this result not as nouns being less likely to be switched, but as the model relying more on the other predictors when it encounters a noun, since the values related to POS would be zero.

5 Discussion

In this study we explore the effect of word surprisal and entropy on CS. The computation of these measures relies on language models trained with sufficient and appropriate data, which is non-trivial in the case of bilingual text. Moreover, even assuming a large bilingual corpus, utterances tend to appear in segments of the same language, so any code-switch is likely to cause an increase of surprisal on the word where the switch occurs. For example, in: "洗衣房在basement。", we expect "basement" to have high surprisal at least partly because an English word does not tend to follow a sequence of Chinese words. This makes it difficult to assess whether the increase of surprisal causes the code-switch or vice versa. Conversely, by using a translated version, if there is an increase of surprisal, it would not be because of the code-switch, but possibly because the concept or Chinese word is infrequent.

Considering these aspects, we used a monolingual Chinese language model to see whether word surprisal at the CS-point can predict CS assuming that the switch never occurred, as previously described. While we performed several steps to verify the quality of the translations, it is possible that the translation process could introduce uncommon constructions, increasing the surprisal values. However, in more than half of the code-switched sentences, the switch corresponded to a single word, likely resulting in few translation alternatives. Moreover, the verification survey showed that the translations are fluent and semantically correct. Consequently, we expect the sentences that we used to be appropriate for our study.

Using this new Chinese-English CS dataset, we tested several factors that have been shown to affect CS. We found that long and infrequent words are more likely to be code-switched, which is compatible with previous findings suggesting that words with these characteristics are less accessible and more likely to be code-switched.

Another important predictor was sentence length, where longer sentences are more likely to be code-switched. This could reflect the effort related to produce longer sentences, as they require more structures to be retrieved and handled.

Critically, word surprisal was also a relevant factor for predicting CS. Since surprisal has been related to cognitive effort in language comprehension (Hale, 2001) and to some extent in language production (e.g., Kello and Plaut, 2000), we may interpret word surprisal as the degree to which the words that were previously produced facilitate production. From this view, word surprisal would index context-dependent accessibility, in contrast to word frequency, which would index context-free word accessibiity. Then, words with high surprisal would be less accessible, similar to infrequent words.

Unexpectedly, word entropy did not seem to predict CS. If word entropy indexes the effort of choosing among multiple possible word continuations, then our calculation did not reflect the true probability distribution, as the number of candidate words would be drastically reduced by the semantics the speaker tries to convey. Something similar would happen with word surprisal, in that case, we interpret the results observed here as the facilitation that the previous words have on the production of the next word, beyond the effect that the semantics could have. We expect that entropy calculations conditioned on semantics could give different results.

The current model of bilingual representations suggest that bilinguals actively inhibit the alternative language when speaking in a second language (Green, 1998; Meuter and Allport, 1999). Nonetheless, bilinguals often code-switch, suggesting that inhibition might depend on the context, available resources, and even audience design. For example, when a bilingual is with other bilinguals, he/she might feel more free to code-switch using the most accessible words, knowing that the audience would understand both languages (Blanco-Elorrieta and Pylkkänen, 2018); while in monolingual situations, the bilingual would use the appropriate language, inhibiting the alternative one (e.g., Blom and Gumperz, 1972). Alternatively, if the bilingual is under high cognitive load, the available resources for inhibition would be less, reducing inhibition and increasing the probability of CS (e.g., Dornic, 1978).

Most predictors in our model can be related to production effort: long and infrequent words are less accessible and therefore harder to produce; longer sentences require more structures to be handled; assuming word surprisal reflects context-dependent accessibility, high-surprisal words would also be harder to produce. In this context, we propose CS as a result of high cognitive effort, leaving less resources for inhibition, and thus increasing the probability of CS.

Another explanation for CS, within audience design, is one of strategy, where CS can highlight segments with high information density in order to emphasize their content. This might increase the probability of successful comprehension. To study that possibility, Myslín and Levy (2015) used *predictability of meaning* (PoM). This notion relates to word surprisal perhaps in a similar way to how word surprisal relates to word frequency: they both encode how unlikely a given word is, correlating to some degree, however they are not completely redundant. PoM is calculated by asking comprehenders to guess possible word continuations – in either language – given a context and within a limited number of guesses. PoM corresponds to the accuracy of that guess. This is where our study contributes important data using the automatic, neutral, and well-studied surprisal metric. Word surprisal is calculated using a language model, permitting more nuanced online estimations. In terms of their interpretation, PoM is, as its name suggests, more related to semantics, as it is language-independent and elicited offline (without time pressure). By contrast, word surprisal is language-specific and encodes both how likely a meaning and a specific word are. Thus, an example in which these metrics would differ is when a meaning is predictable, but the related word in a specific language is not, giving a high PoM but also a high word surprisal.

Further modeling is needed to disentangle these explanations (cognitive effort, audience design). It is possible that if a speaker has difficulties during production, CS would be more likely to occur; and at the same time, if the speaker believes CS would help to communicate his/her message (as proposed by Myslín and Levy, 2015), then he/she might choose to code-switch. As previous findings show, CS is a phenomenon that can be affected simultaneously by a wide variety of factors.

6 Conclusion

We investigated the effect of word surprisal and word entropy on the probability of code-switching (CS) in Chinese-English written communication. A corpus of text containing Chinese-English conversations was collected and its code-switched phrases were translated back into their context language. The translations of the code-switched sentences were compared to sentences with similar syntactic structure but without code-switches, in order to see what factors affected the propensity to CS.

Surprisal predicts CS. Since surprisal has been associated with cognitive effort during language production (e.g., Kello and Plaut, 2000) and comprehension (Hale, 2001), we can relate CS to situations in which the speaker faces difficulties; and/or similar to Myslín and Levy (2015), situations where the speaker uses CS as a strategy to emphasize highly informative content to the comprehender in order to facilitate communication.

We found no evidence showing that entropy, as opposed to surprisal, predicts CS. This may be due to the formulation of entropy that we used, which does not consider the semantics the speaker tries to convey.

This paper makes two specific contributions. The first one is the finding that **CS in *written language*** is reliably affected by **sentence length, word length,** arguably **word frequency**, and, most importantly **word surprisal**. The second contribution is a **new Chinese-English CS dataset**, which includes translations to the dominant language, which we hope will be used in further models of CS.

Acknowledgements

This research was funded by grant 1734304 from the National Science Foundation. We thank Matthew Kelly, Saranya Venkatraman, Zixin Tang, and Michael T. Putnam for their comments, and Jing Wang for providing a Chinese language model. We thank the anonymous reviewers whose comments have greatly helped improve this work.

References

Jeanette Altarriba and Azara L Santiago-Rivera. 1994. Current perspectives on using linguistic and cultural factors in counseling the hispanic client. *Professional Psychology: Research and Practice*, 25(4):388.

Peter Auer. 1995. The pragmatics of code-switching: A sequential approach. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 115–135.

Inma Munoa Barredo. 1997. Pragmatic functions of code-switching among basque-spanish bilinguals. *Retrieved on October*, 26:2011.

Leslie M. Beebe and Howard Giles. 1984. Speech-accommodation theories: A discussion in terms of second-language acquisition. *International Journal of the Sociology of Language*, 1984(46):5–32.

Esti Blanco-Elorrieta and Liina Pylkkänen. 2018. Ecological validity in bilingualism research and the bilingual advantage. *Trends in cognitive sciences*, 22(12):1117–1126.

Jan Petter Blom and John J. Gumperz. 1972. Social meaning in linguistic structures: Code switching in northern norway.

Michael H Bond and Tat-Ming Lai. 1986. Embarrassment and code-switching into a second language. *Journal of Social Psychology*, 126(2):179–186.

Jonathan R Brennan and John T Hale. 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741.

Vincent Calcagno, Claire de Mazancourt, et al. 2010. glmulti: an R package for easy automated model selection with (generalized) linear models. *Journal of statistical software*, 34(12):1–29.

Jesús Calvillo and Matthew Crocker. 2015. A rational statistical parser. *Natural language processing and cognitive science: Proceedings 2014*.

Michael Clyne. 1991. *Community languages: the Australian experience*. Cambridge University Press.

Simonetta D'Amico, Antonella Devescovi, and Elizabeth Bates. 2001. Picture naming and lexical access in italian children and adults. *Journal of Cognition and Development*, 2(1):71–105.

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

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Stanislav Dornic. 1978. The bilingual's performance: Language dominance, stress, and individual differences. In *Language interpretation and communication*, pages 259–271. Springer.

Gerard HJ Drieman. 1962. Differences between written and spoken language: An exploratory study. *Acta Psychologica*, 20:78–100.

Eva Duran Eppler. 2011. The dependency distance hypothesis for bilingual code-switching. In *Proceedings of the International Conference on Dependency Linguistics*.

Kenneth I Forster and Susan M Chambers. 1973. Lexical access and naming time. *Journal of verbal learning and verbal behavior*, 12(6):627–635.

Stefan L Frank and Roel M Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126, Cambridge, MA. MIT Press.

James W Gibson, Charles R Gruner, Robert J Kibler, and Francis J Kelly. 1966. A quantitative examination of differences and similarities in written and spoken messages. *Communications Monographs*, 33(4):444–451.

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.

David W Green. 1998. Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and cognition*, 1(2):67–81.

François Grosjean. 1982. *Life with two languages: An introduction to bilingualism*. Harvard University Press.

John J Gumperz. 1982. *Discourse strategies*, volume 1. Cambridge University Press.

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 on Language technologies*, pages 1–8. Association for Computational Linguistics.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive science*, 30(4):643–672.

Janet G. van Hell and Annette MB De Groot. 1998. Disentangling context availability and concreteness in lexical decision and w ord translation. *The Quarterly Journal of Experimental Psychology: Section A*, 51(1):41–63.

John M Henderson, Wonil Choi, Matthew W Lowder, and Fernanda Ferreira. 2016. Language structure in the brain: A fixation-related fmri study of syntactic surprisal in reading. *Neuroimage*, 132:293–300.

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

David M Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968.

Martha Sif Karrebæk. 2003. Iconicity and structure in codeswitching. *International Journal of Bilingualism*, 7(4):407–441.

Christopher T Kello and David C Plaut. 2000. Strategic control in word reading: evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(3):719.

Amy R Lederberg and Cesáreo Morales. 1985. Code switching by bilinguals: Evidence against a third grammar. *Journal of Psycholinguistic Research*, 14(2):113–136.

Viorica Marian. 2009. Language interaction as a window into bilingual cognitive architecture. *Multidisciplinary approaches to code switching*, pages 161–185.

Renata FI Meuter and Alan Allport. 1999. Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of memory and language*, 40(1):25–40.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Carol Myers-Scotton. 1993. *Social Motivations for Codeswitching: Evidence form Africa*. Clarendon Press.

Mark Myslín and Roger Levy. 2015. Code-switching and predictability of meaning in discourse. *Language*, 91(4):871–905.

Sarah Elizabeth Petersen. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplication for Bilingual Education*. Ph.D. thesis, University of Washington.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Michael T. Putnam, Matthew Carlson, and David Reitter. 2018. Integrated, not isolated: Defining typological proximity in an integrated multilingual architecture. *Frontiers in Psychology: Language Sciences*.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 324–333. Association for Computational Linguistics.

Suzanne Romaine. 1995. *Bilingualism*. Wiley-Blackwell.

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

The Stanford NLP Group. 2018. Stanford word segmenter (version 3.9.2). `https://nlp.stanford.edu/software/segmenter.shtml`.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Ana Celia Zentella. 1997. Growing up bilingual: Puerto rican children in new york. *Lingua*, 1(103):59–74.

George Kingsley Zipf. 1935. The psycho-biology of language: An introduction to dynamic philology.