R250 Replication Project Report

Which Measures of Uncertainty Predict Code-Switching?

Zébulon Goriely, Queens' (zg258)

Word Count: 3972[1]

1 Introduction

Code-switching is the alternation of two or more language within a single discourse, identified by the integration of another language's phonology, morphology and/or syntax within the context of that conversation (Poplack, 2000). Originally considered a substandard use of language (Weinreich, 2010), it is now understood to be a natural linguistic phenomenon present in written and spoken multilingual conversation.

Despite being well studied in linguistics, psycholinguistics and sociolinguistics, code-switching is a relatively understudied phenomenon in the field of natural language processing (NLP), despite being prevalent on social media platforms (Mave et al., 2018). Even identifying code-switching can be difficult, given that it can occur at the sentence, clause, word and even sub-word level. This causes difficulties for monolingual models used to perform standard NLP tasks such as named entity recognition and part-of-speech tagging, especially given the lack of annotated data. As a result, the study of code-switching in NLP is considered low-resource and non-standard (Aguilar and Solorio, 2020).

Calvillo et al. (2020) contribute to the study of code-switching for NLP in two ways. Firstly, they provide a dataset of Chinese-English text with 1476 clean code-switched sentences, sourced from a student forum for Chinese-English bilinguals. The dataset also contains translations of the code-switched sentences back into Chinese and aligns these sentences with non-code-switched sentences, allowing for linguistic analysis at the point where code-switching occurs that was not possible with previous datasets.

The second contribution of Calvillo et al. (2020) was to train a logistic regression model using well-known predictors, then add word entropy (Roark et al., 2009) and word surprisal (Hale, 2001) to find that surprisal, but not entropy, is a significant predictor for code-switching. Their model also corroborates previous findings, following the hypothesis that multilinguals actively inhibit code-switching, but that high cognitive load may reduce the resources available for this inhibition. They argue that surprisal is a good model for cognitive effort in this case, as also suggested by studies of language production (Kello and Plaut, 2000) and language comprehension (Hale, 2001).

In this report, I replicate the work of Calvillo et al. to investigate whether surprisal is a significant predictor for code-switching. I also train a Chinese language model on a different corpus and calculate my own values for surprisal, finding that these values corroborate their findings. I also use this language model to calculate mutual information at the location of the code-switch in each sentence. As an alternative measure of uncertainty, I find that mutual information is also a significant predictor of code-switching.

2 Background

2.1 Factors that Predict Code-Switching

A single speaker may code-switch not just due to their proficiency with the language, but also because of who they are speaking to, what information they are trying to convey, and variety of other competing and contextually-dependent factors. These factors can broadly be labelled as **speaker-related**, **comprehender-related**, **linguistic** and **sociocultural**. In this report, I follow Calvillo et al. (2020) in focusing on the speaker-related factors that may predict code-switching, specifically factors related to cognitive load or limitations of working memory. To study these factors, the authors produced a dataset designed to try to account for the other types of factors, as described in the next section.

[1] `texcount report/report.tex`

2.2 Chinese-English Code-Switching Dataset

The corpus produced by Calvillo et al. contains 1476 pairs of sentences, sourced from a publicly available bulletin board system for Chinese students in Pennsylvania. The first sentence in each pair is one that contains code-switching. The translation of these sentences back into Chinese are also provided, referred to as *CS-sents*. The second sentence in each pair does not contain code-switching (it is entirely in Chinese) and is referred to as a *nonCS-sent*. The instances of code-switching are described as "clean", all occurring when a clear equivalent in Chinese existed, rather than because the English word is a proper noun or internet slang that does not have a clear translation. The authors argue that these other instances of code-switching may occur for completely different reasons and so are not of interest to the study of code-switching caused by cognitive load.

A major feature of the dataset is the way in which the CS-sents and nonCS-sents are paired. As the authors were interested in studying the speaker-related factors that influence code-switching, the data collection process was designed to account for most of the other factors that can predict it. By collecting sentences from speakers of the same community, they assume that sociocultural factors are homogeneous in the dataset. To account for linguistic factors, they designed the dataset so that CS-sents were paired with nonCS-sents that had the most similar syntactic structures. The goal was to align pairs syntactically to facilitate information-theoretic analyses comparing the point at which code-switching occurs in the CS-sent to the point where code-switching would have occurred, but did not, in the nonCS-sent. This point is defined as the *CS-point*, the location of the first word that is code-switched in the CS-sent.

This alignment process involved taking the Levenshtein similarity of the POS sequences of the sentences to find candidate nonCS-sents, only pairing CS-sents with nonCS-sents when the trigram of POS tags around the CS-point matched. They also preferred pairs of sentences whose dependency relation labels at the CS-point matched, which was possible for 92.2% of the pairs. This resulted in pairs of CS-sents and nonCS-sents that have very similar syntactic structures, especially around the CS-point. An example of such a pair can be seen in fig. 1.

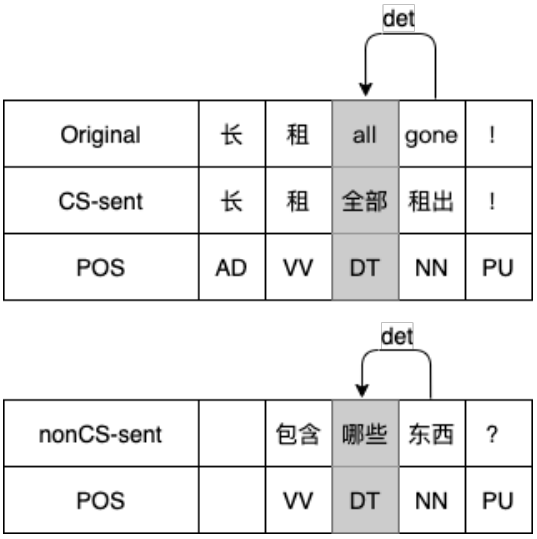


Figure 1: Two aligned sentences in the Chinese-English code-switching dataset, with the CS-point in grey. The 3-gram of POS tags around the CS-point are identical, as is the label of the dependency relation to the word at the CS-point.

The dataset contains not just these pairs of sentences but also a variety of features extracted from the sentences that can be used to train and test a model that predicts code-switching. Many of these features are used as control variables, based on findings documented in the code-switching literature:

- **Word Frequency**: The relative frequency of the word at the CS-point, calculated using the Google 1-gram corpus and stored as a negative logs.
- **Word Length**: The number of Chinese characters making up the word at the CS-point.
- **Sentence Length**: The number of words in the sentence (not previously studied before the authors, but used as a measure of sentence complexity in other work).
- **Part-of-Speech Tag**: The POS of the word at the CS-point, considering just the classes "noun", "verb" and "other".
- **Dependency Relation**: The tag of the dependency relation connecting the word at the CS-point to its governor, considering just the relations that occurred more than 100 times.
- **Dependency Distance**: The difference in index between the word at the CS-point and that word's governor, when the governor is to the left of the CS-point, otherwise 0.

- **Location**: A discrete variable to encode whether the CS-point is located in the beginning, middle or end of a sentence using a 10-80-10 split.

These features are all motivated by previous studies that found them to be good predictors for code-switching. Longer and more infrequent words are harder to access (D'Amico et al., 2001), increasing the likelihood of code-switching (Gollan and Ferreira, 2009; Forster and Chambers, 1973). Nouns and verbs are more likely to be code-switched over other parts of speech (Myers-Scotton, 1995). Longer dependency relations increase the likelihood of code-switching due to memory limitations (Eppler, 2011). The dependency relation, location and sentence length were not directly linked to previous studies, but were presumably included to account for the remaining syntactic structure. The authors included sentence length in particular because previous studies had linked it to sentence complexity, so it could be linked to cognitive load and the inability to inhibit code-switching.

The actual dataset contains many additional features not mentioned in the paper, such as the index of the dependency governor, whether the word at the CS-point is the root of the sentence and other encodings of location (25-50-25, 30-40-30, first-middle-last). Through initial statistical analyses carried out in the R code in their repository, they found these features not to be relevant and in the case of the location variable, found the 10-80-10 split to be a better representation than the others. Part of my replication will involve revisiting these analyses.

The dataset also includes the variables of interest to the study: entropy and surprisal. To calculate entropy and surprisal, they used a Chinese 5-gram corpus trained on Chinese wikipedia. Word surprisal measures unpredictability of a word w_i given its context. In this case, the context is the previous four words:

$$\text{surp}(w_i) = -\log P(w_i|w_{i-1}, \ldots, w_{i-4})$$

Word entropy measures the uncertainty of the sentence by finding the expectation of word surprisal over the vocabulary. In this dataset, it is calculated before the word w_i:

$$H_{i-1} = \sum_{w \in \text{vocab}} -\log P(w|w_{i-1}, \ldots, w_{i-4}) \times P(w|w_{i-1}, \ldots, w_{i-4})$$

As with the control features, variants of these measures are also available in the dataset (such as average surprisal and entropy after the CS-point) but initial analysis by the author's study found these not to be relevant.

2.3 Mutual Information

To investigate whether other measures of uncertainty could be used to predict code-switching, I calculated the *pointwise mutual information* at the CS-point in each sentence. Pointwise mutual information is an information-theoretic measure of association that has been used to represent uncertainty in other areas of computational linguistics, such as in modelling speech (Jelinek, 1997; Brent, 1999) and part-of-speech tagging (Stratos, 2018). In this report, I use mutual information to represent the association between a word w_i and its four-gram context as follows:

$$\text{MI}_i = -\log_2 \frac{P(w_i \ldots w_{i-4})}{P(w_i)P(w_{i-1} \ldots w_{i-4})}$$

2.4 Logistic Regression Training

To model code-switching, Calvillo et al. use their dataset to train logistic regression model that predict code-switching. Specifically, given a Chinese sentence, their model predicts whether the sentence belongs to the set of CS-sents or nonCS-sents.

Using the features described above, with numerical predictors standardised to have a mean of 0 and a standard deviation of 0.5, they first train a control model using all the control factors and their two-way interactions. To find the best combination of control factors, they use a genetic algorithm provided by the *glmulti* package (Calcagno and de Mazancourt, 2010) to find the subset of parameters that minimises the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) of the model. These are measures of model quality that produce a score based on the maximum likelihood $\hat{L}$ of the model and the number of parameters k. BIC differs slightly to AIC in that it further punishes the number of parameters k as the size of the dataset n increases:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

$$\text{BIC} = k\ln(n) - 2\ln(\hat{L})$$

The genetic algorithm is necessary due to the number of possible models growing exponentially with

the number of parameters to include. The genetic algorithm randomly explores a subset of the possible models, with a bias towards better models, making computation much faster than exploring all possible models.

Once the control model is found, the authors then added the word surprisal and word entropy variables to find whether they acted as relevant predictors of code-switching by performing statistical tests based on the AIC and BIC of the resulting models.

3 Method

In this section I describe the two approaches I took to replicate Calvillo et al. (2020). In the first section, I re-ran the public source code of their experiment to validate their conclusion: that surprisal is a relevant predictor for code-switching. In the second section, I trained my own Chinese 5-gram language model on a different corpus. I use this language model firstly to see if the author's conclusions hold when using a different language model to calculate surprisal and secondly to see if mutual information can be used inplace of and alongside surprisal to act as a significant predictor of code-switching.

The script to produce the new features, enhanced dataset and modified R code can be found in my repository[2].

3.1 Replication of Code-Switching Experiment

To replicate the code-switching experiment, I downloaded the authors' publicly-available open-source repository[3]. This repository contains all the code used to produce the Chinese-English code-switching dataset, the dataset itself, and the R code used to run the experiment. The experimental process is only briefly described in the paper, but the R code was well-documented and legible, so I was able to easily run it one section at a time to confirm their findings. These are the steps to the experiment that I re-ran:

1. **Reading in the dataset**.
2. **Preprocessing the dataset**: Reducing part-of-speech tags to just three categories, removing rare dependency label, calculating dependency distance and other cleaning operations.
3. **Standardising features**: Standardising numeric features to have a normal of 0 and a standard deviation of 0.5.
4. **Single-feature modelling**: Fitting logistic regression models using single features or combinations of two features for initial analysis and for deciding which features to use when there are several options (such as the different variables for encoding CS-point location).
5. **Control model selection**: Using a genetic algorithm to select the control model, selecting the subset of features and two-way interactions of features that minimises AIC. The process is also repeated for minimising BIC.
6. **Adding variables of interest**: Adding entropy and surprisal, individually and together, to the control model.
7. **Determining significance**: Performing likelihood ratio tests comparing the new model to the control model to find whether entropy and surprisal are significant predictors.

I did not have any problems running these steps. The results for steps 4-7 are described in section 4.

3.2 Using Mutual Information to Predict Code-Switching

To investigate whether mutual information could act as a significant predictor for code-switching, I trained a new 5-gram Chinese language model and used it to calculate the mutual information at the CS-point of each sentence, given the previous four-gram context, as described in section 2.3. I also used this language model to provide a new calculation of surprisal at the CS-point to further validate the results of Calvillo et al. (2020).

To train the language model, I used the Wikipedia section of the CLUECorpusSmall corpus (Xu et al., 2020). This section of the corpus contains just over 1 million articles from Chinese Wikipedia, totalling 1.6G of text, which I tokenised using the Stanford Chinese Segmenter (Tseng et al., 2005). To create the 5-gram language model, I used the KenLM Language Model Toolkit (Heafield, 2011) with default parameters.

To calculate surprisal and mutual information, I wrote a python script that parsed each row of the

[2] https://github.com/codebyzeb/r250

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

dataset, queried the language model to fetch estimated ngram probabilities and finally calculated surprisal and mutual information at the CS-point. I then modified the R code of Calvillo et al. to incorporate these features and used this to repeat the experimental steps described in the previous section.

4 Results

In this section, I report the results of re-running the R code of Calvillo et al., then running a modified version of the code to include the newly calculated surprisal and mutual information features. When discussing the significance of parameters in logistic regression models, I use (***) to refer to significance of $p < 0.001$, (**) to refer to $p < 0.01$ and (*) to refer to $p < 0.05$, where p-values are two-tailed from the Z-value of the parameter.

4.1 Single Feature Modelling

Initial analysis not reported by Calvillo et al. involved fitting logistic regression models using single features or pairs of features in order to decide which of several options to use in the main experiment.

I first confirmed that the surprisal at the CS-point is a significant (***) predictor but that the mean surprisal of the sentence is not. This can be understood when looking at the box plots of these features against sentence type; the mean surprisal feature seems to be identically distributed between classes, as seen in fig. 2. The surprisal of the previous word and the deviation from the mean surprisal at the CS-point were also not significant, but the deviation from the previous surprisal at the CS-point was significant (**). When fitting models with pairwise combinations of these different measures of surprisal, only surprisal at the CS-point was found to be significant (***), suggesting that these related metrics do not complement each other and explaining why deviation from previous surprisal was not used in the main experiment.

This analysis was repeated for entropy, frequency, sentence length, part-of-speech tags, dependency features and location features. The initial analysis at this stage already confirmed that various measures of entropy were not significant predictors, either when combined with surprisal or when used on their own.

The last part of this initial analysis was to select which location feature to use, firstly by comparing the AIC score of models using each different encoding for location (10-80-10, 25-50,25, 30-40-30, first-middle-last) then using the genetic algorithm in the `glmulti` package to find which location feature was most significant. Replicating this, I confirmed that the 10-80-10 split was the best encoding to use.

4.2 Control Model Selection

The variables used in the control model selection process were location (using the 10-80-10 split), dependency label, dependency distance, dependency distance to the left (set to 0 if the governor is the right), part-of-speech tag, word length, sentence length and frequency. Considering these 8 variables and and their pairwise interactions gives $2^{2^8} = 1.16 \times 10^{77}$ possible models, justifying the use of the genetic algorithm for the search procedure.

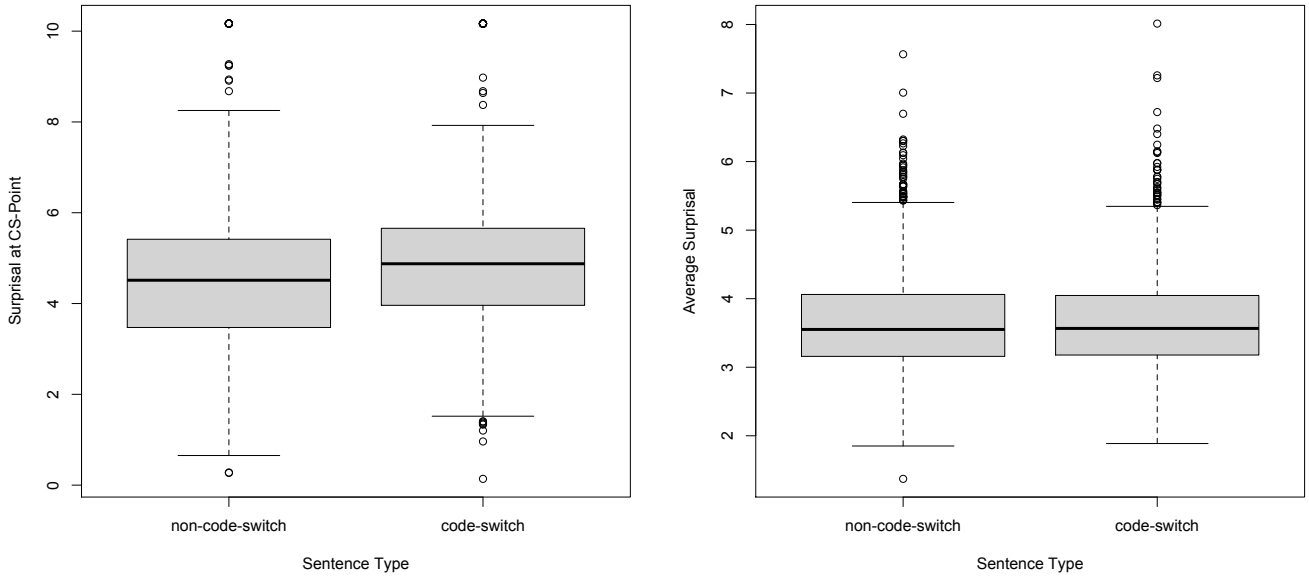
I ran the genetic algorithm 10 times using these variables and their pairwise interactions optimising for AIC. In 6 runs, the algorithm converged to the following model, with an AIC of 3939 and a BIC of 4053:

$$\text{CS} \sim \text{postag} + \text{w\_length} + \text{s\_length} + \text{freq} + \text{w\_length} : \text{depencency\_distance\_length} + \text{location} : \text{s\_length} + \text{postag} : \text{w\_length} + \text{postag} : \text{freq} + \text{dependency\_label} : \text{freq}$$

I then repeated this process, optimising for BIC. The algorithm converged in 7 out of 10 runs to the following model, with an BIC of 4003 and an AIC of 3967:

$$\text{CS} \sim \text{postag} + \text{w\_length} + \text{s\_length} + \text{freq}$$

This is the model that was selected by Calvillo et al.. It contains the same variables as the AIC model, just without the interactions. They chose to use this model as their control model because all variables were found to be significant and they preferred using harsher BIC criterion to remove the interaction terms. I confirmed this, finding that all four parameters were significant (***). As expected, dependency distance, dependency label and location were not selected due to being balanced by alignment. It is odd, however, that the part-of-speech tag is still included in this model, given that the distribution of part-of-speech tags is identical between the two classes due to the alignment process. The authors do not explain this phenomenon, but it may be that the part-of-speech tag can help distinguish code-switched sentences after distinguishing using the other variables.



(a) Surprisal at CS-point.

(b) Average Surprisal of Sentence.

Figure 2: Box plots for various surprisal features provided by the Chinese-English dataset.

4.3 Variables of Interest

Using this control model, I then added surprisal and entropy to see if they improved the quality of the control model. I confirmed that surprisal does improve quality, reducing the AIC from 3967.3 to 3954.5 and the BIC from 4003.2 to 3996.4. This is confirmed by a likelihood ratio test ($\chi^2(1) = 14.81, p < 0.001$). I also confirmed that entropy did not improve the quality of the model, nor does it reach significance when added. The authors find this surprising, but explain that it is likely because entropy is related to the difficulty of selecting any word, rather than being limited by the semantics that they speaker is attempting to convey. I agree with this conclusion.

4.4 Comparing Mutual Information to Surprisal

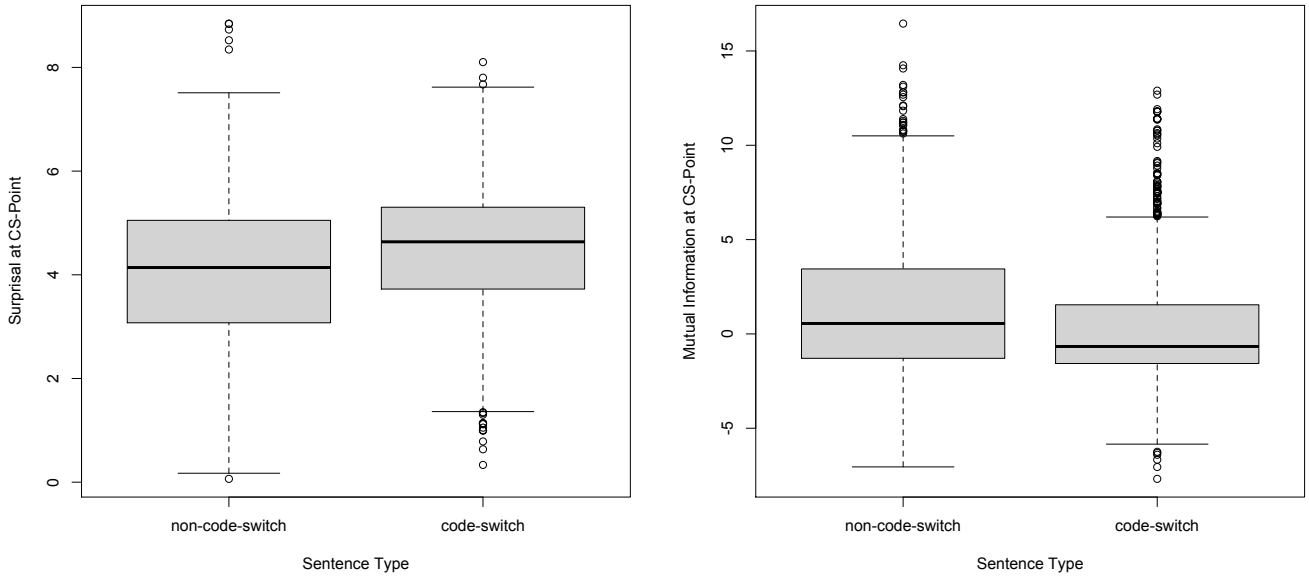
Having confirmed the findings of Calvillo et al., I then modified their R code to conduct similar analyses using the surprisal and mutual information features that I calculated.

Fitting logistic regression models using just the new surprisal feature and mutual information feature, I found both to be significant predictors (***). This is clear given the difference in distribution of these features between the two classes, seen in fig. 3. To compare my calculation of surprisal to the dataset's surprisal values, I fitted a logistic regression using both and found that only my newly calculated surprisal was significant (***). This suggests that my calculation was correct, as I would not expect one of the surprisal features to add any more information over the other. This also suggests that my calculated surprisal feature is a better predictor.

Fitting a model using both new features, I found both surprisal (*) and mutual information (**) to be significant. This suggests that mutual information is not just a good predictor, but may even complement surprisal.

I then added both features to the control model, individually and together. Adding the new surprisal feature improves the quality of the control model, reducing AIC from 3967.3 to 3929.8 and the BIC from 4003.2 to 3971.7. This is in fact a better improvement than the dataset's provided surprisal feature, and is confirmed to be significant using a likelihood ratio test ($\chi^2(1) = 39.47, p < 0.001$). This corroborates the authors' conclusion that surprisal is a good predictor for code-switching.

Adding mutual information to the control model also improves the quality, reducing AIC from 3967.3 to 3899.1 and reducing the BIC from 4003.2 to 3941.0, an even better improvement than the new surprisal feature and confirmed to be significant ($\chi^2(1) = 70.21, p < 0.001$). Adding both mutual information and surprisal to the control model improves the quality even further, reducing the AIC to 3884.2 and the



(a) Surprisal at CS-point.

(b) Mutual Information at CS-point.

Figure 3: Box plots for the newly calculated surprisal features.

BIC to 3932.1, a significant improvement over the model with mutual information ($\chi^2(1) = 16.88, p < 0.001$). This shows that mutual information does act as a significant predictor.

Finally, I examined the correlation between surprisal and mutual information. This gave a Spearman's $\rho = -0.84$ ($p < 0.001$), indicating a strong negative correlation. This makes sense, as they are very similar measures of contextual uncertainty, calculated using the same n-gram. Calvillo et al. carried out a similar analysis, finding that surprisal is correlated with frequency and also with word length. To assess whether this collinearity may have an impact on the quality of the model, they use the Generalised Variance Inflation Factor (GVIF), getting values below 2. They state that any value below 5 indicates that although collinearity may exist, it is not problematic for the model's results. When I calculated the GVIF on the model with my surprisal and mutual information features, all values were below 4.

4.5 Final Code-Switching Prediction Model

The final model incorporating the control variables and the new surprisal and mutual information variables can be seen in table 1. Nearly all parameters retain significance and the coefficients of most parameters agree with the hypothesis that multilinguals code-switch when under cognitive load: words with lower frequency are more likely to be code-switched ($\beta = 0.79$), as are longer words ($\beta = 0.49$) and words in longer sentences ($\beta = 0.62$). This is similar to the final model of the authors. The model also indicate that words with low mutual information with respect to their context have a tendency to be code-switched ($\beta = -0.55$).

Strangely, the final parameter for surprisal in the model that includes mutual information is negative ($\beta = -0.75$), suggesting that words with *low* surprisal are more likely to be code-switched. This is the opposite effect to what the authors observed and to what I observed in the model with surprisal but without mutual information. I believe that this is due to the interaction between mutual information and surprisal: despite the GVIF value of 3.46 (below 5), it may be that surprisal is being used to distinguish between words with low mutual information, oppositely to how surprisal would be used without mutual information present, as a result of the strong negative correlation.

5 Discussion

In this report, I have successfully replicated the work of Calvillo et al. (2020). My results are consistent with the view that multilinguals actively inhibit code-switching, but will stop inhibition when under cognitive load, due to the difficult to retrieve a certain word. These results are also consistent with a

	Parameter Estimates		Wald's Test		Likelihood Ratio Test	
Predictor	Coef.β	SE(β)	Z	p_z	χ^2	p
(intercept)	-0.11	0.05	-2.10	< 0.05		
surprisal	-0.75	0.11	-6.76	< 0.001	16.88	< 0.001
mutual information	-0.55	0.05	-4.07	< 0.001	47.61	< 0.001
frequency	0.79	0.15	5.15	< 0.001	27.75	< 0.001
word length	0.49	0.10	4.90	< 0.001	24.38	< 0.001
sentence length	0.62	0.08	7.71	< 0.001	62.62	< 0.001
POS=verb	0.39	0.11	3.67	< 0.001	13.60	< 0.01
=other	0.12	0.11	1.11	.267		

Table 1: Summary of the logistic regression model after adding surprisal and mutual information: coefficient estimates β with standard error, Wald's z-scores and their significance level, contribution to likelihood χ^2 and its significance level. AIC/BIC before introducing surprisal and mutual information: 3967.3/4003.2; after introducing surprisal and mutual information: 3884.2/3932.1.

comprehender-related factor of code-switching, that a speaker may code-switch to highlight segments of a sentence with high information density in order to emphasise them, as explored by Myslín and Levy (2015).

These results should not be overstated. Firstly, this experiment was conducted only on written, online communication, and so the conclusions may not hold for other forms of communication. Secondly, this model may not be directly useful for real-world applications, as usually NLP systems need to process sentences where code-switching has already occurred, rather than analysing sentences where code-switching may occur. Even for systems that may want to predict code-switching (such as a foreign language tutoring system), the current model still requires knowledge of the CS-point. Further work would be required to compare the surprisal and mutual information at the CS-point to the rest of the sentence for this model to be useful in a real application. Nevertheless, this work is evidence that surprisal and mutual information are useful in producing cognitive models of the factors at play when a multilingual code-switches.

6 Conclusion

I have confirmed the results of Calvillo et al. to find that surprisal, but not entropy, can act as a significant predictor of code-switching beyond other well-known predictors. I confirmed this firstly by re-running their public R code and secondly by re-calculating surprisal using a new language model. I further found that the mutual information of a word with respect to its previous context can also act as a significant predictor of code-switching and that it retains significance when used along with surprisal, despite their strong negative correlation.

References

Aguilar, G. and Solorio, T. (2020). From English to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8033–8044, Online. Association for Computational Linguistics.

Brent, M. R. (1999). An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning*, 34(1):71–105.

Calcagno, V. and de Mazancourt, C. (2010). glmulti: an r package for easy automated model selection with (generalized) linear models. *Journal of statistical software*, 34(12):1–29.

Calvillo, J., Fang, L., Cole, J., and Reitter, D. (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.

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

Eppler, E. D. (2011). The dependency distance hypothesis for bilingual code-switching. In *Proceedings of the International Conference on Dependency Linguistics*, pages 145–154.

Forster, K. I. and Chambers, S. M. (1973). Lexical

access and naming time. *Journal of verbal learning and verbal behavior*, 12(6):627–635.

Gollan, T. H. and Ferreira, V. S. (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.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Hult, F. M. (2014). Covert bilingualism and symbolic competence: Analytical reflections on negotiating insider/outsider positionality in swedish speech situations. *Applied Linguistics*, 35(1):63–81.

Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT press.

Kello, C. T. and Plaut, D. C. (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.

Martin, J. N. and Nakayama, T. K. (2013). *Intercultural communication in contexts*. McGraw-Hill New York, NY.

Mave, D., Maharjan, S., and Solorio, T. (2018). Language identification and analysis of code-switched social media text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 51–61, Melbourne, Australia. Association for Computational Linguistics.

Myers-Scotton, C. (1995). *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

Myslín, M. and Levy, R. (2015). Code-switching and predictability of meaning in discourse. *Language*, pages 871–905.

Poplack, S. (2000). Sometimes i'll start a sentence in spanish y termino en español: Toward a typology of code-switching. *The bilingualism reader*, 18(2):221–256.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (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*, pages 324–333.

Stratos, K. (2018). Mutual information maximization for simple and accurate part-of-speech induction. *arXiv preprint arXiv:1804.07849*.

Tseng, H., Chang, P.-C., Andrew, G., Jurafsky, D., and Manning, C. D. (2005). A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

Weinreich, U. (2010). *Languages in contact: Findings and problems*. Number 1. Walter de Gruyter.

Xu, L., Zhang, X., and Dong, Q. (2020). Cluecorpus2020: A large-scale chinese corpus for pre-training language model.