Singaporean Conversational English-Malay Code-Switching Speech: An Analysis Based on Code-switching Points and Part-of-Speech

Kshitij Gupta*, Chaiyasait Prachaseree†,
Thi Nga Ho†, Kyaw Zin Tun†, Jia Xin Koh†, Ying Ying Tan†, Eng Siong Chng†, Chalapathi GSS*
*BITS Pilani
{f20190212, gssc}@pilani.bits-pilani.ac.in
†Nanyang Technological University, Singapore
{prac0003@e.ntu.edu.sg}{ngaht, ztkyaw, kohj0058, yytan, aseschng}@ntu.edu.sg

***Abstract*—This paper investigates various code-switching properties of conversational speech from bilingual English-Malay Singaporean speakers with data obtained from the National Speech Corpus (NSC) and provides baseline language models for various combinations between English-Malay monolingual and codeswitching transcripts. Specifically, the study analyzed the correlation between code-switching patterns and (i) trigger words and code-switched word pairs at code-switching points, and (ii) wordwise POS and pairwise POS tags. Our analysis shows there is a certain set of words that frequently "triggered" code-switching behavior, and speakers tend to code-switch more frequently around nouns. Additionally, we provide perplexities for language models built on the selected datasets. These perplexities could serve as baselines for future language models for Singaporean speech, especially, English-Malay code-switch speech.**

***Index Terms*—Code-Switching, Language Modeling, Part-of-Speech**

I. Introduction

Code-switching is the act of alternating between two or more languages in a single conversation or text. This may happen within the same utterance or sentence (intra-sentential code-switching) and beyond utterance or sentence boundaries but within the same discourse (inter-sentential code-switching). Code-switching is a common linguistic phenomenon found in multilingual communities, for example, in English and Spanish in the United States [1], Cantonese and English and (more recently) Mandarin in Hong Kong [2], Mandarin and Minnan in Taiwan [3], Malay and English for Malay Malaysians, as well as Mandarin, Malay, English for Chinese Malaysians [4], [5].

In multi-ethnic and multilingual Singapore, studies have increasingly found code-switching, which is often thought to be primarily restricted to informal contexts, to have become a defining feature of Singaporean speech, also occurring in formal domains [6]. This work builds upon the usage of parts-of-speech from [7] work on code-switching for English-Mandarin. However, the English-Malay language pair has not been researched in terms of language modeling and linguistic spectra with respect to the Singaporean community. With the pervasiveness of code-switching in everyday communication, especially for Singaporeans, studying this phenomenon as applied to NLP tasks is essential to further the performance of automatic speech recognition systems when used on natural conversational data.

In addition to speech recognition, code-switch analysis can be useful in tasks such as sentiment analysis, named entity recognition, and machine translation. For example, by identifying and analyzing code-switched expressions in text, sentiment analysis models can better capture the nuanced emotions and attitudes of the speakers; also named entity recognition models can better recognize and classify named entities across languages. Furthermore, in machine translation, code-switch analysis can be used to identify and translate code-switched expressions, leading to more accurate and natural-sounding translations.

This paper investigates English-Malay code-switching points on a subset of the National Speech Corpus (NSC) published by the Info-communications and Media Development Authority of Singapore. The study focuses on two aspects: (i) Code-Switching points and (ii) Part-of-Speech (POS). The work is motivated by linguistic theories that code-switching points within a sentence are not completely random, and certain grammatical characteristics may predict the switches [8]. Thus, we used exploratory analysis on demographic-based data and POS tagging to conclude specifically for Singaporean English-Malay speakers.

The rest of this paper is organized as follows: Section II provides an overview of the selected English-Malay code-switching corpus as provided as part of the NSC corpus. Section III presents an analysis of the frequencies, trigger words, and word pairs of code-switching points. The subsequent section, Section IV, presents the analysis based on POS tags. The results of the baseline language modelling are discussed in Section V. Finally, Section VI presents the conclusion and future work.

II. Corpus Overview

The NSC corpus [9] consists of both audio and transcripts, along with a range of demographic information for each

979-8-3503-3078-6/23/$31.00

speaker. Moreover, the language tags are provided whenever there is a codeswitch from one language to another, thus code-switched sentences can be determined. A selected 200-hour subset of conversations obtained from English-Malay participants from NSC corpus, which will be referred to in the rest of the paper as NSC-ENGMAL-ALL, was used for the analysis. NSC-ENGMAL-ALL consists of a total of 206,714 utterances, with 132,558 (64.1%) English-only utterances and 7,967 (3.85%) Malay-only utterances. The remaining 66,189 (32%) utterances are code-switching English-Malay utterances. Of the code-switching utterances, 21,653 (32.7%) start with Malay, while the rest, 44,536 (67.3%), start with English. As this work focuses primarily on code-switching, we will refer to utterances with intra-sentential code-switching as NSC-ENGMAL-CS. While monolingual English and monolingual Malay utterances are NSC-ENGMAL-ENG and NSC-ENGMAL-MAL, respectively.

In the corpus, the number of utterances containing only English significantly outnumbers the code-switched and Malay-only utterances. Additionally, monolingual Malay utterances were extremely short, having only slightly less than three Malay words per utterance as shown in Table I. In contrast, code-switched utterances were longer sentences on average, having almost fifteen words per utterance.

TABLE I: Overview of utterances in NSC-ENGMAL-ALL

Utterances	# Sentences	Avg. Length	Std Dev Lengths
English	132,558	6.99	8.54
Malay	7,967	3.11	2.63
CS	66,189	14.68	13.90
All	206,714	9.31	11.09

III. Analysis based on Code-Switching Points

This section presents our analysis of possible patterns that happen around a code-switching point, specifically words that trigger code-switching behavior and code-switched word pairs that frequently appear together. NSC-ENGMAL-CS has a total of 187,429 code-switching points in the corpus, with fairly equal English to Malay switching frequency at 94,117 and Malay to English at 93,252. When normalized over 66,189 utterances, the switching frequency per sentence is 1.42 and 1.41 for English to Malay and Malay to English, respectively. The average length for consecutive words of the same language of English phrases is 4.78 words, in Malay phrases 2.42 words, and in other languages 1.15 words per phrase.

A. Words Triggering Code-Switching

In this section, words that appear in front of the language switches were examined. Only words with frequencies higher than 50 that appear before the switch were selected; the words with the highest ratios that indicate a switch are shown in Table II and Table III for English to Malay and Malay to English respectively. For each trigger word, the frequency and the ratio it appears in NSC-ENGMAL-CS are also presented, as well as the average number of words in the same language preceding the trigger word. The average length and Standard dev length represent the number of words for both English and Malay in Table II and III.

Some examples in the data echo previous findings regarding the role of code-switching in adding emphasis to an utterance. For instance, speakers in the data are observed to use "confirm" to emphasize their next word, which has a switch ratio to Malay of more than 50 percent. This is exemplified in the following sentences from the corpus, with "ah" in (2) being a discourse particle:

1) *"the price confirm makin mahal"*
   The price is definitely more expensive.
2) *"confirm lawa gila ah"*
   Definitely crazy beautiful ah

TABLE II: English Trigger Words that Switch to Malay

Word	Switch Freq.	Switch Ratio	Avg. Length	Std Dev Length
confirm	170	0.53	2.70	3.86
toilet	53	0.39	2.06	2.50
post	50	0.38	2.32	2.82
movie	70	0.36	2.44	2.75
colour	71	0.36	2.01	2.60
order	72	0.35	1.88	2.07

TABLE III: Malay Trigger Words that Switch to English

Word	Switch Freq.	Switch Ratio	Avg. Length	English Meaning
iya	84	0.82	1.49	yes
takpe	126	0.64	1.49	Never-mind
merepek	121	0.62	1.99	nonsense
punya	3093	0.62	2.43	have
kesian	117	0.61	1.57	pity

B. Frequent Code-Switched Word Pairs

Frequent word pairs of different languages appearing in NSC-ENGMAL-CS as shown in Table IV and Table V, both from English to Malay and Malay to English, were explored. Similar to the previous section, only word pairs appearing more than 50 times in the corpus were ranked according to the ratio. We present here only content words and not those with filler words or discourse particles.

TABLE IV: English to Malay Word Pairs

English Word	Malay Word	Corpus Freq.	Switch Ratio
then	lepas (off)	171	0.93
I	rasa (feel)	168	0.52
so	bila (when)	60	0.43
one	yang (which)	87	0.42
time	aku (me)	113	0.34

IV. Analysis based on Part-of-Speech (POS)

POS tagging is the task of classifying each word of a given text sentence into its grammatical categories. Most languages and dialects have similar or comparable part-of-speech tags: notably having nouns, verbs, adjectives, and adverbs.

TABLE V: Malay to English Word Pairs

Malay Word	English Word	Corpus Freq.	Switch Ratio
kat (at)	Singapore	67	4.00
kalau (if)	let's	51	2.54
nak (son)	try	50	1.15
cakap(talk)	okay	68	0.52
aku (me)	just	182	0.32

Part-of-speech (POS) analysis can be used in code-switching research to identify patterns and patterns of usage of certain word classes, such as nouns, verbs, adjectives, and adverbs. This allows for a more fine-grained analysis of language use and code-switching patterns. [10] and [11] have provided evidence in their respective works that incorporating part-of-speech (POS) tags into language models can yield benefits for code-switching modelling and result in a decrease in perplexity for language modelling tasks. For example, researchers can examine whether certain types of words are more likely to be code-switched than others or if there are any differences in the frequency of code-switching across different POS categories.

In this work, we used NLTK [12] perceptron tagger and Malaya [13] library for the part-of-speech analysis. Parts of speech in both English and Malay were distributed into 12 categories: *Noun, Pronoun, Verb, Adverb, Adjective, Conjunction, Determiner, Particle, Number, Symbol, Other*.

A. Wordwise POS Analysis

This subsection shows the analysis considering the POS of each word in conversations. Figure 1 shows an example of wordwise POS of a code-switched utterance. The following text roughly translates to "*Let me guess you are here for work*".

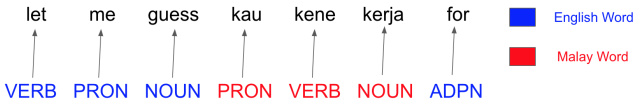

Fig. 1: Wordwise POS Example

For our analysis, we have employed a metric known as the **Malay Word Rate** to analyze code-switching behaviour. The metric is defined as follows:

$$\textbf{Malay Word Rate} = \frac{\textbf{\# Malay Words}}{\textbf{\# Total Words}} * \textbf{100}$$

This metric enables us to quantify the frequency of Malay word usage in the code-switched sentences and facilitates the comparison of code-switching behaviour across different part-of-speech.

Figure 2 shows the results of the Word-by-Word Part-of-Speech analysis on code-switched NSC-ENGMAL-CS for POS having significant data. The analysis reveals that among the different parts of speech, the category of *Particle* exhibits the highest *Malay Word Rate*, indicating that words belonging to this category are more likely to be substituted with their Malay counterpart during code-switching. Particles are a category of words that fall outside of the eight traditional parts of speech. They are frequently used as prepositions in conjunction with other words to form phrasal verbs, which makes them very likely to occur at code-switching points. The category of *Conjunction* also shows a relatively high Malay Word Rate. On the other hand, the category of *Determiner* exhibits the lowest Malay Word Rate, indicating that words belonging to this category are less likely to be substituted with their Malay counterpart and are instead predominantly used in English.

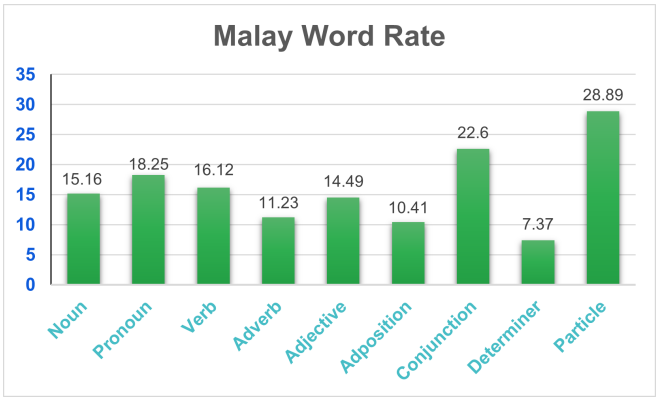

Fig. 2: Malay Word Rate for POS

B. Pairwise POS Analysis

Since code-switching is not a rule-based phenomenon and can vary among individual speakers, relying solely on simple word-wise part-of-speech analysis may not provide enough information to identify code-switching patterns accurately. Hence, Wordwise POS analysis is insufficient for obtaining additional information on these patterns, as code-switching is highly dependent on its context. We obtain more about context using the POS of both words at the code-switching point in a code-switched utterance. Figure 3 shows an example of pairwise POS, where there are two code-switching points:

- **English to Malay:** "*guess*" → "*kau*"
- **Malay to English:** "*kerja*" → "*for*"

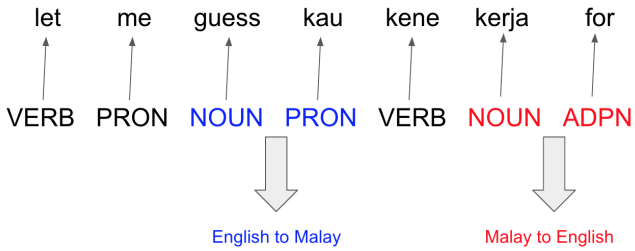

Fig. 3: Pairwise POS Example

This analysis is more accurate in understanding the code-switching patterns in Singaporean Malay speakers using Part-of-speech. We have made the code and detailed results publicly

available on our GitHub repo[1]. We quantified this analysis using the following metric for English to Malay and Malay to English.

$$\textbf{Conversion Rate} = \frac{\textbf{\# Code-switched Instance}}{\textbf{\# Total Instances}} * \textbf{100}$$

Figure 4 and Figure 5 show the top results of English to Malay and Malay to English, respectively. In these results ***POS*** || ***POS*** represents ***Part-of-Speech of first word*** || ***Part-of-Speech of second word***.

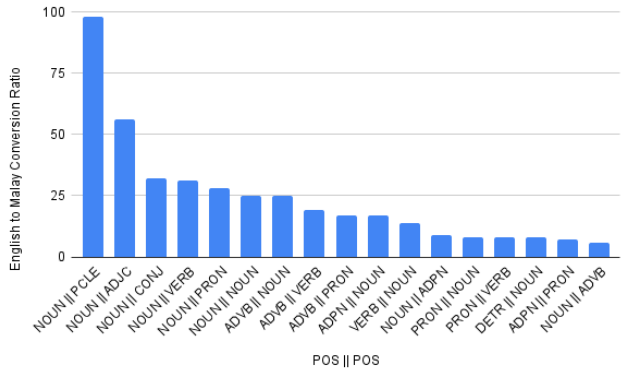

Fig. 4: English to Malay Code-switching Results

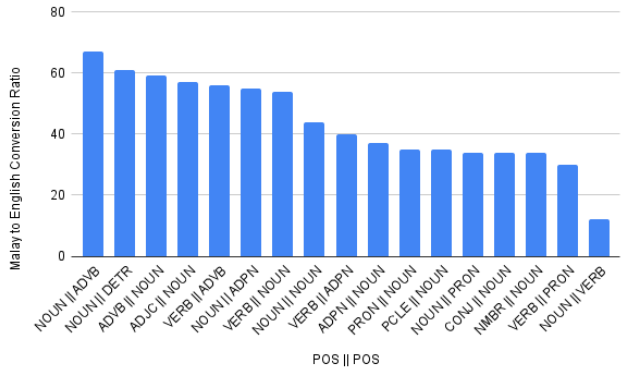

Fig. 5: Malay to English Code-switching Results

Based on this analysis, it can be concluded that Singaporean speakers tend to code-switch more frequently around *Nouns* and *Adjectives*. This analysis can generate more realistic code-switch data, which is closer to conversational data for conversational agents, chatbots, virtual assistants, etc. It can also be used for various language modellings tasks, such as machine translation, speech recognition, and natural language understanding.

V. Language Model Baselines

This section highlights the perplexity performance on the test set using different combinations of the NSC-ENGMAL-ENG, NSC-ENGMAL-MAL, and NSC-ENGMAL-CS training partitions. The subcorpus is partitioned with 80 percent for training and 20 percent for evaluation. The n-grams were trained using SRILM [14] toolkit with an order of 4 and Kneser-Ney smoothing.

In Table VI, a comprehensive list of LM abbreviations is provided, which aids in understanding the Language Modeling results showcased in Table VII. It is worth noting that the perplexity score on code-switched texts obtained solely from the analysis of Code-switched (CS) texts is on par with the perplexity score achieved when incorporating a combination of CS texts, Monolingual Malay, and English texts. However, the differences in English and Malay test texts were substantial.

TABLE VI: LM Abbreviations

LM Abbreviation	Training Texts
LM-CS	CS texts only
LM-ENG	English texts only
LM-MAL	Malay texts only
LM-CS-ENG	CS and Mono English texts
LM-CS-MAL	CS and Mono Malay texts
LM-ENG-MAL	Mono English and Mono Malay texts
LM-CS-ENG-MAL	All training texts

This implies that the utilization of CS texts alone, without the inclusion of additional Monolingual Malay and English texts, yields comparable perplexity scores. The perplexity score is a measure of how well a language model predicts the next word or sequence of words in a given text. Therefore, this finding indicates that the CS texts possess sufficient linguistic context and information to achieve results comparable to those obtained when combining multiple language sources. This can be inferred that the CS texts in isolation are capable of providing a robust foundation for language modeling tasks. This observation may be particularly relevant in scenarios where the availability of Monolingual Malay and English texts is limited or when the focus is specifically on code-switching phenomena. Hence, the findings support the notion that CS texts alone can serve as a valuable resource for language modelling and exhibit comparable performance when compared to a combination of CS, Monolingual Malay, and English texts. Overall, the inclusion of more texts reduces perplexity on the test partitions, but the portion of CS texts reduces when adding monolingual texts.

TABLE VII: Language Modeling Results

LM Data	CS	ENG	MAL
LM-CS	159.8	230.7	202.6
LM-ENG	1243.1	103.9	13535.7
LM-MAL	9526.5	29468.2	57.0
LM-CS-ENG	161.4	111.9	275.8
LM-CS-MAL	159.7	240	105.4
LM-ENG-MAL	448	105.3	134.5
LM-CS-ENG-MAL	157.5	109.8	127.2

VI. Conclusion and Future Work

This work has shown that there exists a set of "trigger" words that frequently appear as code-switching points in

[1]https://github.com/kjgpta/NSC-Code-Switch-Analysis.

a codeswitching conversation. For example, 53% of words after the word "confirm", a word Singaporeans use to show higher significance, switch to Malay. The work also confirmed that Part-of-Speech influences code-switching frequencies in Singaporean English-Malay code-switching speech. The accuracy of code-switched sentence parsing is limited by the performance of part-of-speech taggers(NLTK [12] perceptron tagger and Malaya [13] library) on conversational text. POS taggers are typically trained on formal sentences, which differ from conversational text. For example, our analysis has shown that phrases ending in nouns have the highest frequency of switching from English to Malay and Malay to English. The analysis further confirms linguists' theory that language switches should follow some syntactic rules and not by random alternation.

One of the possible future work is to utilize these factors in downstream linguistic and NLP tasks. The analysis output could also be an insight into the domain of code-switching data generation, which is currently in urgent need of improving ASR and spoken language processing for the codeswitching languages. Another possible usage is to follow a similar pattern of this particular corpus to generate more natural code-switching English-Malay data. This corpus can be used for a variety of multilingual language modeling tasks, such as machine translation and summarization, especially in conversational domains where there is limited data available, not only for English-Malay but also for other low-resource language pairs. Furthermore, the baseline LM models presented in this work may be used for future comparison of language models for Singaporean English-Malay code-switching speech.

ACKNOWLEDGMENT

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE2019-T2-1-084). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

We would like to acknowledge the High Performance Computing Centre of Nanyang Technological University Singapore, for providing the computing resources, facilities, and services that have contributed significantly to this work

REFERENCES

[1] J. G. Cox, A. LaBoda, and N. Mendes, ""i'm gonna spanglish it on you": Self-reported vs. oral production of spanish–english codeswitching," *Bilingualism: Language and Cognition*, vol. 23, no. 2, pp. 446–458, 2020. DOI: 10.1017/S1366728919000129.

[2] K. L. R. Chan, "Trilingual code-switching in hong kong," *Applied Linguistics Research Journal*, vol. 3, Sep. 2019. DOI: 10.14744/alrj.2019.22932.

[3] C.-H. Hsiao, "Code-switching between typologically similar languages: Data from mandarin-taiwanese code-switching," Ph.D. dissertation, Jun. 2022.

[4] M. David, W. Yee, Y. Ngeow, and G. Hui, "Language choice and code switching of the elderly and the youth," *International Journal of The Sociology of Language*, vol. 2009, pp. 49–74, Nov. 2009. DOI: 10.1515/IJSL.2009.044.

[5] M. Hadei, "Social factors for code-switching-a study of malaysian-english bilingual speakers," *International Journal of Language and Linguistics*, vol. 4, p. 122, Jan. 2016. DOI: 10.11648/j.ijll.20160403.15.

[6] W. Bokhorst-Heng and I. Caleon, "The language attitudes of bilingual youth in multilingual singapore," *Journal of Multilingual and Multicultural Development - J MULTILING MULTICULT DEVELOP*, vol. 30, pp. 235–251, May 2009. DOI: 10.1080/01434630802510121.

[7] G. Lee, "Cross-lingual language modeling: Methods and applications," Ph.D. dissertation, National University of Singapore (Singapore), 2021.

[8] C. Lee, "Motivations of code-switching in multi-lingual singapore," *Journal of Chinese Linguistics*, vol. 31, pp. 145–176, Jan. 2003.

[9] J. X. Koh, A. Mislan, K. Khoo, *et al.*, "Building the singapore english national speech corpus," in *Interspeech*, 2019.

[10] H. Adel, N. T. Vu, K. Kirchhoff, D. Telaar, and T. Schultz, "Syntactic and semantic features for code-switching factored language models," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 3, pp. 431–440, Mar. 2015, ISSN: 2329-9290. DOI: 10.1109/TASLP.2015.2389622. [Online]. Available: https://doi.org/10.1109/TASLP.2015.2389622.

[11] G. I. Winata, A. Madotto, C.-S. Wu, and P. Fung, "Code-switching language modeling using syntax-aware multi-task learning," in *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 62–67. DOI: 10.18653/v1/W18-3207. [Online]. Available: https://aclanthology.org/W18-3207.

[12] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[13] Z. Husein, *Malaya-speech*, https://github.com/huseinzol05/malaya-speech, Speech-Toolkit library for bahasa Malaysia, powered by Deep Learning Tensorflow, 2020.

[14] A. Stolcke, "Srilm - an extensible language modeling toolkit.," in *INTERSPEECH*, J. H. L. Hansen and B. L. Pellom, Eds., ISCA, 2002. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2002.html#Stolcke02.