Analyzing Triggered Code-switching on a Marathi-English Code-Mixed Corpus

Advait Sankhe

Institute of Artificial Intelligence, The University of Georgia – ams57464@uga.edu

Abstract

This paper tests Clyne's original triggering hypothesis on the low-resource Indian language, Marathi. The hypothesis suggests that code-switches occur due to the knowledge of certain cognates, causing the activation of the non-selected language immediately before or after certain trigger words. The hypothesis is tested out and shown to be partially true on a twitter-based, code-mixed corpus through statistical analysis. It also brings to light new types of trigger words that may exist in an online context.

1 Introduction

Code-Switching is the process of switching between multiple languages within the context of a single conversation or even a sentence. The phenomenon of code-switching is widely observed among bilingual speakers throughout the world and has intrigued linguists, psychologists and anthropologists alike. The way that one language flows into the next, seemingly conforming to the structure of both languages rather effortlessly is a fascinating way of studying how different languages are organized in the mind.

An interesting topic to be explored then is, why do bilinguals code-switch? A commonly cited reason is that bilinguals seem to lack proficiency in a single language. Heredia et al. (2001) address certain problems with this point of view. Three major problems were brought up. The first is that the frequency of certain words used in the context of one language is much higher than that of the other, indicating that its just easier to retrieve those words in the context of that language. Secondly, this does not explain how code-switched words are adapted perfectly to the grammatical structure of either language. Lastly, how is proficiency defined? Most bilinguals receive their education in a single language (commonly English). This creates a disparity between the reading/writing proficiency versus the spoken proficiency of the languages.

A more reasonable hypothesis is that the knowledge of certain cognates triggers a code-switch. Michael Clyne, in his original triggering hypothesis (1967), suggested that code-switching is facilitated by the anticipation or consequence of the utterance of certain trigger words (words that have similar form and meaning in both languages). This hypothesis was first statistically tested by Broersma et al. (2006) using a Dutch – Moroccan Arabic speech corpus. They also suggested an adjusted triggering hypothesis, based on modern speech production theory.

This paper aims to test the triggering hypotheses in Marathi-English code-switching. Marathi is an Indian language, primarily spoken in the state of Maharashtra with approximately 99 million speakers in the world. Marathi is still considered to be a low-resource language (Pingle et al., 2023), as is commonly seen in languages of developing countries, due to a lack of corpora

and morphological analyzers. It serves as an interesting base to study code-switching, due to a rapidly increasing trend of English-based education in the state. With Hindi also being commonly spoken in the northern parts of the country, many native Marathi speakers grow up bilingual or trilingual. Thus, code-switching is frequently observed in day-to-day conversations between Marathi speakers. To my knowledge, this is the first paper to explore the triggering hypotheses for Marathi. Since the corpus used in this study consists of tweets, it also opens up certain possibilities for what might be considered to be trigger words in online discourse.

2 Background

2.1 Clyne's Triggering Hypothesis

Clyne, in his original triggering hypothesis, put forth the idea that code-switching occurs due to the presence of trigger words (words that are common to two or more of the speaker's languages). These words share similar form or meaning in the two languages. He suggests that such words cause the speaker to lose their linguistic orientation, resulting in a code-switch either immediately after or in anticipation of such words. Trigger words are split into three types, 1) Lexical transfers (aka loanwords), 2) Bilingual homophones, 3) Proper nouns. This study was done in 1967, based on tapes of German-English and Dutch-English bilinguals in Australia. No statistical analysis was performed.

2.2 Broersma's Adjusted Triggering Hypothesis

Broersma's Triggering hypothesis builds upon the original hypothesis based on knowledge about modern speech production theory. They suggest that the language selection of lemma occurs before the surface structure (the final order in which the words will be uttered) is formed. The presence of a trigger word does not necessarily mean that a word adjacent to it will be code-switched. Instead, they posit that language-specific lemmas are chosen at the lexical selection stage. Thus, the effect of trigger words is observed at the smallest processing unit, i.e., the basic clause. They studied the effects of the presence of trigger words in basic clauses and also performed the first statistical analysis of

the original triggering theory. The analysis was done using a Moroccan Arabic-Dutch corpus.

3 Methodology

3.1 MeLID Corpus

The MeLID dataset (Tanmay et al., 2023) was used as the corpus for this study. It consists of 11813 code-mixed tweets where each word is tagged as English, Marathi or 'Other'. The 'Other' category consists of mainly of proper nouns and numbers. The dataset was developed for Language Identification tasks for Machine Learning Models. It is a subset of a larger dataset that was developed for training different types of ML models. It is unclear what percentage of these tweets were written in the original Devanagari script for Marathi, but the final dataset contains the transliterated forms of all the tweets. Punctuation and twitter @USER tags were removed from the tweets.

3.2 Pre-processing

Most of the pre-processing and querying was done in Python. The *pandas* library was used to group together tweets, get spans, determine adjacency, create contingency tables, perform chi-square analysis, etc. All words tagged as "OTH" were considered to be trigger words. The *scikit-learn* library was used to perform logistic regression for feature importance. Figures were created using *matplotplib*.

3.3 Testing Clyne's Triggering Hypothesis

To test Clyne's triggering Hypothesis, a methodology similar to what was done by Broersma et al. was adopted. Any word in a tweet that belongs to a different language than the word preceding it was tagged as a code-switch. Single lexical items also constitute a code-switch. Based on the arguments made by the authors, lexical transfers were not considered to be trigger words, as, for an individual speaker, it is impossible to determine which English words are part of the Marathi lexicon and vice versa. Thus, the assumption made here is that the creators of the corpus have tagged such words to belong to the language that is accepted by the larger community. Due to the transliterated

nature of the corpus and the vastly different phonetic systems of the two languages, identifying bilingual homophones was an infeasible task and were not considered to be trigger words. Although this is a valid limitation, based on the examples and observations by Broersma et al., a large majority of the trigger words are indeed proper nouns. Trigger words in this study therefore mainly consist of proper nouns and numbers (the case for the latter is made in the further sections).

4 Results and Discussion

4.1 Triggering Hypothesis Observations

Based on the above-mentioned methods, the following results were seen.

Table 1: Number of Code-switched words that precede (and do not follow) a Trigger word

		Precedes a Trigger Word	
		No	Yes
Code-Switch	No	106876	5543
	Yes	22746	1225

The results show that the likelihood of a code-switch preceding a trigger word is nearly identical to code-switches that do not border trigger-words ($\chi^2 = 1.35$, $p = 0.85$).

Table 2: Number of Code-switched words that follow (and do not precede) a Trigger word

		Follows a Trigger Word	
		No	Yes
Code-Switch	No	107450	4969
	Yes	21387	2584

"*Sarkar aapli madat pahije*, win8 update system is complete failure for me"

[Sir, I need your help, win8 update system is a complete failure for me]

The results here show that words that follow trigger words are much more likely to be code-switched than words that do not border a trigger word (χ^2 = 1527.60, p << 0.0001).

Table 3: Number of Code-switched words that border a Trigger word on both sides vs words that are followed by Trigger words

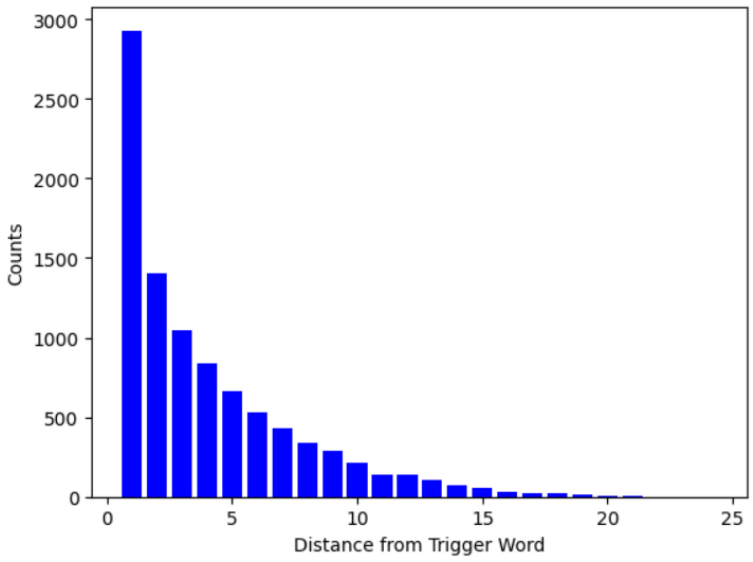
		Bordered by Trigger words on both sides	
		No	Yes
Code-Switch	No	4969	570
	Yes	2584	345

For words that were followed by a trigger word, the additional influence of preceding a trigger word does not seem to affect its likelihood of being code-switched (χ^2 = 4.40, p = 0.35). Through all these observations, there is no evidence of anticipational triggering. These observations align with those seen by Broersma et al.

4.2 Activation Effect of Trigger Words

The activation threshold hypothesis (Paradis, 1998) suggests that the activation of a lexical element decreases over time. This suggests that the probability of a code-switch occuring is inversely proportional to its distance from the nearest trigger word.

Figure 1: Distance of code-switched words from the nearest Trigger word (only appearing after Trigger word utterance)



4.3 Numbers as Trigger Words

The online, written context of this corpus brings up two unique types of lexical elements that were not mentioned in the original hypothesis, but do fit the definition of a trigger word. Numbers have similar form and meaning in both languages.

Table 4: Code-switched words that precede a number

		Precedes a number	
		No	Yes
Code-Switch	No	111492	927
	Yes	23757	214

Table 5: Code-switched words that follow a number

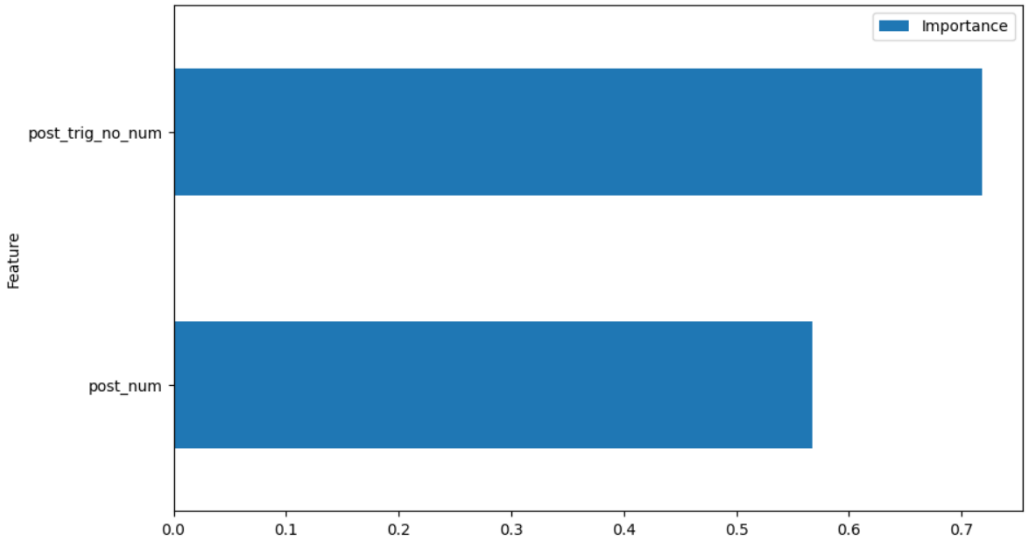
		Follows a number	
		No	Yes
Code-Switch	No	111543	23661
	Yes	876	310

"*Ho na* **2** sessions *khelayla pahije hote*"

[Yeah, 2 sessions should've been played]

Table 4 (χ^2= 1.10, p = 0.89) and Table 5 (χ^2 = 60.55, p << 0.0001) show similarities with the previously seen trigger word observations.

Figure 2: Logistic Regression Feature Importance for predicting code-switch (post non-number trigger word, post number)



5 Conclusion and Limitations

The study done here shows that Clyne's Hypothesis is partially true for Marathi-English code-switching, two vastly different languages. It also adds merit to the original definition of a trigger word, introducing numbers as a possibility to consider in modern, informal, text-based discourse. This is also the first time this hypothesis was tested out on the low-resource Marathi language. The corpus however, is limited to short, disconnected tweets. The lack of long-form conversations does not allow for testing of code-switching metrics that analyze the span of a code-switch. The absence of annotations also makes it difficult to perform clause-based analyses, as done in studies of other higher-resource languages. As the number of Marathi-English bilingual speakers grows, better corpora, allowing for further research in the domain will surely help in understanding the mental processes behind language production in speakers.

References

Heredia, R. R., & Altarriba, J. (2001). Bilingual Language Mixing: Why Do Bilinguals Code-Switch? Current Directions in Psychological Science, 10(5), 164-168. https://doi.org/10.1111/1467-8721.00140

Clyne, M. G. (1980). Triggering and language processing. *Canadian Journal of Psychology / Revue canadienne de psychologie, 34*(4), 400–406. https://doi.org/10.1037/h0081102

Broersma, Mirjam & De Bot, Kees. (2006). Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. Bilingualism: Language and Cognition. 9. 1-13. 10.1017/S1366728905002348.

Pingle, A., Vyawahare, A., Joshi, I., Tangsali, R., Kale, G., and Joshi, R., "Robust Sentiment Analysis for Low Resource languages Using Data Augmentation Approaches: A Case Study in Marathi", arXiv e-prints, 2023. doi:10.48550/arXiv.2310.00734.

Tanmay Chavan, Omkar Gokhale, Aditya Kane, Shantanu Patankar, Raviraj Joshi, "My Boli: Code-mixed Marathi-English Corpora, Pretrained Language Models and Evaluation Benchmarks", arXiv e-prints, 2023. https://doi.org/10.48550/arXiv.2306.14030

Paradis, M. (1998). Acquired aphasia in bilingual speakers. In M. Sarno (ed.): Acquired aphasia (3d ed.), 531–549. San Diego: Academic Press.

Mario Filho, « How To Get Feature Importance In Logistic Regression » https://forecastegy.com/posts/feature-importance-in-logistic-regression