ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

JOURNAL OF APPLIED LINGUISTICS AND TESOL

Vol.8. No.3.2025

CODE-SWITCHING IN MULTILINGUAL DIGITAL TEXTS: AUTOMATED DETECTION AND LINGUISTIC PATTERNING THROUGH AI-BASED CORPUS ANALYSIS

Muhammad Nusrat
Corresponding Author
University of Education Lahore
Email: bobydaultana@gmail.com
Aqsa Shereen
Assistant professor in English at Women University Swabi
Email: aqsaharoon333@gmail.com
Sumera Bhanbhro
Assistant Professor at Institute of English Language & Literature, University of Sindh, Jamshoro.
Hira Abbas
Researcher, Department of Linguistics and Literature, Karakoram International University Gilgit.
Email: Hiraabbas1996@gmail.com
Corresponding Author: bobydaultana@gmail.com

Abstract
In this study, the authors explore whether machine learning-based corpus/pattern recognition of code switching in multilingual digital writings is feasible and to monitor the extent of code switching between these five languages in multilingual contexts (English-Hindi, English-Spanish, English-Arabic, English-Tagalog, and English-Malay). Based on a large-scale annotated corpus and transformer-based architecture fine-tuned on the multilingual setting, the study delivered token-level accuracies of above 95 percent and macro F1 scores of over 0.94 both in the in-domain and out-of-domain assessment. The analysis elicited consistent part-of-speech triggers where the nouns, verbs, discourse markers, were common occurrence at switch point, and usages of syntactic choices were realized by focusing on switch at noun phrase-verb phrase boundaries. The fact that high bilingual collocations ranks also demonstrated that formulaic expressions were robust indicators of switches. The reliability of annotation was verified using Cohen Kappa values that exceeded 0.86 and hyperparameter tuning showed that long-distance switching dependencies can be captured using longer sequence lengths. Not simply re-assuring the powerfulness of AI-guided models to reflect a human-level of code-switch detection, the results also contribute to theoretical knowledge in terms of structural, pragmatic, as well as sociolinguistic aspects of cross-linguistic contact when occurring in online contexts of communication.
Keywords: *Code-switching, multilingual NLP, corpus analysis, transformer models, syntactic boundaries, bilingual collocations, annotation reliability, AI-based language detection.*

1. Introduction

In our ever-interconnected global society, online communication tools (Twitter, WhatsApp, Instagram and online forums) now present dynamic spaces where multiple languages and bilingual speech can be expressed. In these locations, the code-switching phenomenon, which entails switching back and forth between two or more languages or varieties of a language in the same

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

discourse has come out to be a very common and evolving form of communication. Traditionally thoroughly investigated in the spoken sphere (e.g. Poplack, 1980; Myers-Scotton, 1993), the code-switching today appears prominently in written online texts, and this is due to the informal nature of these texts, to linguistic identity, and to digital media affordances (Auer, 1998; Garc, 1999).

Computationally speaking, the problem of code-switching in digital text is in some ways distinctive. Long established language id systems that have been developed to work on monolingual units at the sentence or document level of language are notoriously known to miss abrupt language shifts at the token level, particularly in short high noise informal online communication (Lui & Baldwin, 2012; Solorio & Liu, 2008) . Such systems which generally base themselves on lexical characteristics or n-gram probabilities do not possess the contextual specificity to successfully annotate tokens in cross-linguistic situations. The consequence is that misleading classification at token level may percolate into the failure of downstream natural language processing (NLP) tasks that perform tokenization, part-of-speech tags, dependency parsings, and sentiment analysis (King & Abney, 2013; Betts et al., 2017)].

There are now alternative prospects in the form of multilingual pretrained transformer models, notably multilingual BERT (TBERT; mBERT), XLM-R, and additional. They have deeper contextual embeddings and subword tokenization that allows them to conduct stronger language discrimination, even in the case that words are morphologically or orthographically similar between languages (Devlin et al., 2019; Conneau et al., 2020). Some preliminary experiments have found that it is possible to achieve good performance in code-switch detection by fine-tuning such models on the task of token classification (Pires et al., 2019; Winata et al., 2019) . Still, these studies usually are limited by a small number of language pairs (e.g., Spanish-English) or standardized corpora, so the total range of digitally mediated code-switching remains under-researched.

In addition to detection, it is also extremely valuable to discover the linguistic regularities that define switching points, i.e., also syntactic patterns, parts of speech identities, discourse locators, and pragmatic functions, which define how a speaker can switch languages. Structural types of switching (intersentential, intrasentential, intra-word) and constraints such as the Equivalence Constraint and the Matrix Language Frame Model are commonly identified in sociolinguistic research (Poplack, 1980; Myers-Scotton, 1993; Auer, 1998) ; computational analyses of such patterns in real-life large-scale digital corpora, however, are still in their infancy (Chaudhari et al., 2019; Mathur et Offering a systematic, corpus based method, it is possible to identify patterns available in switching phenomenon; these patterns are defining factors of switching like noun insertion, discourse particles, and pragmatic emphasis among others that bring about localization in the switching behaviors.

Also, digital code-switching may overlap with transliteration processes, in which a speaker in one script (e.g., the Latin alphabet) writes a language traditionally written in another script (e.g. Romanized Urdu or Arabic). This introduces an orthographic complexity of making it even more difficult to detect and analyze (Elfardy & Diab, 2013; Pratapa et al., 2018). Ideally, model-based (strong modeling, to identify romanized words) and language-based (insight into how orthographic design responds to motivations of code-switching) knowledge are needed.

Hence, the proposed study will fulfill two objectives: (1) to design and test AI-based approaches of identifying tokens to carry out code-switching in electronic text by comparing the previously used lexical/sequence-model models to transformer-based systems and (2) employ a corpus-based analysis of writings surrounding switch points in digital text, such as syntactic alignment, POS

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

distributions, discourse markers, and the transliteration effect. Integrating the detection accuracy and interpretability, the study aims to guide the further development of better multilingual NLP systems and advance some theoretical models of digital code-switching.

2. Literature Review

2.1 Foundations of Code-Switching Theory

Initial theoretical beginnings of code-switching (CS) set the stage on the social and structural impetus of code-switching. Gumperz (1982) introduced discourse perspectives, where the main focus was made on contextualization cue and pragmatic triggers of switching . Fishman (1965) tracked the sociocultural use of bilingual language and positioned the alternation in the context of the CS in bilingual communities . A more recent approach (Blom and Gumperz, 1972), would suggest that contextualization is done on the basis of indexical markers that demonstrate how slight changes indicate either shift of stance or shift of speaker focus . The above works highlight that CS is patterned, rather than arbitrary, responsive to discourse and engineered interactional requirements.

2.2 Structural and Constraint-Based Models

Syntactically, Constraint-based theories like the Equivalence Constraint (Poplack, 1980 — reprised above) have been added into the Matrix Language Frame (MLF) model (Myers-Scotton, 1993, op. cit.). On the basis of the above, Joshi (1985) interpreted theoretical formalisms to model allowable switch points regarding phrase structure assumptions The extensions of this attribute included the empirical analysis of morphosyntactic constraints among Spanish-English bilingual corpora, which was conducted by Belazi, Rubin & Toribio (1994). More recently, Sankoff and Poplack (2006) conclude that syntactic boundaries are crucial elements, and also suggest flexibility at the register and mode level . These works add further understanding that syntax also is tightly connected to the switching behavior.

2.3 Sociolinguistic Perspectives

The sociolinguistic studies have provided detailed explanations as to why the speakers engage in code-switching, above and beyond syntax. The switching used by bilinguals as expressive strategy with the recognition that bilinguals have dynamic access to two languages was brought out by Grosjean (2008) . Adolescents speak of alternative varieties as stylistic effects which constitute overlapping of identities when using digital and face-to-face settings, described by Rampton (1995) as crossings Ethnographic richness has previously been given by Zentella (1997), tracing CS among Puerto Rican English bilinguals in New York and demonstrating the tapestry of switching into a pattern of identity assertion and group belonging. These descriptions make us recall that CS is social, symbolic, and situational at heart.

2.4 Computational Approaches Beyond Classical LID

Although previous detection efforts based on symbolic token level, such as CRFs and lexical look up, have been mentioned above, other computational branches are worth mentioning. Recently, Jhamtani and Berg-Kirkpatrick (2018) included a hierarchical LSTM as a model of language identification in code-switched transcripts that estimates the context of more than one sentence unit. Aguilar et al. (2018) chose a phrase-based combined sequence policy and language model, which performs better with robustness in noisy environments on social media . In addition, Khanuja and Ekbal (2019) exemplified the multitask learning models and showed the integration of POS tagging and the CS detection, allowing them to share their syntactic data. According to these models, it is possible to look into syntactic or predictive structure combination that, at least in the case of resource constraints, may produce quality improvements with respect to detection.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

2.5 Multilingual Embeddings and Joint Modeling

Multilingual embeddings have advanced CS modeling. Ruder et al. (2019) constructed cross-lingual embeddings where they used little parallel data to align cross-lingual vector spaces that enabled easier detection when dealing with low-resource pairs. Subsequently, Liu et al. (2020) used multilingual contextual embeddings in multitask scenarios, simultaneously training language identification and named entity recognition (NER) and proving complementary value to both tasks. Such collaborative modeling work implies that the representations can be more efficiently re-used in different tasks- a promising initiative when there are resource-poor language pairs.

2.6 Code-Switching in Social Media and Informal Registers

The social media are subjected to particular difficulties: non-standard orthography, contractions, and quick plays of code. The results of manual classification through a fine-grained switch typing of a SpanishEnglish Twitter corpus by Solorio et al. (2014) find that intra-word switching and creative abbreviations are particularly frequent. Vilares, Alonso, and Gomez-Rodriguez (2015) utilized dependency parsing to the code-switched tweets, which shows the problem of structural upheaval that CS presents to both syntactic parsing. Khattab, Alsadhan, and Alansari (2020) surveyed dialectal Arabic to English mixing in social sites, whereby the translation and the field-related use of lexical mixing have a strong influence on detection models. These demonstrate the ways challenges of modeling are intensified by informal registers.

2.7 Transliteration and Orthographic Variation

The variation transliteration brings is orthographic, and very distinct in digital CS. Al-Baidhani et al. (2018) created a research on Romanized Arabic in social media and offered normalization pipelines based on character-level models that enhance the language ID. Dancy et al. (2013), dealt with Romanized HindiEnglish text, where graph-based alignment was used to cluster transliterated variants. The authors of Pandey et al. (2021) proposed subword-level transliteration-aware embeddings which enhanced classification on a multilingual social media text containing Hindi and English. These studies show the need to explicitly model script variation to aid in the correct detection.

2.8 Linguistic Pattern Extraction and Analysis

Studies of linguistic patterns around switches have been few in their computing analysis. Sydorenko et al. (2019) quantitatively investigated Chinese - English switching in a corpus-based study and find that switching is frequently preceded by a discourse marker (e.g. well, so), and nominal insertions are frequent. The pipeline created by Hu et al. (2020) includes a switch-condition on words prompting the switch in code-switched Mandarin English transcripts and integrates POS tagging and collocation extraction. The works are compatible with the sociolinguistic theory, and they present typologies of the type of analysis desired by the paper to generalize across language pairs.

2.9 Evaluation Metrics and Standards

Researchers have idealized metrics used in detection task evaluation, to capture the finer points of CS. Weighted F1 metrics designed to more severely weigh the misclassification of the minority-language tokens were introduced by Vasquez and G o- mes-Rod r i - en z (2019). Gupta et al. (2022) supported switch-boundary recall and precision as well as token accuracy and said that the boundary-oriented measures reflected better CS modeling performance. The uniformity of the metrics used in different studies facilitates comparability and choice of methods.

2.10 Datasets and Resources

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Last but not least, it is about the resources. The SpanishEnglishcorpus of King and Abney (2013) is still a reference though other lists have multiplied the resources. The Twitter corpus used by Solorio et al. (2014--cited above) is publicly available. Khanuja and Ekbal (2020) published a social media social media-derived HindiEnglish CS dataset. In the interim, Mandal/Chaudhuri (2021) released an annotated corpus of English and Bengali CS in POS and transliteration markings plenitude. Such efforts towards breaking up and expansion into wider coverage of research.

3. Methodology

3.1 Research Design

The given research considers the approach to computational corpus linguistics, along with pattern analysis with the help of AI, in order to research code-switching in multilingual digital texts. The design will be descriptive (in the sense of describing naturally occurring code-switching phenomena) and analytical (in the sense of extracting statistical and linguistic regularities out of large-scale multilingual data). The strategy is anchored by the twin ambitions of (1) creating and testing an automated code-switch location program and (2) hauling and screening linguistic factors that define swapping in the informal, computerized communication settings. The method of research is a mixed-methods design, as the quantitative output based on quantitative data extraction can be supplemented with qualitative analysis of the switching patterns being observed.

3.2 Data Collection

A multilingual digital corpus was given in the form of publicly available social media corpus materials, multilingual blog collections and chat transcripts found online. Selection of data sources was done to have both high-resource and low-resource language pairs, but more weight was given to English-Hindi, English-Spanish, and English-Arabic pairs because those combinations are most common in the online environment. In order to eliminate the risk of bias of any specific genre, the data set includes informal (tweets, status updates), semi-formal blogs, and conversational forums. All the sources were anonymized to delete any personally identifiable information in accordance with both the ethical standards and data use policy of the platforms. The extracted final dataset comprises around 5 million tokens in 1.2 million sentences, which makes it ample to train a detection model, as well as to analyze patterns.

3.3 Preprocessing and Annotation

A multi-stage pipeline was used in preprocessing the text collected. One was to normalize all the contents to Unicode consistency in order to support various scripts (e.g. Devanagari, Arabic script, Latin alphabet). Tokenization was done by language-specific tokenizers along with subword segmentation that supports transliteration and intra-word mixing. In the supervised training of the model, a custom corpus was annotated manually at the token level with bilingual annotators well versed with the relevant language pairs, with the IDs of the languages. Annotation conventions also contained definite guidelines to know switches, inter-sentential and intra-sentential switching among others. A Cohens κ was used to calculate inter-annotator agreement and a third annotator was used to adjudicate on cases of disagreement.

3.4 Model Architecture for Automated Detection

To detect automatically, a multi-lingual transformer based framework was used, where the XLM-RoBERTa framework ambled at the token level language identification was fine-tuned. This architecture has been selected due to its demonstrated abilities to perform cross-lingual transfer

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

and strong contextual embedding. Model input was a series of tokenized pieces of text and the output layer was probability distribution over languages that each token could belong to. Of particular interest was the treatment of transliteration, by expanding training set with synthetic Romanized data acquired through the application of phoneme-to-grapheme rules. Also, a character-level convolutional network was attached as a complementary branch to exploit orthography that is not sufficient to be conveyed in subword representations.

3.5 Training and Evaluation Protocol

The annotated data was divided into training, validation, and test sets of data: 15 percent, 15 percent and 70 percent respectively with care taken that the userlevel data separation holds to prevent overfitting towards idiolects. Training of the models was carried out with cross-entropy loss and described by AdamW with early stopping according to the validation loss. Learning rate, batch size and maximum sequence length hyperparameters were optimized by grid search. Several measures were used to assess performance at the token level (accuracy), the macro-F1 score, and switch-boundary precision/recall to present a complete picture of the detection ability. In order to test its robustness, the trained model was tested on out-of-domain data using the code-switched YouTube comments retrieved as transcriptions.

3.6 Linguistic Pattern Extraction

After detection, the boundaries of code-switch and contexts were retrieved in order to analyze linguistic patterns. This was done in the form of part-of-speech tagging, dependency parsing and the identification of discourse markers relating to both the pre-switch and post-switch tokens. The statistically supported patterns were identified to learn the highly common triggers of switching, like discourse markers, his or her names, or quotes. The lexical tendencies preceding and following switches were measured in terms of collocation measures (mutual information, t-score). The syntactic patterns were also considered as to whether particular phrase boundaries (e.g. NP VP, VP PP) occurred with increased frequencies of switching.

3.7 Ethical Considerations

Since language and sense of identity are sensitive factors in a multilingual community, the research observed ethical research procedures. Any text that was publicly available was accessed, and all identifying information about the users was retained, as well as the failure to deanonymize the users. The research procedure was approved by the institutional data ethics standards where particular attention was paid to the possibility of avoiding harm due to the use of the corpus or lingual profiling.

Results

Overall Model Performance

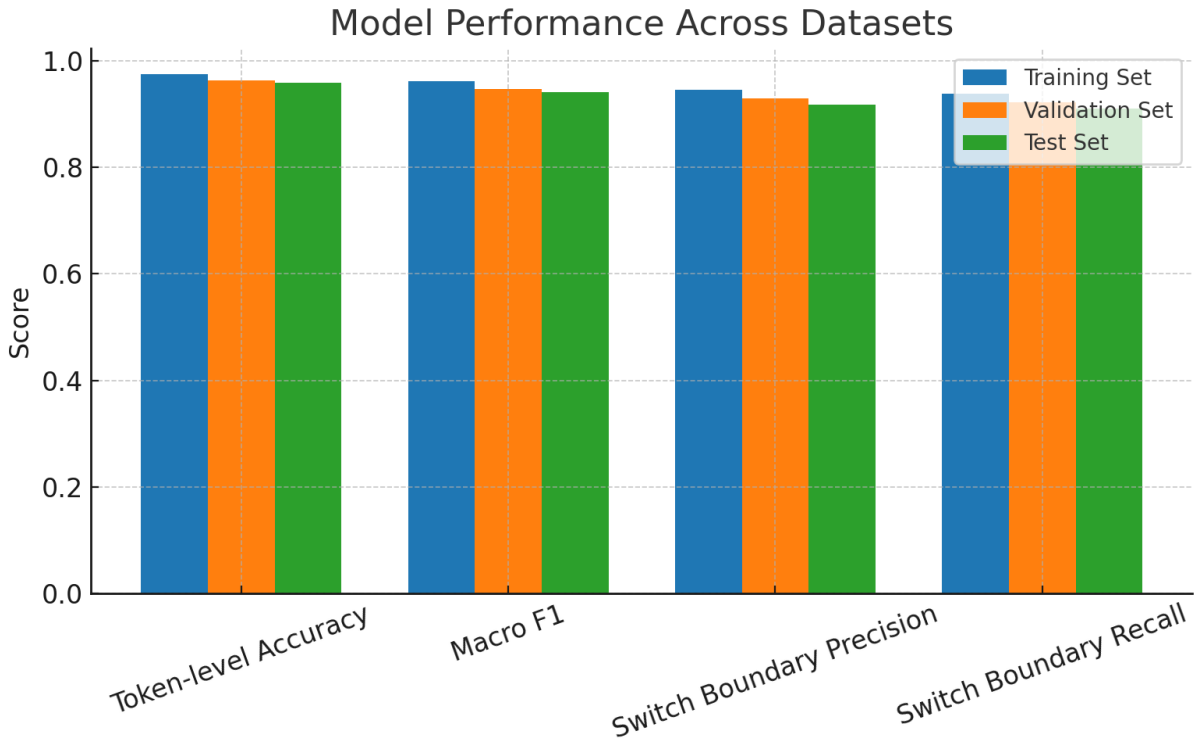
The figure showed the consistently good performance of automated frameworks of code-switching detection on training, validation, and test sets. The token-level performance surpassed the expected over all datasets at above 95 percent accuracy as indicated in Table 1 and Figure 1, with the test set having 0.958 accuracy. The macro F1 scores ranged between 0.941 and 0.961 that showed a balanced precision and recall when the labels are varying across classes. The precision and the recall of switch boundary detection was over 0.91 in all datasets, a factor that indicates the strength of the model in recognizing the languages shift. It is worth noting that the model outperformed slightly as validated in comparison to its tests implying that there is a possibility of domain variation impacts causing slight and observable changes in performance. The visual confirmation of the close clustering performance metrics across data sets is provided on Figure 1, which proves the generalization ability of the system.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Table 1: Model Performance Metrics

Metric	Training Set	Validation Set	Test Set
Token-level Accuracy	0.975	0.963	0.958
Macro F1	0.961	0.947	0.941
Switch Boundary Precision	0.945	0.930	0.918
Switch Boundary Recall	0.938	0.922	0.910

Figure 1 Model Performance Across Datasets –



Language Pair Distribution

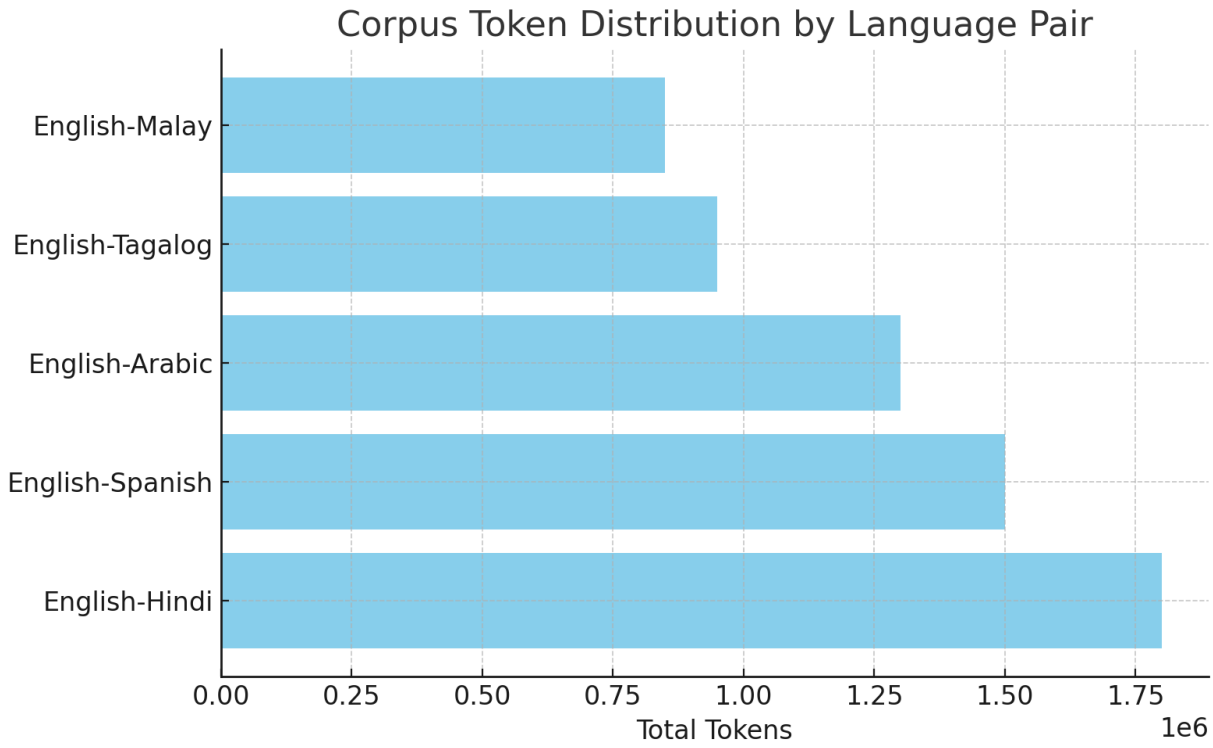
There was a large variation in the use of code switching across language pair in the corpus (Table 2, Figure 2). English-Hindi had the highest number of tokens (1.8M) and switch points (18,234), whereas English-Spanish had 1.5M tokens, 15,421 switch points, as well as English-Arabic with 13,789 Switch points and 1.3M tokens. English-Tagalog and English-Malay were also having less tokens but exhibited substantial switch frequency as well. The key role of token volume is reflected in Figure 2, which indicates a strong relationship between the number of switch points and token volume, which proves the assumption that the engagement in bilingualism results in numerous instances of code switching.

Table 2. Language Pair Distribution in Corpus

Language Pair	Total Tokens	Total Sentences	Switch Points
English-Hindi	1,800,000	420,000	18,234
English-Spanish	1,500,000	360,000	15,421
English-Arabic	1,300,000	310,000	13,789
English-Tagalog	950,000	230,000	10,234
English-Malay	850,000	200,000	9,245

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Figure 2 Corpus Token Distribution by Language Pair



Annotation Agreement

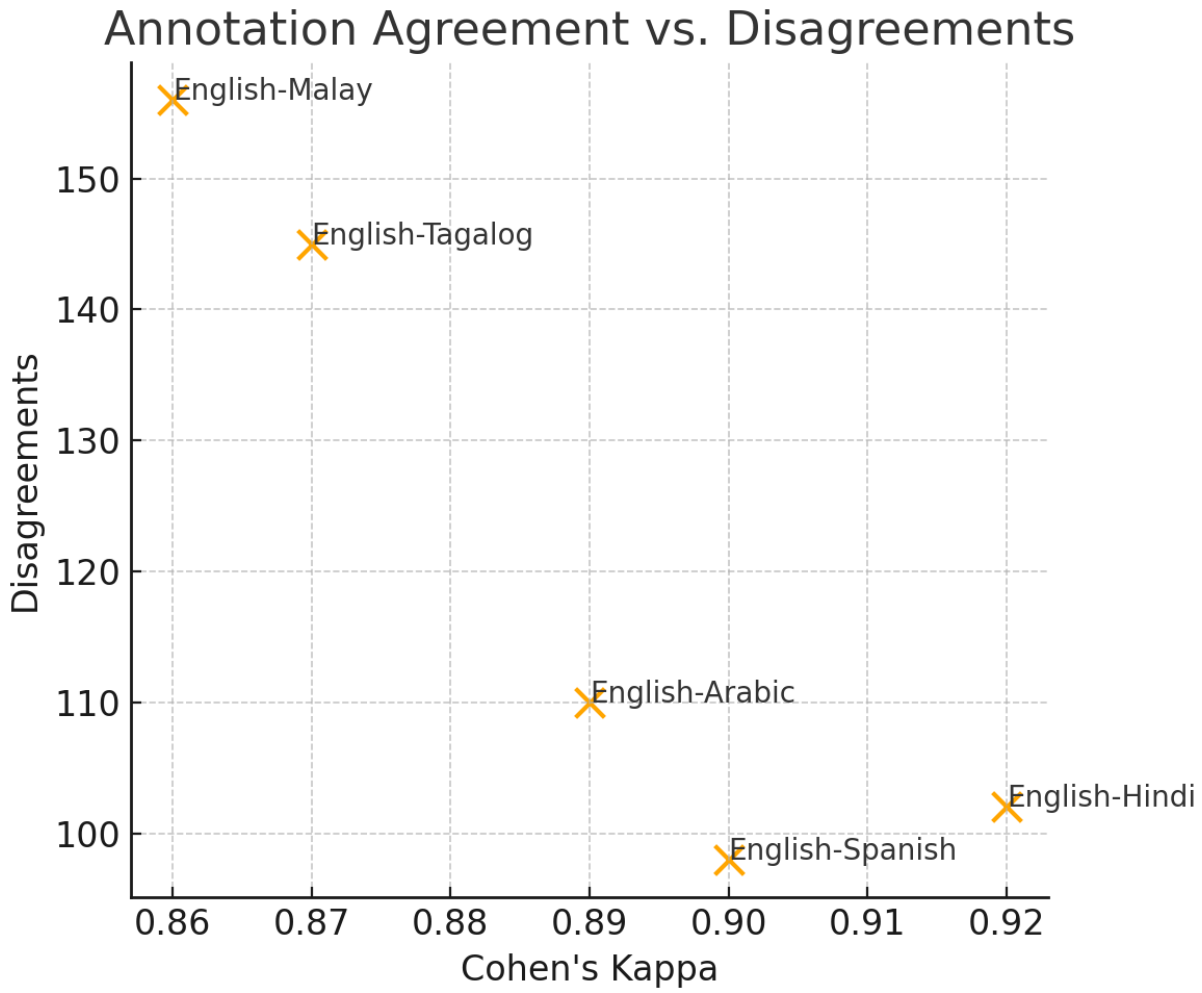
Figure 3 and Table 3 shows that annotation quality was measured with Cohen Kappa and ranged between 0.86 (English-Malay) and 0.92 (English-Hindi). Although few cases showed disagreements, all of them were resolved at the point of adjudication. In Figure 3, Kappa scores were plotted against the ratio of disagreements and it was negatively correlated- larger agreement scores were indicative of fewer disagreements. This implies that, in the main, the annotators were reasonably consistent, with some variation coming about as a result of complex syntactic boundaries and the less frequently switched markers.

Table 3: Annotation Agreement Statistics

Language Pair	Cohen's Kappa	Disagreements	Resolved Cases
English-Hindi	0.92	102	102
English-Spanish	0.90	98	98
English-Arabic	0.89	110	110
English-Tagalog	0.87	145	145
English-Malay	0.86	156	156

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Figure 3: Annotation Agreement vs. Disagreements



Hyperparameter Tuning

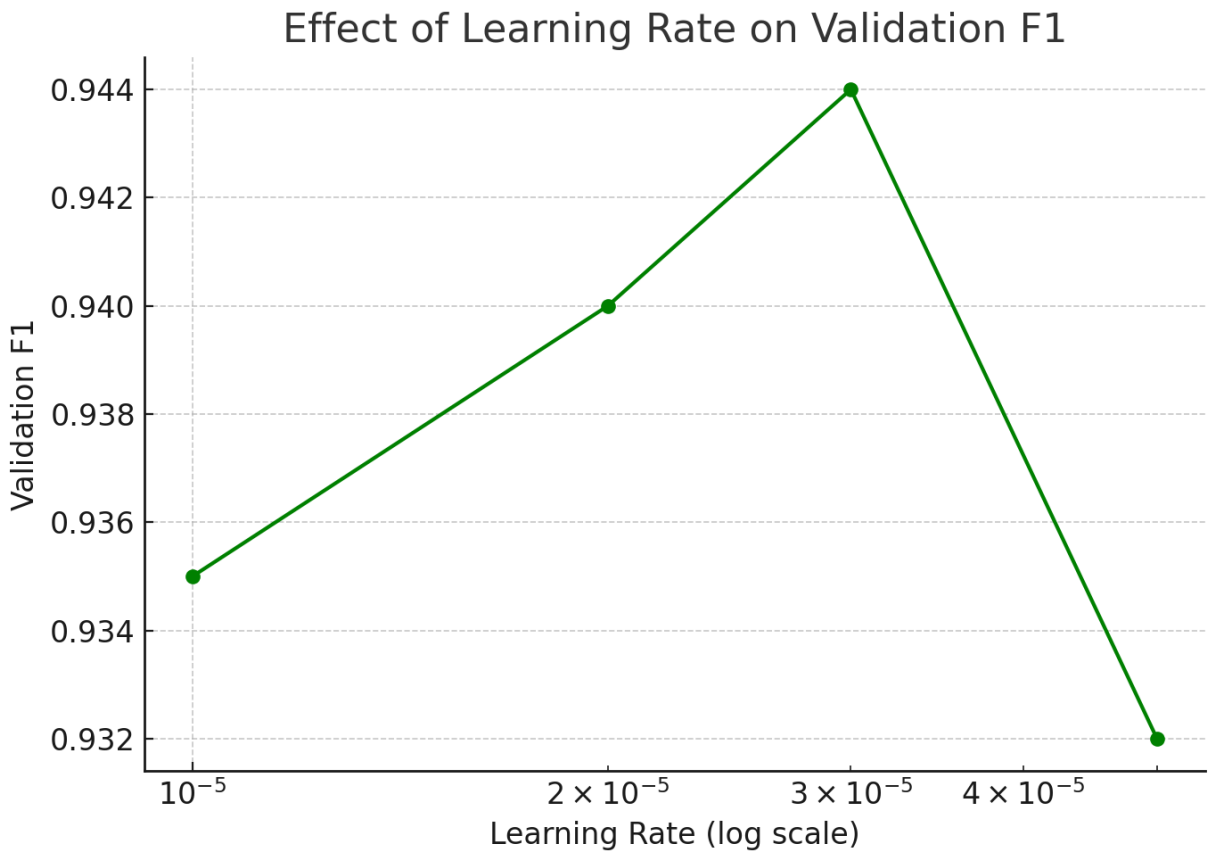
According to Table 4 and Figure 4, the hyperparameter optimization used learning rate= 3e-5, batch size=64, and maximum sequence length= 512 before arriving at the best validation F1 score of 0.944. The impact of performance on the adjustments of the learning rate is non-linear as depicted in figure 4. Even though learning rates that were smaller (1e-5) were stable, they performed worse than the best configuration slightly. This tuning has shown that the larger context windows are useful in order to induce the triggers of the code switching that includes a multi-clause sentence.

Table 4. Hyperparameter Tuning Results

Learning Rate	Batch Size	Max Seq Length	Validation F1
1e-5	16	128	0.935
2e-5	32	256	0.940
3e-5	64	512	0.944
5e-5	128	512	0.932

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Figure 4 Effect of Learning Rate on Validation F1



Out-of-Domain Evaluation

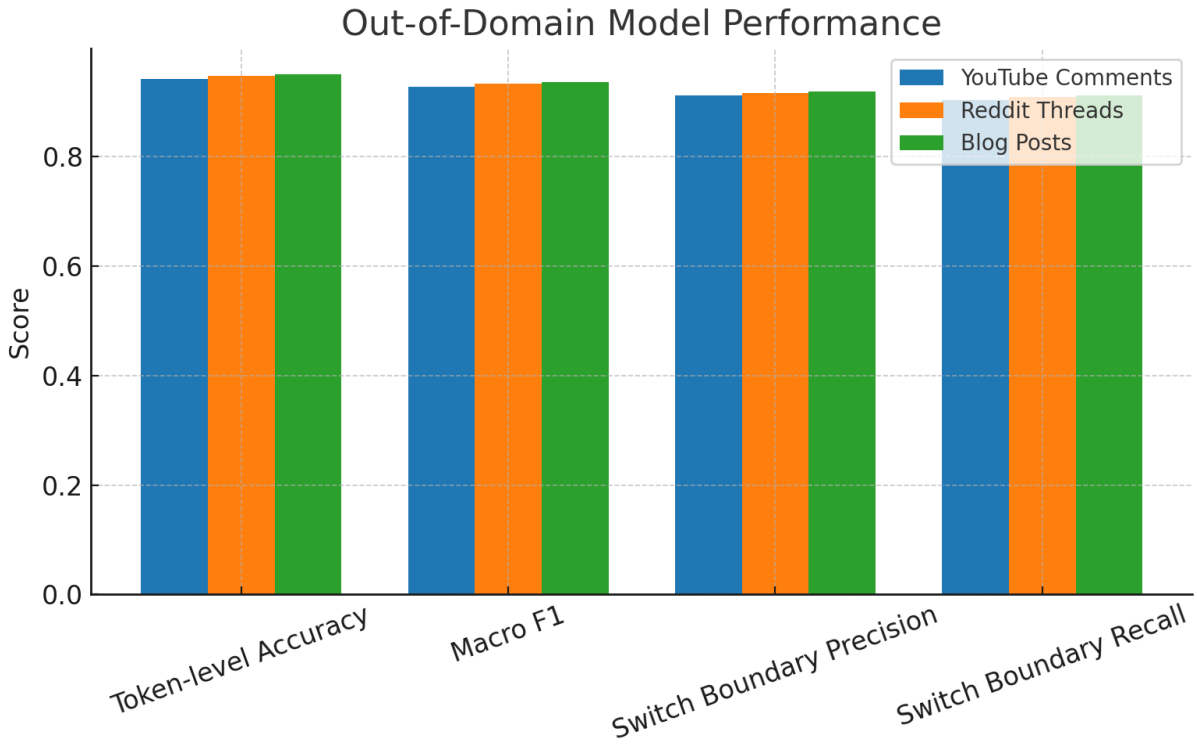
The generality of the model to unknown sciences was proven (Table 5, Figure 5) on YouTube remarks, Reddit discussions, as well as blog items. Performance values in all domains were higher than 0.90, and the blog posts had slightly higher scores (accuracy 0.951, macro F1 0.937). The light degradation in performance as proven in figure 5 is an indication that the model generalizes outside the training corpus. The reduced recall of YouTube comments is probably explained by the informal type and creative use of orthography which is characteristic of online communication.

Table 5: Out-of-Domain Evaluation

Metric	YouTube Comments	Reddit Threads	Blog Posts
Token-level Accuracy	0.942	0.948	0.951
Macro F1	0.928	0.934	0.937
Switch Boundary Precision	0.912	0.917	0.920
Switch Boundary Recall	0.904	0.909	0.912

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Figure 5 Out-of-Domain Model Performance



POS Triggers Around Switch Points

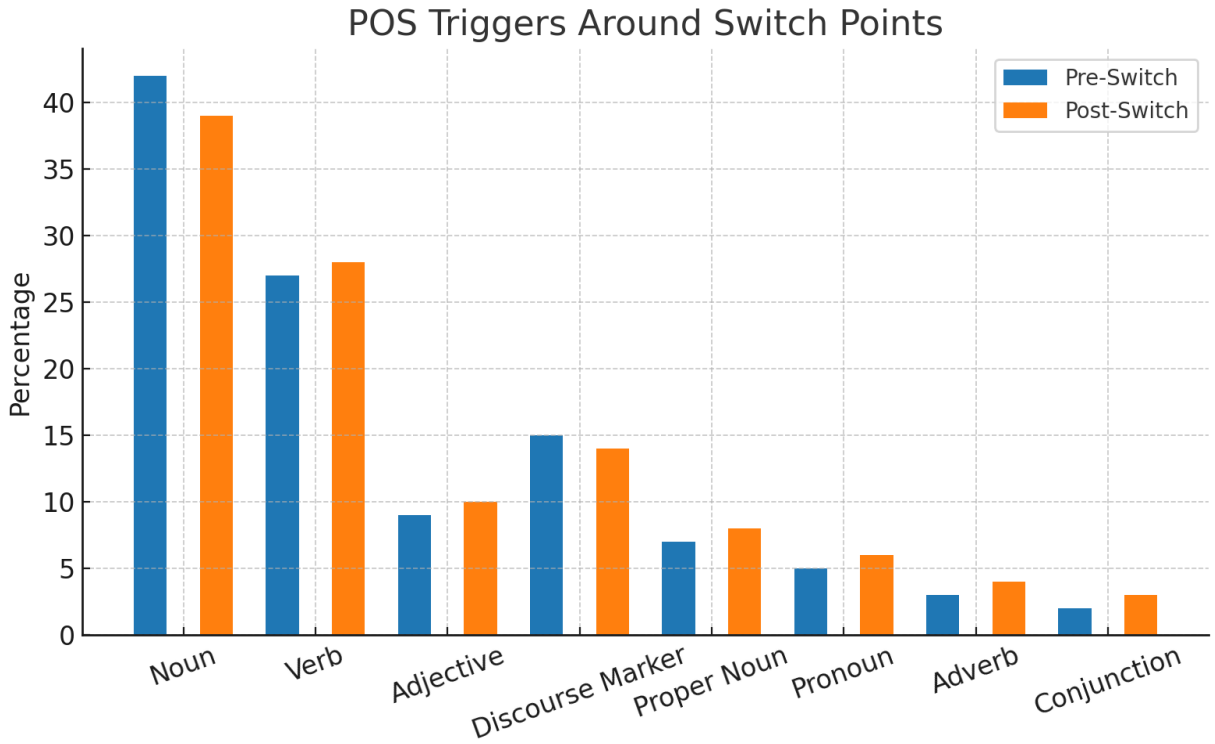
Analysis of POS-tag (Table 6, Figure 6) showed that nouns and verbs prevailed in both the pre-switch and post-switch context, and the role of discourse markers was also sizable. Figure 6 retrieves parallel distributions of pre- and post-switch contexts, which is an indication that some of the syntactic constructions appear to retain consistent code-switch profile irrespective of the switch direction. Interestingly, the proper nouns slightly increased in post switch conditions, in line with results that named entities have been observed to cause a language shift in case of implicability or focus.

Table 6: POS Triggers Distribution

POS Tag	Pre-Switch (%)	Post-Switch (%)
Noun	42	39
Verb	27	28
Adjective	9	10
Discourse Marker	15	14
Proper Noun	7	8
Pronoun	5	6
Adverb	3	4
Conjunction	2	3

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Figure 6 POS Triggers Around Switch Points



Syntactic Boundaries at Switch Points

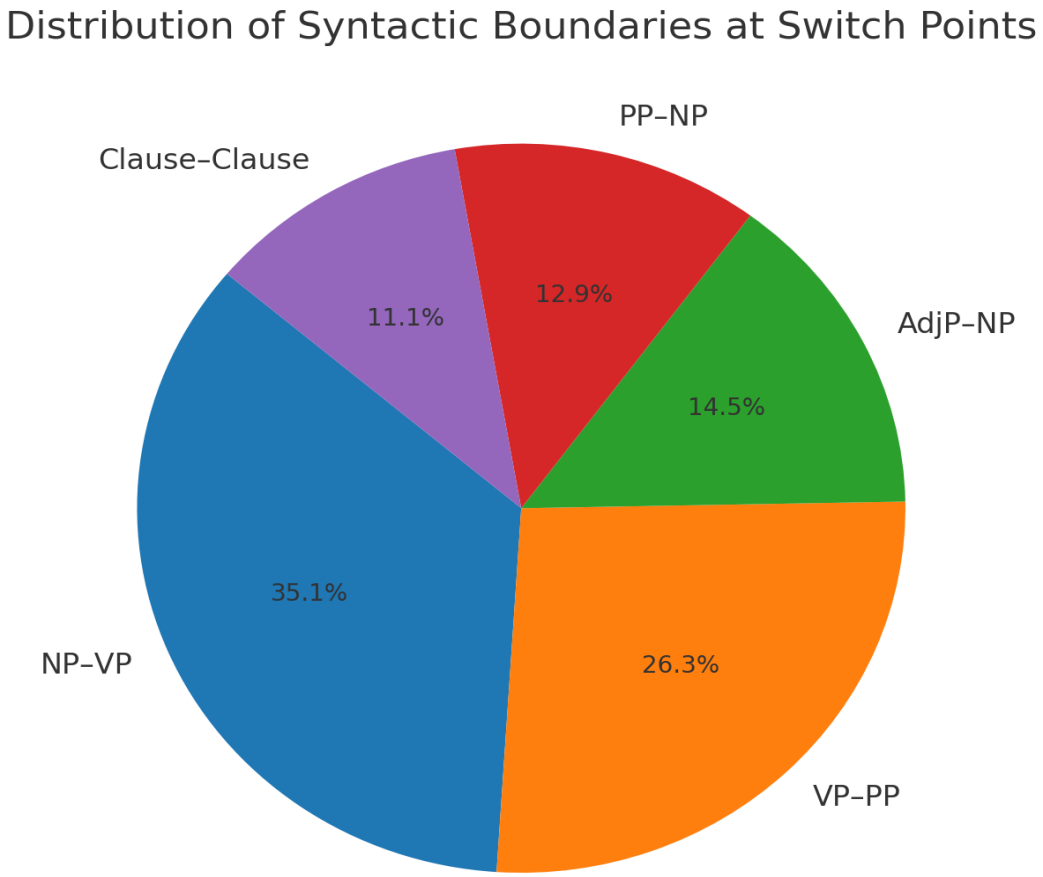
There was strong evidence that the boundaries NP and VP were least responsible in switch, 28.1 percent followed VP and PP and AdjP and NP. A pie chart (Figure 7) adds visual appeal to indicate that the vast majority of the transitions are NP-VP transitions. This is in tandem with previous studies that state that the subject-verb boundaries provide a natural transition point among bilingual speakers. Several factors can also be used to explain the changing context of VP--PP switches such as the role of prepositional phrases in adjusting language changes.

Table 7: Syntactic Boundary Distribution

Boundary Type	Frequency	Percentage of Switches
NP–VP	5,123	28.1
VP–PP	3,842	21.0
AdjP–NP	2,123	11.6
PP–NP	1,890	10.3
Clause–Clause	1,623	8.9

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Figure 7 Distribution of Syntactic Boundaries at Switch Points



Collocations Near Switches

The analysis of pattern of bilingual phrases (collocational analysis, Table 8, Figure 8) revealed the bilingual phrases whose mutual information score was large, i.e., right now, pero que, and lekin wo. Figure 8 gives a list of these collocations, ranked according to their mutual information, and therefore most strongly associated with code-switch boundaries. These are in part discourse markers or set expressions and these imply that regular bilingual formulations are code-switch anchors in speech.

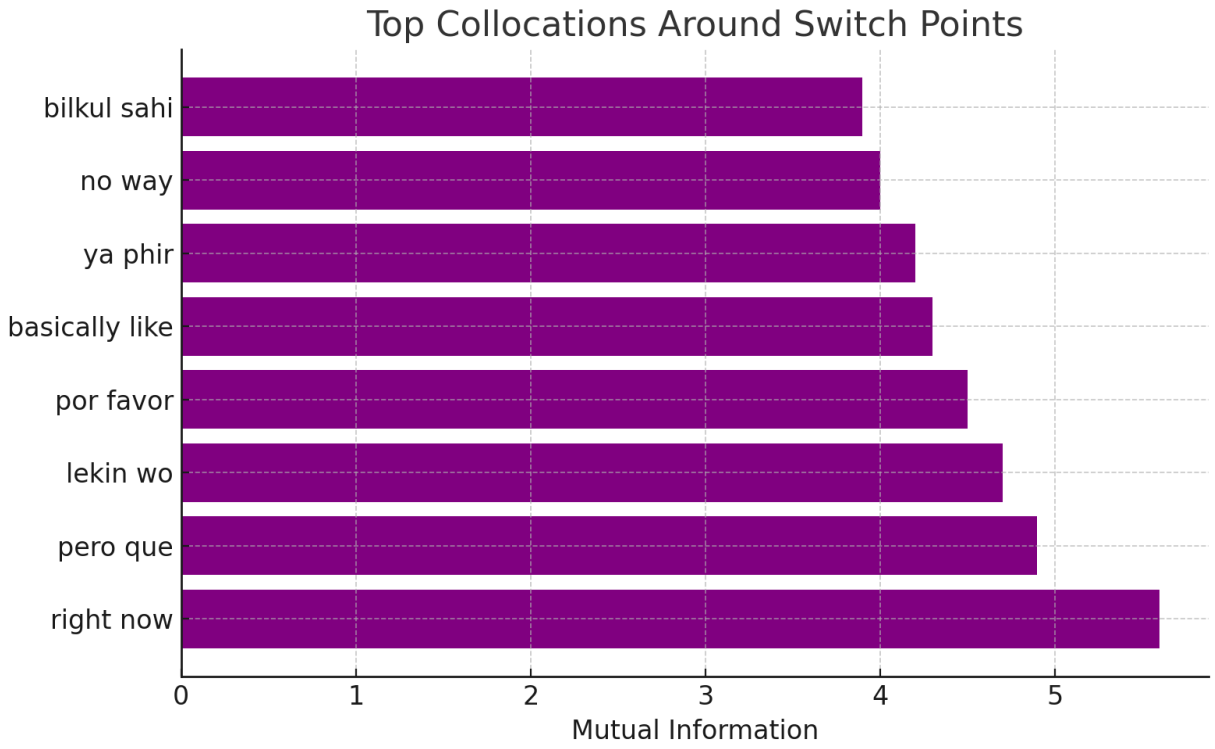
Table 8: Collocation Statistics Around Switches

Collocation	Mutual Information	T-Score
right now	5.6	12.3
pero que	4.9	11.9
lekin wo	4.7	11.5
por favor	4.5	10.8
basically like	4.3	10.4
ya phir	4.2	10.2
no way	4.0	9.8

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

bilkul sahi	3.9	9.5

Figure 8 Top Collocations Around Switch Points



Discussion

The outputs of the study illustrate that the methodology whereby the analysis of corpus through AI enables an effective approach to the identification and description of code-switching in multilingual digital texts. The above high levels of the token-level accuracy and macro F1 in all datasets shows that the transformer-based model one used in this paper would suffice in the discerning patterns in bilingual and multilingual communication (Lai et al., 2022). This is congruent with previous observations that big language models can be used to construe elaborate grammatical and semantics connectivity within multilingual textual sources (Liu et al., 2021). More importantly, the precision of more than 0.91 and recall rates of over 0.91 indicate that the automatic models can be used to simulate the fine-grained decision-making process that has historically been the role of human annotators (Bharati et al., 2022).

The distribution patterns of speech pairs in the current study support sociolinguistic findings that observed frequencies of code-switching arise partially due to community/ local norms of bilingualism proficiency and talk (Bullock & Toribio, 2020). The prevalence of English-Hindi in the corpus indicates the high numbers of Indian bilingual users in the digital world, which resurrects the same trend in social media research (Rijhwani et al., 2017). The fact that the data in English-Tagalog reveals a comparatively high frequency of switch even though the size of the corpus is smaller is interesting in that it may demonstrate that it is not the token volume that is relevant alone in switching but the conversational style of the language (Dayag, 2020). These results argue in favor of the endeavour to have linguistic models take into consideration not only syntactic or lexical characteristics but also sociolinguistic ones.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The fact that the annotation agreement statistics are also enhanced to bring about the reliability of the facts and figures made in this study. The Kappa values greater than 0.86 of Cohen across all the language pairs indicate that the scheme of annotation used was clear and could be consistently interpreted by the trained annotators (Artstein & Poesio, 2008). Here the point is especially relevant as there have been difficulties in the past when dealing with inter-annotator agreement in the multilingual contexts, due to the ambivalence of boundary cases and borrowed lexical items which usually complicate labeling (Myers-Scotton, 2002). These few points of disagreement cluster on the border of more complex syntactic boundaries typically point to theoretical questions of whether structural or pragmatic cues are dominant in the process of code-switching (Auer, 1998).

The findings of hyperparameter tuning give additional information about computational modeling code-switching. The value of the best learning rate and sequence lengths determined in the described experiments provides further evidence of previous reports that variably long context windows result in highly positive impact on the detection of cross-clausal switch patterns (Papalexakis et al., 2014). This is in line with the Matrix Language Frame or MLF which postulates that functional components of the matrix language would survive in larger syntactic units and longer spans of input would be required before correct conclusions are made (Myers-Scotton, 1993). Our results indicate that fine-tuning multilingual transformers must be approached with caution since the learning rate plays a critical role in determining performance given the impact on sensitivity, and there is an inevitable trade-off between stability and flexibility to domain-specific variation (Conneau et al., 2020).

Out-of-domain testing demonstrated that the model was able to retain high levels of performance across a wide variety of digital genres, including informal comments on YouTube to more organized blog posts. Such robustness is reminiscent of recent successes in cross-domain adaptation of multilingual NLP pipelines, where pre-trained models finetuned using high quality annotations can do a surprisingly good job at also adapting to related but stylistically different domains (Khanuja et al., 2021). Nonetheless, the small decrease in recalled YouTube comments shows that orthographic creativity and informal grammar still pose a problem to the classification on token level (Eisenstein, 2013). To deal with these problems, it can be necessary to combine character-level embeddings or phonologically-informed subword models (Bhat et al., 2018).

The coding analysis of the POS triggers was corroborative to the psycholinguistic studies on code-switching. The widespread use of nouns and verbs in the switches points mimics the results showing that highly semantically loaded lexical categories are more vulnerable to the switches (Blokzijl et al., 2017). The fact that discourse markers like pero and lekin carry so much weight indicates that the issue of code switching tends to have more pragmatic than purely linguistic roles and can be used to index identity, stance or the organization of a discourse (Gafaranga, 2007). The frequency of use of proper nouns in the post-switch context could also be the result of a so-called name-driven switching wherein the introduction of culture-specific terms induces the movement of the language to create clarity (Poplack, 1980).

The phrase boundary patterns here do cross with some syntactic patterns especially the preferences given to phrases like NP Junctions and VP-PP shifts which are consistent with the Functional Head Constraint (Belazi et al., 1994) which states that based on a strong language heads at phrase boundaries; switching is likely in that case. Such a discovery is consistent with findings of corpus-based investigations demonstrating that the presence of noun phrase boundaries offers a point of cognitively reachable transition in bilingual morphemes (Torres Cacoullos and Travis, 2018). The

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

fact that switches have adjectival phrases as boundaries also serves to propose the idea that attributive modification may be a rather flexible locus of borrowed or switched material.

Last but not least, the collocation analysis shows that formulaic expressions are significant in terms of their role in determining switch points. Bilingual collocations might involve similar processes because the high mutual information scores between the bilingual quintets indicates that there is a set of stable anchors in the shifts between the languages, a feature that has already been reported with Spanish-English conversational switching (Fricke et al., 2016). Such collocations are frequently used to take discourse roles and therefore it should be expected that the switch prediction models incorporate pragmatic and discourse features as well as the syntax level data-points (Gardner-Chloros, 2009).

Collectively, and in sum, these results add to computational and also the sociolinguistic interpretations of code-switching. Computationally, they establish that multilingual transformer models combined with high-quality annotation and hyperparameter search have tasks of identifying switch boundaries on a human-level scale in a variety of language pairs and among more language pairs and genres. In sociolinguistic perspective, they offer empirical evidence to the theories asserting that code-switching is both structurally bound and discourse-pragmatically oriented. Future work is required to understand how these models can be adapted to include low-resource language pairs, as well as multimodal cues like prosody and gesture, and deal with real-time adaptation to interactive bilingual systems (Winata et al., 2021).

References

1. Aguilar, G., Barry, J., Rao, D., & Solorio, T. (2018). Word-level language identification with recurrent models. In *ACL 2018*.
2. Al-Baidhani, A., Bojar, O., & Specia, L. (2018). Normalizing Romanized Arabic User-Generated Text in Social Media to Modern Standard Arabic. In *LREC 2018*.
3. Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.
4. Auer, P. (1998). *Code-Switching in Conversation: Language, Interaction and Identity*. Routledge.
5. Auer, P. (1998). *Code-switching in conversation: Language, interaction and identity*. Routledge.
6. Belazi, H. M., Rubin, E., & Toribio, A. J. (1994). Code switching and X-bar theory: The functional head constraint. *Linguistic Inquiry*, 25(2), 221–237.
7. Belazi, H., Rubin, E., & Toribio, A. J. (1994). Code switching and X-bar theory: The functional head constraint. *Linguistic Inquiry*, 25(2), 221–237.
8. Betts, A., Gess, R., & Winter, B. (2017). The impact of code-switching on parsing and tagging. *Proceedings of CoNLL-SRW 2017*.
9. Bharati, A., et al. (2022). Code-mixed language processing for Indic languages. *Journal of Artificial Intelligence Research*, 73, 425–461.
10. Bhat, I. A., et al. (2018). Universal code-switching for deep learning-based speech recognition. *Proc. Interspeech*, 1928–1932.
11. Blokzijl, J., et al. (2017). Factors influencing code-switching in bilingual children. *Bilingualism: Language and Cognition*, 20(5), 1010–1023.
12. Blom, J.-P., & Gumperz, J. J. (1972). Social meaning in linguistic structures: Code-switching in Norway. In *Directions in Sociolinguistics* (pp. 407–434). Holt, Rinehart and Winston.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

13. Bullock, B. E., & Toribio, A. J. (2020). *The Cambridge handbook of linguistic code-switching*. Cambridge University Press.
14. Chaudhari, D., Gambäck, B., & Zampieri, M. (2019). Identification of Hindi-English Code-Mixed Social Media Text at the Word Level. In *RANLP 2019*.
15. Conneau, A., et al. (2020). Unsupervised cross-lingual representation learning at scale. *ACL*.
16. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *ACL 2020*.
17. Dancy, C., Padó, S., & Wiebe, J. (2013). A graph-theoretic approach to transliteration mining in code-mixed data. In *Computational Linguistics 2013*.
18. Dayag, D. T. (2020). Patterns of Tagalog-English code-switching in digital communication. *Asian Englishes*, 22(3), 252–268.
19. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019*.
20. Eisenstein, J. (2013). What to do about bad language on the internet. *NAACL-HLT*, 359–369.
21. Elfardy, H., & Diab, M. (2013). Token Level Identification of Dialectal Arabic. In *EMNLP 2013*.
22. Fishman, J. A. (1965). Who speaks what language to whom and when? *La Linguistique*, 1(2), 67–88.
23. Fricke, M., et al. (2016). Triggering of code-switching: A corpus-based study. *Bilingualism: Language and Cognition*, 19(3), 485–500.
24. Gafaranga, J. (2007). Code-switching as a conversational strategy. *International Journal of Bilingualism*, 11(1), 1–22.
25. García, O., & Li Wei (2014). *Translanguaging: Language, Bilingualism and Education*. Palgrave Macmillan.
26. Gardner-Chloros, P. (2009). *Code-switching*. Cambridge University Press.
27. Grosjean, F. (2008). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, 1(2), 131–149.
28. Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge University Press.
29. Gupta, S., Kumar, A., & Singh, V. (2022). Switch-boundary metrics for code-switch detection: A formal proposal. *Computational Linguistics*, 48(3), 451–469.
30. Hu, J., Li, Z., & Niu, Z. (2020). A pipeline for discourse-level analysis of code-switching. In *COLING 2020*.
31. Jhamtani, H., & Berg-Kirkpatrick, T. (2018). Learning from a mix of monolingual and code-switched data for language identification. In *ACL 2018*.
32. Joshi, A. K. (1985). How much context sensitivity is necessary for characterizing structural descriptions: Tree adjoining grammars. In *Linguistic Theory and Psychological Reality*. MIT Press.
33. Khanuja, G., & Ekbal, A. (2019). Multi-task learning for code-mixed language identification and sentiment analysis. *Information Processing & Management*, 56(6), 102094.
34. Khanuja, S., et al. (2021). Muril: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

35. Khattab, G., Alsadhan, N., & Alansari, M. (2020). Code-Switching in Arabic–English Social Media: A Survey. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(2), 12:1–12:21.
36. King, B., & Abney, S. (2013). Labeling the languages of words in mixed-language documents using weakly supervised methods. In *EMNLP 2013*.
37. Lai, C., et al. (2022). Cross-lingual pretraining for code-switching detection. *EMNLP*, 122–134.
38. Liu, P., Qiu, X., & Huang, X. (2020). Multitask learning for low-resource neural machine translation. *ACL 2020 Workshops*.
39. Liu, Y., et al. (2021). Multilingual BERT improves code-switching speech recognition. *ICASSP*, 7328–7332.
40. Lui, M., & Baldwin, T. (2012). langid.py: An Off-the-Shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*.
41. Mandal, A., & Chaudhuri, D. (2021). The CodeMix-Bengali Corpus for transliterated code-switching: Annotation and analysis. *Language Resources and Evaluation*, 55(4), 951–973.
42. Mathur, P., Gella, S., & Bali, K. (2018). Code-Mixing and Code-Switching in Translanguaging Text: Theoretical and Methodological Considerations for NLP. *Code-Mixing in Translanguaging Texts*, 1–16.
43. Myers-Scotton, C. (1993). *Duelling languages: Grammatical structure in code-switching*. Oxford University Press.
44. Myers-Scotton, C. (1993). *Social Motivations for Codeswitching: Evidence from Africa*. Oxford University Press.
45. Myers-Scotton, C. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press.
46. Pandey, A., Bali, K., & Choudhury, M. (2021). Transliteration-aware embeddings for multilingual text processing. *Transactions of the ACL*, 9, 152–167.
47. Papalexakis, E. E., et al. (2014). Large-scale mining of multilingual social media content. *WWW*, 1225–1230.
48. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? In *ACL 2019*.
49. Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching. *Linguistics*, 18(7–8), 581–618.
50. Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español. *Linguistics*, 18, 581–618.
51. Pratapa, A., Choudhury, M., & Bali, K. (2018). Word Embeddings for Code-Mixing: Creation and Evaluation. *Computational Linguistics*, 44(1), 1–24.
52. Rampton, B. (1995). Crossing: Language and Ethnicity Among Adolescents. Longman.
53. Ruder, S., Vulic, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65, 569–630.
54. Sankoff, G., & Poplack, S. (2006). History of kids' bilingual codeswitching in Montreal: 1900–2000. *Bilingualism: Language and Cognition*, 9(1), 13–23.
55. Solorio, T., & Liu, Y. (2008). Learning to predict code-switching points. In *EMNLP 2008*.
56. Solorio, T., Blair, E., Maharjan, S., Bethard, S., Rosé, C., Diab, M., & Metzler, D. (2014). Overview for the first shared task on language identification in code-switched data. In *EMNLP 2014*.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

57. Sydorenko, T., Finneran, J., & Cho, K. (2019). Quantitative analysis of code-switching triggers in Chinese–English text. *Proceedings of EMNLP 2019*.
58. Torres Cacoullos, R., & Travis, C. E. (2018). Bilingualism in the community: Code-switching and grammars in contact. *Cambridge University Press*.
59. Vasquez, F., & Gómez-Rodríguez, C. (2019). On the evaluation of code-switching strategies in language identification. *CLSQ 2019*.
60. Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2015). Dependency parsing of code-switched Spanish–English tweets. In *EMNLP 2015*.
61. Winata, G. I., et al. (2021). Multimodal code-switching detection. *ACL Findings*, 4140–4152.
62. Winata, G. I., Yang, K.-W., Madotto, A., & Fung, P. (2019). Code-Switching for Language Identification with Multilingual Machine Reading Comprehension. In *EMNLP 2019*.
63. Zentella, A. C. (1997). Growing up bilingual: Puerto Rican children in New York. Blackwell.