Discovering System Morphemes in Conversational and Literary Code-Switched Text

Anonymous ACL submission

Abstract

Code-switching (CS) is the process of speakers interchanging between two or more languages. To describe CS, the Matrix Language Frame (MLF) theory defines system morphemes, which contribute to a sentence's grammatical structure. The discovery of the system morphemes can improve automatic CS text generation and analysis. This work introduces a novel approach for discovering system morphemes from analysing frequencies of Parts of Speech (POS) in monolingual data based on the MLF theory. The method is implemented and evaluated on the Matrix Language (ML) determination task for SEAME and Miami datasets. Applying the System Morpheme Principle approach with the extracted system morpheme set revealed that the conventional system words (pronouns, conjunctions, determiners, auxiliaries) fall into the top 50% most frequent word frequencies when averaged over POS. Moreover, the novel approach for system morpheme extraction has also revealed the ranking of the POS with respect to the ML determination task, showing the importance of particles and adpositions. The performance of ML determination is tested on literary data, and the proposed method leads to an absolute 1.22% F1-score increase. Finally, correlation analysis is performed between the CS fragments and stylistic function, which revealed a correlation between monolinguality and thematic stylistic function, as well as between highlights in poetic texts and adherence to MLF.

1 Introduction

Code-switching (CS) is the process of speakers switching between several languages in spoken or written language. CS data is typically scarce, therefore, models for CS analysis and recognition often yield poor performance in comparison to monolingual models. Despite being poorly documented, the use of CS is widespread (e.g India, South Africa, Nigeria) (Diwan et al., 2021; Ncoko et al., 2000; Rufai Omar, 1983), and therefore it is essential to develop Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) technologies for processing both CS speech and text.

To describe code-switching the Matrix Language Frame (MLF) theory was introduced (Myers-Scotton, 1993). It introduces the concept of a main, i.e. dominant language and a secondary, inserted language to describe CS sentences. These languages are called Matrix Language (ML) and Embedded Language (EL), respectively. The MLF theory proposes two rule-based methods for ML determination. *The Morpheme Order Principle* (MOP) aims to identify the surface morpheme order for a CS sentence if it consists of singly occurring EL lexemes and any number of ML morphemes. For example, in the sentence "哦你 post 在你的那个 blog" the English words "post" and "blog" are singly occurring and the surface morpheme order is that of Mandarin, therefore the ML is Mandarin. *The System Morpheme Principle* (SMP) states that system morphemes which have grammatical relations external to their head constituent will come from ML. System morphemes are a type of morpheme that primarily serve a grammatical function rather than carrying lexical meaning. For example, in the sentence "im okay with **the** 蛋黄" the English determiner "the" is present, therefore the ML in this sentence is English. There are no known methods for automatic detection or determination of system morphemes. In the linguistic literature whole single-morpheme words belonging to such Part of Speech (POS) as coordinating and subordinating conjunctions, auxiliaries, determiners and pronouns are commonly cited as system morphemes, yet the researchers do not provide a definitive or comprehensive closed set. MLF theory provides the framework for identifying the "main" or "dominant" language in a CS sentence and may bring valuable insights for CS data, such as language or morpheme distributions, but it has been

rarely used for automatic analysis of conversational and literary texts.

In this paper, a novel approach for discovering system morphemes is introduced. An approach to system morpheme extraction, which computes average POS frequencies from monolingual data, is developed to improve the performance of SMP over the conventional system morpheme set. MOP is used to assess the ML determination performance of SMP with the extracted system morpheme sets. POS rankings of contribution to grammar are extracted based on the novel approach. The extracted frequency-based POS rankings are compared to the conventional system morpheme set. As a result, when using SMP along with the novel approach for system morpheme discovery, it has performed significantly better in the textual Matrix Language Identification (MLID) task than the conventional system morpheme set. Furthermore, the performance of ML determination is evaluated on literary data, where the proposed method yields an improvement in the macro-averaged F1 score. Finally, a correlation analysis between CS fragments and stylistic functions reveals a relationship between monolinguality and the thematic stylistic function, as well as between highlights in poetic texts and adherence to the MLF model.

The remainder of the paper is organised as follows. Section 2 introduces relevant literature for the applications of MLF. Section 3 provides a detailed description of the methods used. This is followed by Section 4, which provides information on datasets, detailed implementation, experiment descriptions, as well as a discussion of results. Section 5 summarises and completes the paper.

2 Related Work

Following the MLF, one can state that CS is a **composition of two languages** and not a manifestation of a new language. This way, multilingual data should be sufficient for building an ASR model that can recognise CS. In reality, multilingual models cannot surpass the quality of systems trained on CS data (Li et al., 2019; Shan et al., 2019). This is due to the systems not being introduced to cross-lingual information, which helps with CS processing. This additional information could be derived from applying the MLF theory to CS data.

Although the MLF model has not been used explicitly in NLP, the ideas of the dominant ML have been used for CS data simulation. MLF model has been used for simulating CS sentences from monolingual texts for them to be further used in language modelling, as in Hu et al. 2019 and Lee et al. 2019, or constructing a self-supervised training procedure for a machine learning algorithm in such a way that it encourages the generation of utterances with CS (Chang et al., 2019). There have been attempts to determine the influence of different attributes of CS speech and text, which are in line with the linguistic theories: Brown word vectors (Adel et al., 2015), POS and cognate word pairs (Soto and Hirschberg, 2019). The latest multilingual systems for CS achieve the best results when Byte Pairwise Encoding (BPE) is used for tokenisation, in contrast to grapheme (character) (Hou et al., 2020; Chowdhury et al., 2021) or word tokenisation, which also supports the MLF theory since BPE tokens frequently coincide with morphemes.

While distinguishing between content/system morphemes is important for the ML determination task, it can serve a different purpose in NLP. According to the Morpheme Sorting Principle of the MLF theory, the four morpheme categories of the 4-M model have different probabilities of being in the EL. The 4-M model is a morpheme classification framework introduced in Myers-Scotton 2002 that describes the order in which a bilingual accesses their mental lexicon, implying that some morphemes and words are formed earlier during language production. This introduces advantages to text and speech generation and processing systems, guiding the generation processes to be more natural and the analysis more precise. However, the distinction between content and system morphemes in these studies was often based on author-defined heuristics with limited theoretical justification. Iakovenko and Hain 2024 introduces precise implementations of the ML determination principles from the MLF theory, but a fixed system morpheme set is used for the SMP implementation. Similarly, Bullock et al. 2018 presents rule-based systems for textual ML determination in CS text based on the token and system POS majorities, but the choice of the system morphemes is not thoroughly justified. Also, in Du et al. 2021, words that belong to the noun and/or verb POS were regarded as content morphemes. However, none of the works specify why these POS were chosen for translation into the EL or for staying in the ML. POS distributions were explored in Hamed et al. 2018 for Arabic/English CS, but the differences in

EL POS distributions between the two languages acting as ML were not explored. Finally, the automated principles from MLF theory have never been, to our knowledge, applied to literary texts.

It must be highlighted that the MLF theory does not perfectly model CS; for example, there have been observations that the MLF theory does not describe the CS of African languages (Auer and Muhamedova, 2005; Nguyen, 2025). An alternative linguistic theory that addresses this issue is the Equivalence Constraint (EC) theory (Sankoff and Poplack, 1981), which relies on the principle of grammatical congruence for modelling and analysing code-switching. Even so, the authors of the present study believe that it is essential to develop MLF-based methods for better understanding and modelling of CS. While pure MLF implementations may not describe all CS production, we believe that their relaxed modifications may be used to fit a variety of CS data.

The above indicates the importance of identifying system morphemes, but it has not been carried out before in the context of ML determination and other downstream tasks. Also, the computational methods from the MLF theory have not been used before for literary text correlation analysis. Therefore, the objective of this study is to advance technologies for conversational and literary CS text data by discovering the system morphemes based on the ML determination principles from the MLF theory.

3 Methods for ML Determination and System Morpheme Set Discovery

Being called "principles for ML determination" (Section 1), MOP and the SMP present three of the features of CS Complementiser Projections (CP) which cannot be used to determine the ML directly. This means that the linguistic principles for ML determination do not introduce a concrete definition of what ML is in a CS CP. Therefore, the principles need to be reformulated to perform only ML prediction based on a set of conditions. Let $\mathbf{x} = (x_1, .., x_n)$ be a CS CP as a sequence of morphemes, $\mathbf{l} = (l_1, .., l_n), l_i \in L_1 \cup L_2$ - a sequence of corresponding morpheme Language Identity (LID) tags, then: a) *The Morpheme Order Principle*: if singly occurring $\mathbf{x}_{i:j}$ lexemes (sequence of morpheme constituents in a lexeme where i is the index of the prefix and j is the index of the suffix of the word) come from the same language L_2 within a context of morphemes from L_1, then L_1 is the ML and L_2 is the EL (a detailed description of the method can be found in Iakovenko and Hain 2024 under the name of P1.1); b) *The System Morpheme Principle*: if $\forall \hat{t} \in \hat{\mathbf{t}}_{L_2},\ \hat{t} \neq T_{sys}$, but $\exists \hat{t} \in \hat{\mathbf{t}}_{L_1}$ such that $\hat{t} = T_{sys}$, then L_1 is designated as the ML and L_2 as the EL, where $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_n)$ is the sequence of morpheme types associated with the system/content morpheme classification, $\mathbf{x} = (x_1, \ldots, x_n)$ and $\hat{\mathbf{t}}_{L_1}, \hat{\mathbf{t}}_{L_2} \subseteq \hat{\mathbf{t}}$ denote the subsets of morpheme types belonging to languages L_1 and L_2, respectively. Below are detailed descriptions of the SMP method variations.

3.1 Implementing the System Morpheme Principle

Compared to MOP, there are fewer issues in adapting SMP the principle for ML determination. However, as highlighted in the earlier section, there are no computational methods for identifying system morphemes. Despite lacking the complete system morpheme set, one can determine system morphemes from a composition of context-free probabilities of morphemes if an ML identity is known for a CP.

It has been previously shown that a correlation was found between the frequency of occurrence and system/content categorisation (Marino et al., 2020). Therefore, let us first assume that system morphemes X_{sys} - the morphemes that contribute to the grammatical structure of the CS CP - are the morphemes that are frequent in the data. Let us further assume that for all $x \in X_{sys}$ the corresponding morpheme type is equal to T_{sys} (system morpheme type), and similarly for all $x \in X_{cont}$ morpheme type is T_{cont} (content morpheme type). Then x being a system morpheme may be approximated by the morpheme frequencies $P(x)$:

$$\hat{t} = \begin{cases} T_{sys}, P(x) > \beta \\ T_{cont}, \text{otherwise} \end{cases} \tag{1}$$

where β is threshold for discriminating system morpheme set X_{sys} and content morpheme set X_{cont}. This approach may also be used to derive system morphemes from monolingual data.

For the approach to better generalise to a variety of morphemes, one can use morpheme frequencies averaged over their grammatical category:

$$\hat{t} \approx \begin{cases} T_{cont}, P(g(x)) > \gamma \\ T_{sys}, \text{otherwise} \end{cases} \tag{2}$$

Once the T_{sys} system morpheme types are obtained, the ML can be predicted effortlessly using the expression from the beginning of Section 3.

4 Experiments

In this section, the methods for discovering the system morphemes are applied. It is important to note that the experiments in this section are conducted at the word level as an approximation to morpheme-level tokenisation. This is done because there are no existing tools or methods to determine grammatical properties of morphemes. The objective is therefore to mine **system words** rather than system morphemes, which serve the same function. Furthermore, the ML is determined on the utterance level for conversational data and on the line and sentence levels for literary data as approximations of the CP-level analysis. This is also related to the limitation of resources and tools for reliable CP segmentation of texts.

4.1 Datasets

Both monolingual and CS datasets are used for the experiments below. For the joint POS+LID tagger training the Universal Dependencies 2.0 (Nivre et al., 2020) dataset is used for Mandarin, English and Spanish languages. The word distributions for the training, validating and testing of the model are given in Table 3. To discover system words from monolingual data, the train sets from the Fleurs dataset (Conneau et al., 2022) are used, and the statistics for the tokens are presented in Table 2.

Although monolingual data cannot directly capture the full dynamics of CS, this approach remains consistent with the assumptions of the MLF theory. The MLF framework conceptualises CS as a conversational production process in which elements from two languages are combined according to a set of structural constraints. Consequently, provided that the grammatical properties of the languages and the relevant constraints are known, monolingual data can still be used to model and analyse aspects of CS behaviour.

In order to test and validate the MLID approach with the extracted system words, real conversational CS data: SEAME (Lyu et al., 2010) and Miami[1] is used for testing and probability estimations. Utterance transcriptions that contain words from two languages: English/Mandarin (SEAME) or English/Spanish (Miami) are chosen for the analysis. CS SEAME subset consists of 57052 sentences and 766525, while CS Miami subset contains 292 sentences and 3589 words.

As a final step of the analysis, the applicability of the method is tested for literary data, namely, poetic and prose data. A small set of 1839 sentence fragments from prose and 1839 lines from poems are annotated for the level of language mixing: 1) monolingual (inter-sentential CS); 2) clear MLID with singular word insertions (adherent to MLF); 3) language alternation with no clear MLID (non-adherenet to MLF). The English/Spanish literary data was also annotated for the stylistic function of the sentence fragment/line: 1) thematic function - the main function of the fragment is to carry across some content, some meaning; 2) highlight - the main function is to emphasise a main idea in the piece; 3) emotion - the main function is to carry across an emotion. Finally, if the CS fragment has a clear MLID, it is determined.

Units of analysis on the level at which the ML is determined in prose data are automatically extracted based on punctuation to approximate CPs. The statistics for the dataset are given in the Table 1 below.

Language pair	Count	MLID
Poem English/Spanish	1293	119
Poem English/Mandarin	545	114
Prose English/Spanish	1839	174

Table 1: CS literary dataset statistics. "Count" column means fragments count for prose and lines count for the poetic data. "MLID" column means lines or fragments for which the MLID was annotated.

4.2 Joint POS and LID Training

It has been shown before that POS tagger models trained on monolingual data can generalise to CS in token classification tasks (Soto and Hirschberg, 2018; Ball and Garrette, 2018), although still underperforming compared to when trained on CS data. Therefore, monolingual English, Mandarin and Spanish datasets from the Universal Dependencies 2.0 are used for joint POS and LID training. The statistics for the splits are given in Section 4.1. For each token in the source sentence, a POS tag and the LID are recognised simultaneously.

To produce an English/Mandarin POS+LID predictor, a pre-trained multilingual BERT (Devlin et al., 2018) with 12 attention heads is fine-tuned on the training subset of the data mentioned above.

[1] https://biling.talkbank.org/access/Bangor/Miami.html

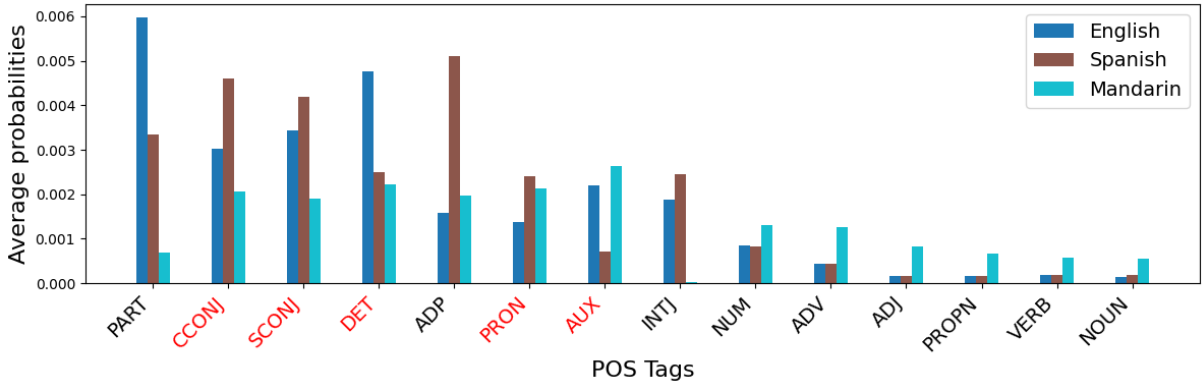


Figure 1: Average word probabilities estimates grouped by POS and sorted by the sum of the average across languages. POS highlighted in red are the POS which are conventionally believed to be represented by system words in linguistics and NLP (Myers-Scotton, 2002; Bullock et al., 2018).

The model is finetuned for 3 epochs with cross-entropy loss. The accuracies on the validation and test subsets are 94% and 93%, respectively, while the F1-scores are 94% and 92%. Calculating the performance metrics on Miami gives F1 score of 80%, which supports the earlier claims of the applicability of monolingual POS systems to CS.

4.3 Data-Driven Discovery of System Words

4.3.1 Average Token Probabilities Estimated from Monolingual Data

For the first experiment, the method described in Section 3.1 is applied to the monolingual Fleurs dataset for three languages: English, Mandarin and Spanish. POS tags are recognised for each of the sentences in the corpora using the joint POS+LID tagger described above. The word probabilities are estimated, and average token probabilities are calculated for every POS tag. Finally, the average probabilities are summed across the three languages to give a per POS tag probability. These are then sorted to demonstrate the similarity with the conventional system word set mentioned in linguistic and NLP literature (Figure 1).

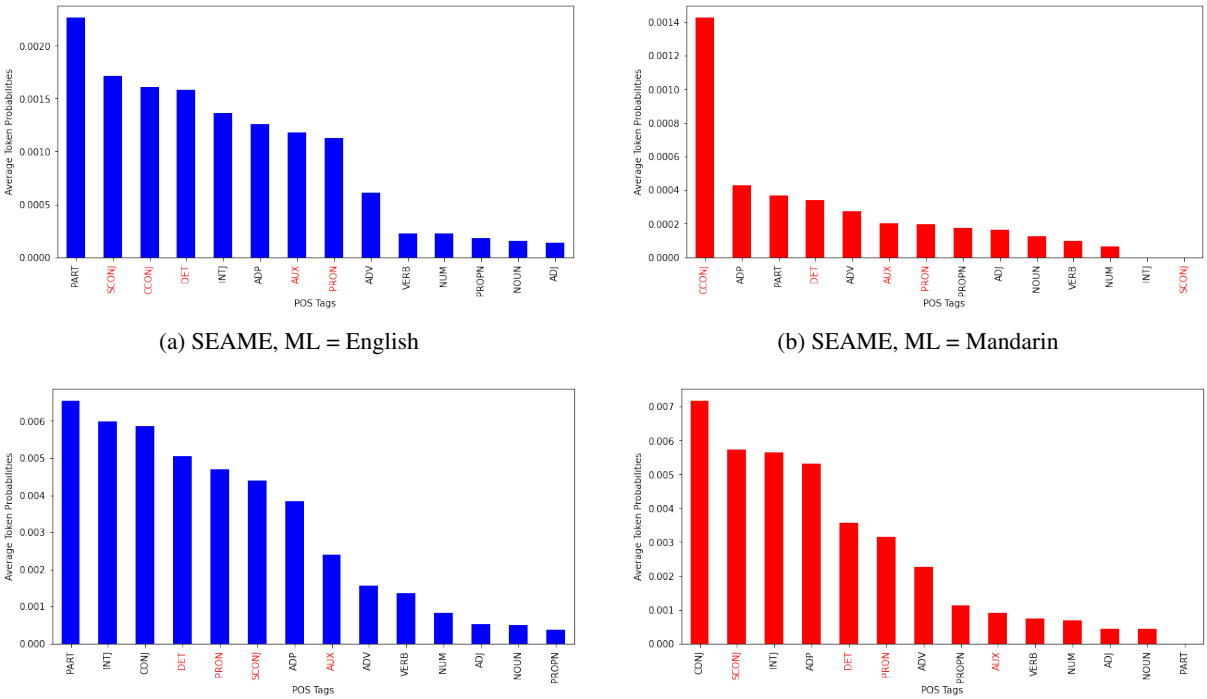
From Figure 1, it can be observed that the conventional grammatical categories that are typically represented by system words auxiliaries (AUX), determiners (DET), coordinating conjunctions (CCONJ), subordinating conjunctions (SCONJ) and pronouns (PRON) seem to be located in the top half of the sorted list. Apart from the conventional aforementioned grammatical categories, particles (PART) and adpositions (ADP) seem to have average probabilities which are comparable to those of the conventional grammatical probabilities.

One way to identify system words is to use mean token probabilities for a given POS. To be specific, the expectation of token probabilities can be used as an indicator of which language is the ML in the utterance. In this way, for each of the three languages, there is an estimated probability ranking for POS, and the top N POS are extracted accordingly. These top N POS suggest the highest grammatical function in language, which is strongly linked to the ML. Examples of the extracted POS sets are given in the Table 4, which will be discussed later in more detail.

4.3.2 Average Token Probabilities Estimated from CS Data

The same approach as above can be applied to a subset of real CS data where the ML can be determined using the MOP method described in Iakovenko and Hain 2024. Similar to Fleurs, token probabilities are estimated and then averaged over POS, but, contrary to the experiment above, averaging of the probabilities is carried out only for the tokens for which the LID is equal to the ML determined using MOP. This means that probabilities are calculated only for the tokens from the ML in CS data. The resulting rankings of POS are displayed in Figures 2a and 2b for SEAME and in Figures 2c and **??** for Miami.

Unlike monolingual, the POS, which are typically represented by system words, are less aligned with average probability rankings. This means that fewer POS that were used for ML determination in earlier studies can be seen at the top of the ranking. However, some conventional system POS still lead in the rankings, such as CCONJ for SEAME



(a) SEAME, ML = English

(b) SEAME, ML = Mandarin

(c) Miami, ML = English

(d) Miami, ML = Spanish

Figure 2: Average token probabilities grouped by POS according to MOP for different Matrix Languages (ML) in the SEAME and Miami corpora.

when the ML is Mandarin, and SCONJ for Miami when ML is Spanish. Furthermore, some similarities with the monolingual data are observed, for example, the leading tendencies of PART and ADP, which may be a reason enough to consider words which belong to these POS as system words.

4.3.3 Measurement of Performance on the ML Determination Task

To measure if the extracted POS can indicate the ML in a CS sentence, they are tested as the X_{sys} set in the proposed SMP method (Section 3). The outcomes of the proposed SMP method with different sets X_{sys} are compared to the baseline approach, where system words are represented by 5 conventional POS (DET, AUX, CCONJ, SCONJ, PRON) following Myers-Scotton 2002 and Bullock et al. 2018. Despite that MOP is highly accurate, and the outputs rarely deviate from human judgment (Iakovenko and Hain, 2024), MOP outputs are still machine-generated. Hence, instead of computing accuracy or F1 on SMP with respect to MOP, Matthew's Correlation Coefficient (MCC) is adopted as the evaluation metric in this experiment to compare the outcomes of the proposed SMP to the outcomes of MOP.

13 Top N POS sets are tested, and the best-performing Top N sets are chosen: 9 for SEAME and 8 for Miami. The best MCC values are obtained using these POS which are 0.22 for SEAME with top 9 extracted POS (a 0.07 increase from the conventional 5-POS baseline) and 0.33 for Miami with top 8 extracted POS (a 0.03 improvement from the baseline).

Language	Sentence count	Token count
English	2518	52602
Mandarin	3246	60622
Spanish	2796	68285

Table 2: Fleurs dataset statistics.

4.4 Applying SMP with the Novel System Word Extraction Method to Literary Texts

In this section, the proposed approach for discovering system words from monolingual texts is incorporated into the novel SMP framework and evaluated on literary data. The objective of these experiments is to examine whether CS and adherence to the MLF contribute to the style in poetic and prose texts. Establishing such relationships would enable the automatic detection of stylistic features, such as emphasis and emotional expression, in CS text. A literary text fragment is adherent to MLF

Language	Sentence count			Word count		
	train	dev	test	train	dev	test
English	32179	5110	7798	523806	76180	7798
Mandarin	7994	3054	3555	859067	93318	3555
Spanish	28474	1000	3147	197232	25326	3147

Table 3: Universal Dependencies 2.0 dataset statistics.

when the principles for ML determination: MOP or SMP, have been applied successfully. In contrast, e.g. when a fragment has singletons or system morphemes from both languages, the ML cannot be reliably determined; therefore, the fragment is not adherent to the MLF.

4.4.1 ML Determination for Literary Texts

Given the ground truth ML values for lines and fragments from poems and prose, correspondingly, one can calculate the following metrics: 1) coverage - the number of samples for which the rule was successfully applied, 2) accuracy - the number of samples predicted correctly, and 3) F1 macro - F1 scores averaged over each class, ensuring class balance during evaluation. Those metrics are calculated for the baseline (SMP using 5 conventional POS) and proposed (SMP using the extracted POS from Table 4) methods. The proposed method yielded a 5.05% absolute increase in coverage (from 59.63% to 64.68%) and a slight 0.82% increase in F1 macro score (from 81.97% to 82.79%) on poetic texts. As for the prose texts, the proposed method yielded a 1.61% increase in F1 macro score (from 86.36% to 87.97%), however showing a 8.19% decrease in coverage (from 63.16% to 54.97%). The decrease in prose data coverage is observed, indicating that fewer CS fragments can be successfully analysed by this variant. The non-ideal outcome can be attributed to the poor quality of the automatic fragment extraction process, which is designed to approximate CPs but does not achieve sufficient accuracy in practice.

4.4.2 Literary Texts Stylistic Function Analysis

Suppose that in literary texts CS may be used as a stylistic choice, conveying emotion or highlight. Then, a correlation analysis between the stylistic function and the language mixing manner can be conducted to quantify the interconnections between the two.

First, the correlations between lines and fragments annotated as monolingual and each of the stylistic function values are computed (Figure 3). As shown in the figure, CS sentences tend to express emotive or highlighting functions slightly more frequently than monolingual ones (0.089 vs 0.034 and -0.116) in poetic texts. In contrast, within prose, monolingual sentences exhibit a marginally higher association with the emotive function than with thematic or highlighting functions, though overall variation is smaller. This trend is expected: poetic language often employs figurative and creative literary devices to emphasise meaning, whereas prose tends to convey stylistic functions through lexical choice or narrative structure.

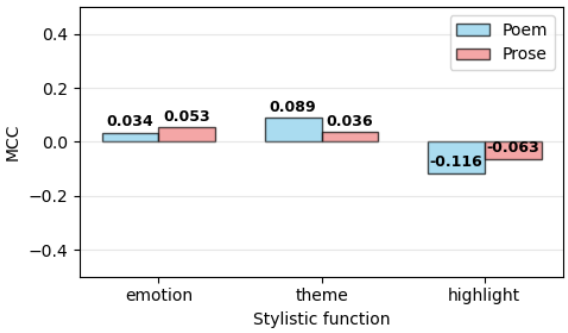


Figure 3: Correlation between ground truth monolinguality and stylistic function.

Subsequently, CS lines and fragments are automatically extracted using the joint POS+LID determination methods described in Section 4.2. The trends observed in the ground truth CS annotations largely persist in the automatically predicted monolingual segments: with 0.118 MCC of monolinguality and thematic stylistic function, compared to 0.039 of emotive stylistic function and -0.149 of a highlight. Similarly for prose, a marginal difference in MCC is observed, with the emotive stylistic function showing the highest correlation with monolinguality (0.06 vs 0.011 and 0.042).

The next stage of analysis examines whether adherence to the MLF theory correlates with the stylistic function of CS fragments. As shown in Fig-

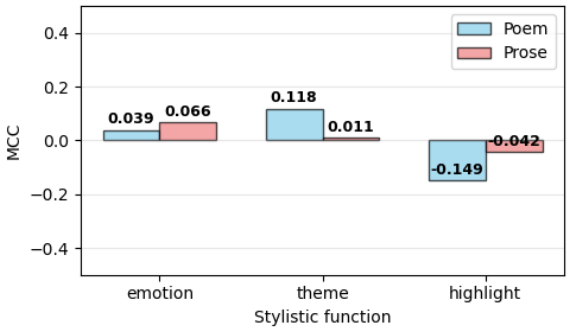


Figure 4: Correlation between predicted monolinguality and stylistic function.

ure 5, a positive correlation is observed between adherence to the MLF and the thematic stylistic function in prose texts, whereas fragments serving a highlighting function exhibit a more variable or "chaotic" pattern of CS. In contrast, for poetic texts, adherence to the MLF shows a stronger association with the highlighting function than with the thematic one, while CS of a more irregular nature (with no adherence to MLF) appears more frequently in fragments expressing emotional stylistic functions.

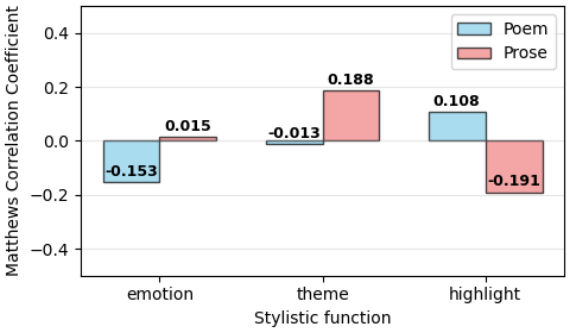


Figure 5: Correlation between ground truth compliance with MLF in CS and stylistic function.

Finally, assuming that adherence to the MLF (Section 4.4) can be automatically derived using the proposed extension of the SMP method, it becomes possible to quantify the correlation between the predicted MLF adherence and stylistic functions. For poetic texts, a positive correlation is again observed between adherence to the MLF and the highlighting stylistic function (0.128), while weaker or negative correlations are found for emotional (-0.067) and thematic (-0.09) functions. In contrast, the results for prose indicate no meaningful relationship between MLF adherence and stylistic functions, suggesting that, at this stage, the method is not suitable for predicting stylistic function based on MLF adherence in prose texts specifically.

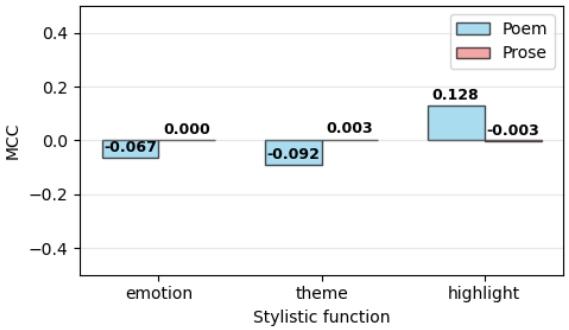


Figure 6: Correlation between predicted compliance with MLF in CS and stylistic function.

5 Conclusion

This study introduces a novel approach for identifying system morphemes in CS text within the MLF framework and applies it to improve the performance of the SMP for ML determination. The proposed method reveals a relationship between conventional system morphemes (pronouns, conjunctions, determiners and auxiliaries) and token frequency distributions across POS categories, while also highlighting the importance of particles and adpositions for ML determination. Leveraging monolingual data to identify POS categories functioning as system morphemes improves MCC from 0.15 to 0.22 for SEAME and from 0.29 to 0.33 for Miami. Furthermore, applying the method to literary data yields a 1.22% absolute improvement in macro-averaged F1 score and enables the analysis of relationships between CS behaviour and stylistic function. The results show that monolingual segments are more strongly associated with thematic stylistic functions, while adherence to the MLF exhibits a positive correlation with highlighting functions in poetic texts, though not in prose. Overall, these findings provide new insights into the grammatical and stylistic roles of system morphemes in CS, contributing toward more robust and interpretable computational models for multilingual language processing.

Limitations

The main limitation of the method is related to the data availability: there is limited ML-annotated CS data openly available to date. Therefore, it is prob-

lematic to assess the quality of ML classification and the feature importance. ML identity can be determined in CS data using the rule-based MOP principle, which has a high accuracy, but the principle can only be applied in the case of singleton EL insertions. Furthermore, the units of analysis have to be approximated: specifically CPs to sentences, utterances and lines, and morphemes to words. This leads to noisier conditions for analysis and overall less correlation.

Acknowledgments

References

Heike Adel, Ngoc Thang Vu, Katrin Kirchhoff, Dominic Telaar, and Tanja Schultz. 2015. Syntactic and semantic features for code-switching factored language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23:431–440.

Peter Auer and Raihan Muhamedova. 2005. Embedded language'and 'matrix language'in insertional language mixing: Some problematic cases. *Rivista di linguistica*, 17(1):35–54.

Kelsey Ball and Dan Garrette. 2018. Part-of-speech tagging for code-switched, transliterated texts without explicit language identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3084–3089, Brussels, Belgium. Association for Computational Linguistics.

Barbara Bullock, Wally Guzmán, Jacqueline Serigos, Vivek Sharath, and Almeida Jacqueline Toribio. 2018. Predicting the presence of a matrix language in code-switching. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 68–75, Melbourne, Australia. Association for Computational Linguistics.

Ching-Ting Chang, Shun-Po Chuang, and Hung yi Lee. 2019. Code-switching sentence generation by generative adversarial networks and its application to data augmentation. In *INTERSPEECH*.

S. A. Chowdhury, A. Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching arabic asr. In *Interspeech*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, and Karthik Sankaranarayanan. 2021. Mucs 2021: Multilingual and code-switching asr challenges for low resource indian languages. pages 2446–2450.

Chenpeng Du, Hao Li, Yizhou Lu, Lan Wang, and Yanmin Qian. 2021. Data augmentation for end-to-end code-switching speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 194–200. IEEE.

Injy Hamed, Mohamed Elmahdy, and Slim Abdennadher. 2018. Collection and analysis of code-switch Egyptian Arabic-English speech corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki. 2020. Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. In *INTERSPEECH*.

Ke Hu, Antoine Bruguier, Tara N. Sainath, Rohit Prabhavalkar, and Golan Pundak. 2019. Phoneme-based contextualization for cross-lingual speech recognition in end-to-end models. In *Interspeech 2019*, pages 2155–2159.

Olga Iakovenko and Thomas Hain. 2024. Methods of automatic matrix language determination for code-switched speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5791–5800, Miami, Florida, USA. Association for Computational Linguistics.

Grandee Lee, Xianghu Yue, and Haizhou Li. 2019. Linguistically motivated parallel data augmentation for code-switch language modeling. In *INTERSPEECH*.

Ke Li, Jinyu Li, Guoli Ye, Rui Zhao, and Yifan Gong. 2019. Towards code-switching asr for end-to-end ctc models. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6076–6080.

Dau-Cheng Lyu, Tien Ping Tan, Chng Eng Siong, and Haizhou Li. 2010. Seame: a mandarin-english code-switching speech corpus in south-east asia.

Caterina Marino, Carline Bernard, and Judit Gervain. 2020. Word frequency is a cue to lexical category for 8-month-old infants. *Current Biology*, 30(8):1380–1386.e3.

C. Myers-Scotton. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press.

Carol Myers-Scotton. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. OUP.

SOS Ncoko, Ruksana Osman, and Kate Cockcroft. 2000. Codeswitching among multilingual learners in primary schools in south africa: An exploratory study. *International Journal of Bilingual Education and Bilingualism*, 3(4):225–241.

Li Nguyen. 2025. Rethinking the matrix language: Vietnamese–english code-switching in canberra. *International Journal of Bilingualism*, 29(4):1067–1093.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection.

Madaki Rufai Omar. 1983. *A linguistic and pragmatic analysis of Hausa-English code-switching (Nigeria)*. University of Michigan.

David Sankoff and Shana Poplack. 1981. A formal grammar for code-switching. *Paper in Linguistics*, 14(1):3–45.

Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie. 2019. Investigating end-to-end speech recognition for mandarin-english code-switching. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6056–6060.

Victor Soto and Julia Hirschberg. 2018. Joint part-of-speech and language ID tagging for code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

Víctor Soto and Julia Hirschberg. 2019. Improving code-switched language modeling performance using cognate features. In *INTERSPEECH*.

A MMC between MOP and different SMP implementations

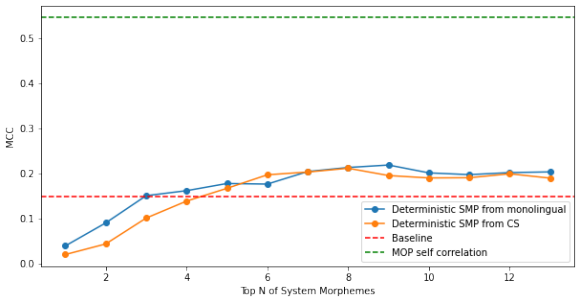


Figure 7: MCC for different SMP implementations on the SEAME dataset. The green dashed line represents the maximum MCC that could have been possible for the SMP implementation: it is not equal to 1 because MOP does not have 100% coverage. The red dashed line is the baseline implementation with 5 conventional POS.

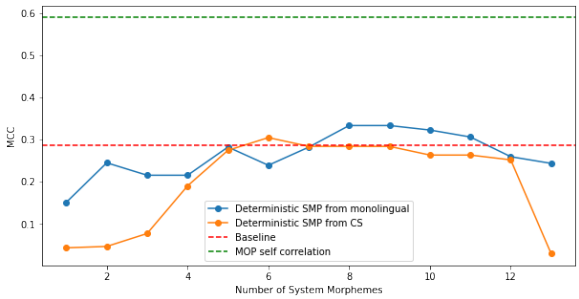


Figure 8: MCC for different SMP implementations for the Miami dataset.

Language	T_{sys}
English	[PART, DET, SCONJ, CCONJ, AUX, INTJ, ADP, PRON, NUM]
Mandarin	[AUX, DET, PRON, CCONJ, ADP, SCONJ, NUM, ADV, ADJ]
Spanish	[ADP, CCONJ, SCONJ, PART, DET, INTJ, PRON, NUM]

Table 4: Extracted grammatical categories of system morphemes for English, Mandarin and Spanish.