Discovering System Morphemes in Code-Switched Text

Anonymous ACL submission

Abstract

Code-switching (CS) is the process of speakers interchanging between two or more languages. The scarcity of CS data drives the development of advanced Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) models, which incorporate ideas from linguistic theory. To describe CS, the Matrix Language Frame (MLF) theory defines a Matrix Language (ML) as the language that provides the grammatical structure for a CS sentence. System morphemes are the type of morphemes that contribute to a sentence's grammatical structure and are used in the MLF theory for ML detection. The discovery of the system morphemes can improve automatic CS text generation and analysis. This work introduces several novel approaches for discovering system morphemes based on the MLF theory. Deterministic and predictive System Morpheme Principle (SMP) implementations are used to discover system words approximating system morphemes through the task of ML determination and prediction for SEAME and Miami datasets. The outputs from the two SMP implementations are compared to the outputs of the Morpheme Order Principle (MOP). Applying the deterministic SMP approach revealed that the conventional system words (pronouns, conjunctions, determiners, auxiliaries) fall into top 50% most frequent word frequencies when averaged over Part of Speech (POS). Moreover, the deterministic SMP has also revealed the ranking of the POS with respect to the ML determination task, showing the importance of particles and adpositions. Using the extracted system words in a CS text simulation task leads to a total of 3.3% fluency improvement, demonstrating the advantages of the statistical analysis of the linguistic properties of data in the deterministic SMP. This study provides valuable insight into the properties of tokens in relation to their grammatical categories in CS data.

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 postulates 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* 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. *The System Morpheme Principle* 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. Morphemes associated with POS, such as coordinating and subordinating conjunctions, auxiliaries, determiners and pronouns are commonly cited as system morphemes, yet the linguistic literature does not provide a definitive or comprehensive closed set. There are no known methods for automatic detection or determination of system morphemes. 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 implemented for NLP or ASR tasks.

In this paper, several novel approaches for discovering system morphemes based on the MLF theory are introduced. Two variations of the System Morpheme Principle (SMP) are developed that discover system morphemes through the task of ML determination and prediction. The Morpheme Order Principle (MOP) is used to assess the ML determination performance of these two SMP implementations. The correlation between the conventional system words that approximate system morphemes (pronouns, subordinating and coordinating conjunctions, determiners, auxiliaries) and word frequencies averaged over Part of Speech (POS) are analysed. The deterministic SMP revealed the ranking of the POS with respect to the ML determination task, demonstrating the big influence of conventional system words on the ML determination task as well as several new unconventional system words for English/Mandarin and English/Spanish CS. The deterministic SMP was used to reveal the ranking of the POS with respect to the ML determination task. A predictive SMP is also trained and compared to the performance of the deterministic SMP.

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 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 sequential (grammatical) 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 further to be 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: POS and cognate word pairs (Soto and Hirschberg, 2019), Brown word vectors (Adel et al., 2015). 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. 4-M model is a morpheme classification framework also introduced in Myers-Scotton 2002, which 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 way content/system morphemes were determined in these studies was usually a poorly argued design decision of the author. 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 poorly argued. 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 of EL POS distributions between the two languages acting as ML were not explored.

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, 2024). 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 this paper believe that it is essential to develop MLF-based methods for better understanding and modelling of CS. While pure MLF implementations may not describe any CS production, we believe that their relaxed modifications may be used to fit a variety of CS data (e.g. ML for parts of CP).

The above indicates the importance of identifying system morphemes, but it was not carried out before in the context of ML determination and other downstream tasks. Therefore, the objective of this study is to advance technologies for CS 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. 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 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 a set of indices of system morphemes in the CS CP $I_{sys} = \{i \mid x_i \in X_{sys}\}$, where X_{sys} is the system morpheme set, $\{x_i \mid i \in I_{sys}\}$ are the system morphemes which have grammatical relations external to their head constituent and $\{l_i = L_1 \mid i \in I_{sys}\}$, then L_1 is the ML and L_2 is the EL. 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 determining system morphemes or a set of 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.

3.1.1 Deterministic approach to SMP

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 $P(x)$:

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

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:

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

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

3.2 Predictive approach to SMP

Alternatively, a predictive approach to determining ML can be defined. Two additional sequences can be derived from CS CP $\mathbf{x}$: a sequence of grammatical categories of morphemes $\mathbf{g} = [g_1, .., g_n]$, $g_i \in G$, and a sequence of morpheme types following the 4-M model (Myers-Scotton, 2002) $\mathbf{t} = [t_1, .., t_n], t_i \in T_{sys} \cup T_{cont}$. All sequences can

be obtained using token classification algorithms and have the same length $|\mathbf{x}| = |\mathbf{g}| = |\mathbf{t}| = |\mathbf{l}|$. The following holds true: $\mathbf{x} \rightarrow \mathbf{g} \rightarrow \mathbf{t}$ and $\mathbf{x} \rightarrow \mathbf{l}$, where the arrow is a many-to-one correspondence. The textual representation $\mathbf{x}$ is language-dependent, while $\mathbf{g}$ and $\mathbf{t}$ are language-independent. Since morpheme types can be unambiguously derived from the grammatical category of a morpheme, $\mathbf{t}$ can be substituted with $\mathbf{g}$ when trying to recognise the ML L:

$$P(L|\mathbf{t}, \mathbf{l}, \theta) = P(L|\mathbf{g}, \mathbf{l}, \theta) \quad (3)$$

With $P^t(L|g, l, \theta_t)$, one can try to recognise the textual ML from the number of occurrences of a singular grammatical category and language combination (g^t, l^t). Consider a model $P^t(L|g, l, \theta_t)$ which predicts ML from a single feature (g^t, l^t). Then, running evaluation with a chosen metric m for a test CS dataset $D_t = [(g_1^t, l_1^t, L_1), .., (g_m^t, l_m^t, L_m)]$ one can calculate feature importance f_t for the task of ML determination:

$$\begin{aligned} f_t = m((&\arg\max_{L_i \in \mathcal{L}} P^t(L = L_i | g_1^t, l_1^t, \theta^t), ..., \\ &\arg\max_{L_i \in \mathcal{L}} P^t(L = L_i | g_m^t, l_m^t, \theta^t)), \\ &(L_1, ..., L_m)) \end{aligned} \quad (4)$$

Obtaining f_t values for all (g^t, l^t) combinations will result in feature importances $[f_1, .., f_t] = F$ may then be used as the content-to-system morpheme scale for a specific language mix and approximate morpheme types $T_{sys} \cup T_{cont}$.

4 Experiments

In this section, the methods for discovering the system morphemes are applied. It is important to highlight that the experiments in this section are carried out on a word-level as an approximation of morpheme-level tokenisation. This is done because grammatical categories of morphemes (e.g. POS tag) are ambiguous, and there are no existing tools or methods to determine grammatical categories of morphemes reliably. Therefore, the objective is to find system morphemes equal to whole words that act as ML markers, and those will be called **system words** from now onwards. Furthermore, the ML determination is carried out on the sentence level as an approximation 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 token distributions for the training, validating and testing of the model are given in Table A3 in the Appendix section. 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 A2 in the Appendix section.

In order to train, test and validate an automatic ML detector from POS+LID tags data is simulated using the 15349 semantically aligned monolingual sentences from the GALE corpus (Liu et al., 2010). Semantic alignment in GALE is the alignment of syntactically congruent words. Real CS data: SEAME (Lyu et al., 2010) and Miami[1] is used for testing and probability estimations. Sentences that contain tokens from two languages: English/Mandarin (Miami) or English/Spanish (Miami) are chosen for the analysis. The statistics for the two CS datasets is given in Table 1

Table 1: CS datasets statistics.

Language	Sentence count	Token count
SEAME	57052	766525
Miami	292	3589

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 monolingual 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 train an English/Mandarin POS+LID predictor, a pretrained multilingual BERT (Devlin et al., 2018) with 12 attention heads is finetuned on the train subset of the data mentioned above. 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

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

F1-scores are 94% and 92%. Calculating the performance metrics on Miami gives F1 score of 80% which supports the earlier claims of 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 3.1.1 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 some 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.

Suppose that the expectation of the token probability that belongs to a certain POS can be used as an indicator for the ML which is present in a CS sentence, then the top N POS can be extracted for each of the three languages from the estimated rankings. Examples of the extracted POS sets are given in the Appendix Table A4, 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. The resulting rankings of POS are displayed in Figures A4 and A5 for SEAME and in Figures A6 and A7 for Miami in the Appendix section.

Although in the case of CS the POS which are conventionally represented by system words are less aligned with average probability rankings, some conventional system POS still lead in the rankings such as CCONJ for SEAME 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 deterministic SMP method (Section 3.1.1). The outcomes of the deterministic SMP method with different sets X_{sys} were 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. The results are presented in Figure A8 for SEAME and Figure A9 for Miami in the Appendix section, where the top N selected POS varies from 1 to 14. The metric for measuring the agreement is Matthew's Correlation Coefficient (MCC) because the outcomes of deterministic SMP are compared to outcomes of MOP. It is not appropriate to use such measures as Accuracy or F1 in this task because MOP outputs are also machine generated, although it is highly accurate and the outputs rarely deviate from human judgment (Iakovenko and Hain, 2024).

In Figure A8 of the Appendix one can see how MCC first increases as the top N increases: this is due to SMP becoming more accurate as the number of top POS for analysis increase. Around 6-9 top N the SMP implementations reach their optimal performance which means that the top N selected usually do not get translated into the EL. After the best 6-9 top N a slight decrease in the MCC values can be observed due to the rest of POS (e.g. nouns or verbs) being used in both ML and EL more frequently and therefore influencing the decision in SMP less or even cause errors.

From the line plots it can be observed that the best results are obtained using monolingual data to

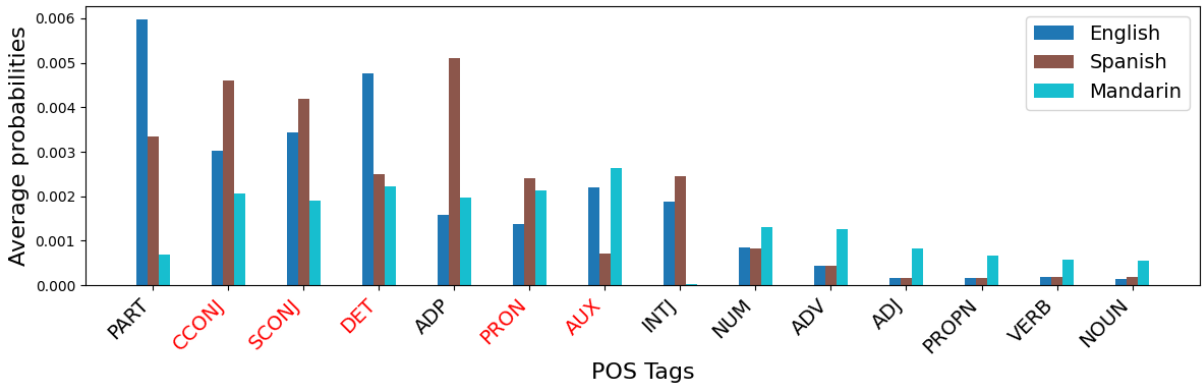

Figure 1: Average word probabilities estimates grouped by POS and sorted by the sum of the average across languages. POS highlighted in red are the POS which are conventionally believed to be represented by system words in linguistics and NLP (Myers-Scotton, 2002; Bullock et al., 2018).

extract grammatical categories that system words belong to. The best performing top N are 9 for SEAME and 8 for Miami. The ability to utilise monolingual data to estimate system words provides advantages when dealing with low-resource or zero-resource data. The extracted POS, which provide the system words for the ML, are displayed in the Appendix Table A4. 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).

4.4 Model-driven discovery of system words

In this section, the application of the predictive SMP towards system word discovery is described. The components are described below in detail as well as the datasets used and their construction.

There is limited ML annotated data available. Although simulated data can be generated using the MLF theory, it is avoided in this part of the experiment to ensure that the influence of the MLF approach is limited strictly to the system word discovery procedure. Therefore, a possible option is to generate a simulated dataset following the EC method described in Rizvi et al. 2021. To use the method, a dependency-level alignment of translations is needed, which is present in the GALE corpus for NMT. For each sentence pair, alignments with semantic links are used to translate parts of sentences from ML to EL. A sentence may have more than one substitution of such substitutions from ML to EL. 100974 simulated CS sentences are generated from the original 15349 sentences of the GALE corpus. The resulting simulated CS sentences are then split into train (114832) and test (26283) subsets. POS tags are generated for all of the above subsets using the POS+LID tagger described previously (Section 4.2) and used as an input for the SMP ML predictor below.

The same baseline determiner as in Section 4.3 that follows the deterministic approach to SMP (Section 3.1.1) and determines the ML based on the 5 conventional POS in a CS sentence is applied to the test subset of the simulated CS data. The system yields 74% accuracy with 24% of CS sentences determined as an "unknown language". 24% test sentences are marked with the "unknown language" label because the SMP method does not have 100% coverage due to some CS sentences containing system words from both languages or not having any system words from any languages. Therefore one of the goals of applying a predictive approach to SMP is to maximise the number of CS sentences for which ML can be determined.

In contrast to the baseline system, a decision tree classifier (DT) is trained to determine pseudo-ML identity (the language of the original non-translated sentence) from POS tags generated from simulated CS data. The classifier yields 98% accuracy on the simulated CS test set while maintaining 100% coverage rate.

4.4.1 Agreement analysis

In order to analyse the properties of the implemented SMP predictor on real CS data agreement analysis for SMP and MOP is carried out. In this experiment only the SEAME dataset is analysed because no English/Spanish translation dataset is manually aligned by dependency groups. Similarly

to the prior experiments, the agreement is measured by MCC. The obtained MCC of 0.18 is higher in comparison to the baseline (MCC=0.15), which appears to show the usefulness of the predictive method for real CS data. However, the method does not seem to outperform the deterministic SMP approach when the POS that are typically represented by system words are derived from monolingual data (MCC=0.22 when top 10 POS are used).

4.4.2 Feature importance analysis

While in Section 4.3 dataset statistics were estimated separately and explicitly for the deterministic SMP approach, in the predictive SMP approach the importance of POS are determined implicitly from task execution performance (Section 3.2). It is possible to measure Gini importance of the classifier, but upon calculation, it only revealed the nature of the simulated data which is not the goal of the experiment. A better strategy for determining the importance of specific (POS, lang) pairs generated from CS text is to train several separate ML classifiers for each of the (POS, lang) features. Having multiple classifiers one can calculate the feature that obtains the highest agreement measured in MCC on real CS data (Figure 2).

MCC values for the two ML determination approaches executed on real CS data are shown in Figure 2. The overall importance of each of the individual features seems to form three groups with noticeable step-changes in MCC. This is visible between Mandarin adverbs (ADV) and English verbs (VERB), and also between English CCONJ and English PRON. However the same tendencies of the conventional system word grammatical categories being important for ML prediction task cannot be observed to the same extent as with deterministic SMP: while English SCONJ and DET, and Mandarin DET and AUX seem to have a big impact on the ML prediction task, the rest of the POS show little to no impact.

The little impact of Mandarin CCONJ and PRON, and English AUX, PRON and CCONJ in the predictive SMP can be attributed to the difference in the training data and the model used. Although EC can facilitate the creation of natural-looking CS sentences, it might not necessarily be representative of the real CS data. Using both EC and MLF theory inspired data simulations would improve the scores beyond the deterministic SMP performance.

4.5 Using discovered system words for creating fluently sounding simulated CS

For the final experiment, a small classifier for automatic fluency assessment is trained on 697 English/Mandarin and English/Spanish CS examples from the CS-Fleurs dataset (Yan, 2025). A fully-connected neural network is trained for 5 epochs with cross-entropy loss and Adam optimiser (Kingma and Ba, 2014) similar to the approach in Kodali et al. 2025. The learning rate is set to 0.01 and the size of the single hidden layer is 50. Such a small neural network was chosen due to the small size of the training set and the noisy nature of the fluency assessment data. The accuracy of the classification into 3 classes of fluency reaches 33.57% on the validation set.

To demonstrate the usefulness of the extracted grammatical system word types, Large Language Model (LLM) (OpenAI ChatGPT version 4o) is prompted with and without system word constraints, similarly to (Kuwanto et al., 2024). System words are extracted from each original non-English monolingual sentence according to the extracted system POS approach (Appendix Table A4). The prompts used are presented in the Appendix Table A5. The generated CS samples for the three approaches are then scored using a trained classifier for automatic fluency. The results are presented in Figure 3.

As can be seen from the figure, the number of simulated CS sentences which are classified as "Unnatural" reduces with the introduction of generic system words in the prompt by 2%, and then is reduced by another 1.3% with the proposed approach when the grammatical categories (POS) of system words are extracted from monolingual data totaling 3.3% improvement from the baseline. This highlights the usefulness of system word identification for text generation in general, and the necessity of ML-based system word determination in contrast to using a fixed system word set.

5 Conclusion

This study introduces several novel approaches for identifying system morphemes in code-switched (CS) text based on the Matrix Language Frame (MLF) theory. Two variations of the System Morpheme Principle (SMP) have been developed to discover system morphemes through the task of ML determination and prediction. To assess ML determination performance across different feature

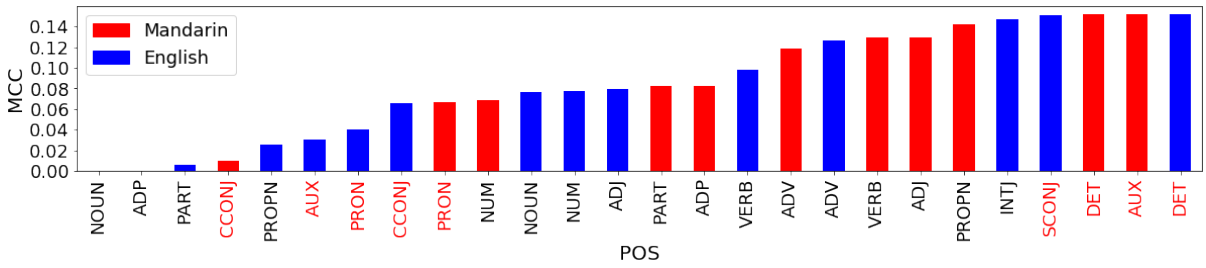

Figure 2: MCC of MOP and predictive SMP outputs on SEAME data. Predictive SMP uses single feature input.

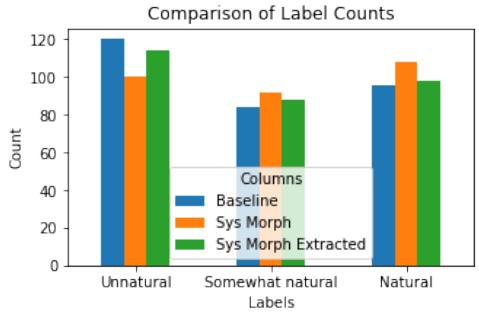

Figure 3: Classification of the CS examples simulated by 3 LLM-based approaches.

sets, the Morpheme Order Principle (MOP) from MLF theory is utilised.

The methods are implemented on a word-level as an approximation of the morpheme-level system morpheme discovery. The proposed deterministic approach highlights a correlation between conventional system morphemes such as pronouns, conjunctions, determiners and auxiliaries, and token frequency averages across Part-of-Speech (POS) categories. It also ranks POS in terms of their importance for ML determination, emphasising the significance of particles and adpositions. Utilising monolingual data to identify POS categories functioning as system morphemes resulted in a 0.07 improvement in Matthew's Correlation Coefficient (MCC) for SEAME (from 0.15 to 0.22) and a 0.04 increase for Miami (from 0.29 to 0.33). Additionally, an alternative predictive SMP model achieved a 0.03 MCC improvement (from 0.15 to 0.18), demonstrating the benefits of linguistic analysis in the deterministic SMP method, leading to higher MCC. Finally, the extracted system morphemes are used in a CS text simulation task, leading to a total of 3.3% fluency improvement, demonstrating the advantages of the statistical analysis of the linguistic properties of data in the deterministic SMP.

Overall, this study provides valuable insights into the relationship between token properties and their grammatical roles in CS data. The presented findings contribute to a deeper understanding of system morphemes and their role in ML determination, paving the way for more accurate computational models in multilingual language processing.

6 Limitations

The main limitation of the method is related to the data availability: there is no ML-annotated CS data openly available to date. Therefore, it is problematic 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. Since there is no ML annotation, simulated data has to be leveraged, but its usage is limited as shown in the paper and additionally requires dependency-aligned parallel data.

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. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Prashant Kodali, Anmol Goel, Likhith Asapu, Vamshi Krishna Bonagiri, Anirudh Govil, Monojit Choudhury, Ponnurangam Kumaraguru, and Manish Shrivastava. 2025. From human judgements to predictive models: Unravelling acceptability in code-mixed sentences. *Preprint*, arXiv:2405.05572.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2024. Linguistics theory meets llm: Code-switched text generation via equivalence constrained large language models. *arXiv preprint arXiv:2410.22660*.

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.

Yi Liu, Pascale Fung, Yongsheng Yang, Denise DiPersio, Meghan Glenn, Stephanie Strassel, and Christopher Cieri. 2010. A very large scale Mandarin Chinese broadcast corpus for GALE project. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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. In *INTERSPEECH*.

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. 2024. Rethinking the matrix language: Vietnamese–english code-switching in canberra. *International Journal of Bilingualism*, page 13670069241254454.

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. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211, Online. Association for Computational Linguistics.

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*.

Brian et al Yan. 2025. Cs-fleurs: A massively multilingual and code-switched speech dataset. In *Accepted to Interspeech 2025*.

A Appendix

Table A2: Fleurs dataset statistics.

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

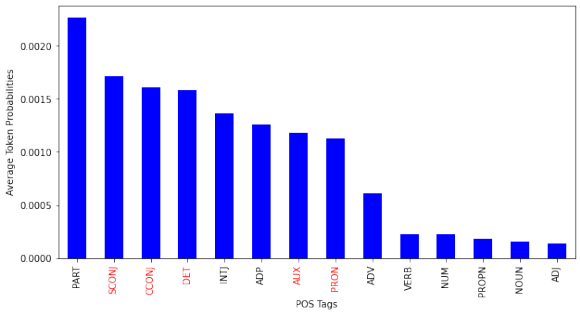

Figure A4: Average SEAME token probabilities grouped by POS for when the ML is English according to MOP.

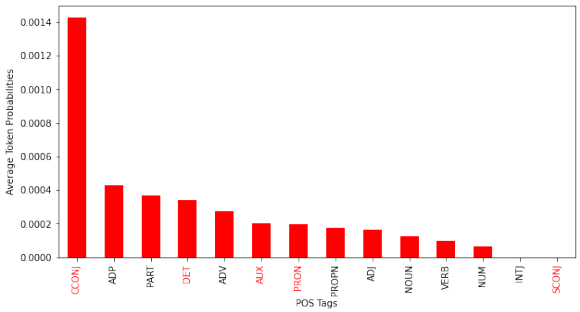

Figure A5: Average SEAME token probabilities grouped by POS for when the ML is Mandarin according to MOP.

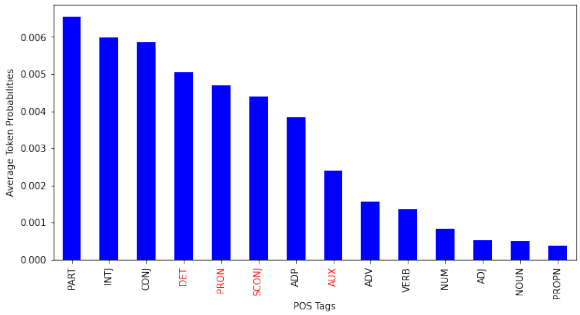

Figure A6: Average Miami token probabilities grouped by POS for when the ML is English according to MOP.

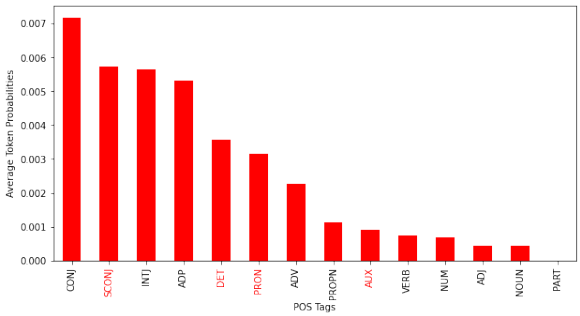

Figure A7: Average Miami token probabilities grouped by POS for when the ML is Spanish according to MOP.

Table A3: Universal Dependencies 2.0 dataset statistics.

	Sentence count			Token count		
Language	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 A4: Extracted grammatical categories of system morphemes for English, Mandarin and Spanish.

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]

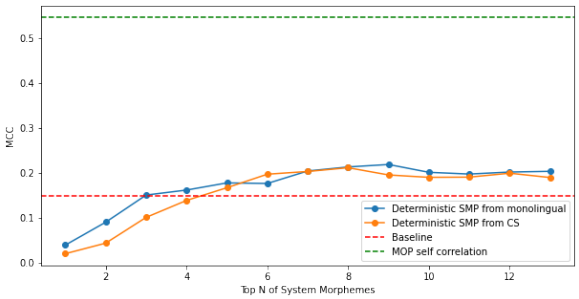

Figure A8: MCC for different SMP implementations on the SEAME dataset. The green dashed line represents the maximum MCC that could have been possible for the SMP implementation: it is not equal to 1 because MOP does not have 100% coverage. The red dashed line is the baseline implementation with 5 conventional POS.

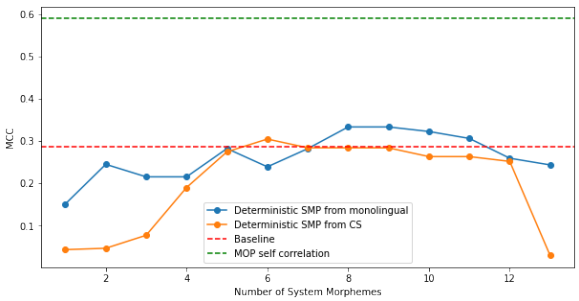

Figure A9: MCC for different SMP implementations for the Miami dataset.

Table A5: Prompts used for the CS text simulation experiment. *Extracted* X_{sys} is the proposed approach which uses extracted system morphemes to guide CS text simulation.

Approach	Prompt
Baseline	You are a Bilingual Spanish-English speaker, you will help translate this Spanish passage into a code-mixed Spanish-English passage. Translate the following passage into a code-mixed Spanish-English passage: Los expertos llegaron a la conclusión de que las materias oscuras se afectan entre sí al igual que lo hace la materia regular. Respond with a code-mixed sentence. Do not explain.
Fixed X_{sys}	You are a Bilingual Spanish-English speaker, you will help translate this Spanish passage into a code-mixed Spanish-English passage, considering the matrix language is Spanish, and the following words are system morphemes are in the passage: la, las, lo, los, que, se; translate the following passage into a code-mixed Spanish-English passage: Los expertos llegaron a la conclusión de que las materias oscuras se afectan entre sí al igual que lo hace la materia regular. Respond with a code-mixed sentence. Do not explain.
Extracted X_{sys}	You are a Bilingual Spanish-English speaker, you will help translate this Spanish passage into a code-mixed Spanish-English passage, considering the matrix language is Spanish, and the following words are system morphemes are in the passage: a, al, de, entre, la, las, lo, los, que, se; translate the following passage into a code-mixed Spanish-English passage: Los expertos llegaron a la conclusión de que las materias oscuras se afectan entre sí al igual que lo hace la materia regular. Respond with a code-mixed sentence. Do not explain.