Discovering of Grammatical Matrix Language Markers in Code-Switched Text

Anonymous ACL submission

Abstract

Code-switching (CS) is the process of speakers interchanging between several languages. CS is a complex process. To better describe CS speech the Matrix Language Frame (MLF) theory introduces the concept of a Matrix Language (ML), which is the language that provides the grammatical structure for a CS sentence. In this work several novel approaches for discovering system morphemes based on the MLF theory were introduced. Deterministic and predictive variations of the System Morpheme Principle (SMP) were developed to discover system morphemes through the task of ML determination and prediction. Morpheme Order Principle (MOP) from the MLF theory was used to assess the ML determination performance from the two SMP implementations. The deterministic approach revealed the correlation between the conventional system morphemes (pronouns, conjunctions, determiners, auxiliaries) and token frequencies averaged over Part of Speech (POS). Moreover, the deterministic approach has also revealed the ranking of the POS with respect to the ML determination task, showing the importance of particles and adpositions. Using monolingual data for discovering the POS that act as system morpheme types has led to a 0.07 Matthew's Correlation Coefficient (MCC) increase compared to the baseline for SEAME and a 0.04 increase for Miami. A predictive SMP was trained and has achieved 0.03 MCC increase 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 processing CS often yield poor performance in comparison to monolingual models. Given that in many countries CS is widespread (e.g India, South Africa, Nigeria) (Diwan et al., 2021; Ncoko et al., 2000; Rufai Omar, 1983), it is essential to develop Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) technologies for processing both CS speech and text.

In order to better describe the process of code-switching the Matrix Language Frame (MLF) theory was formulated (Myers-Scotton, 1997). It introduced 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 introduces two methods for ML determination: *The Morpheme Order Principle* (ML will provide the surface morpheme order for a CS sentence if it consists of singly occurring EL lexemes and any number of ML morphemes) and *The System Morpheme Principle* (all 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. Coordinating and subordinating conjunctions, auxiliaries, determiners and pronouns are actively discussed as the main POS of the system morphemes but a concise closed set is not given in the linguistic literature for a language variety. Furthermore, there are no known methods for automatic detection or determination of system morphemes. Bullock et al. 2018 explores if the same 5 POS can be used for automatic ML determination, however, no impact of the different combinations of POS was observed for the ML determination task.

MLF sets 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 token distributions but has been rarely

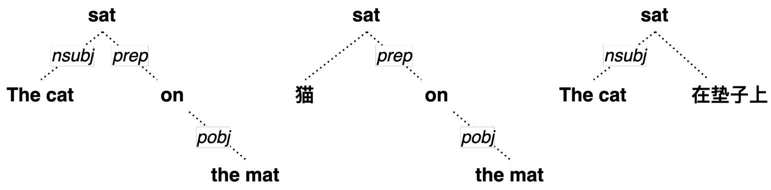

Figure 1: Example of CS simulation (original - left, synthetic - right).

implemented for NLP or ASR tasks. Some of the ideas from the MLF theory were implemented in Lee et al. 2019 and Hu et al. 2020 but the implementations are limited due to the absence of ML annotated data. Otherwise the usage of the MLF theory, specifically in the context of ML determination has been limited.

In this paper, several novel approaches for discovering system morphemes based on the MLF theory are introduced. Deterministic and predictive variations of the System Morpheme Principle (SMP) are developed to discover system morphemes through the task of ML determination and prediction. Morpheme Order Principle (MOP) from the MLF theory is used to assess the ML determination performance from the two SMP implementations. The correlation between the conventional system morphemes (pronouns, conjunctions, determiners, auxiliaries) and token frequencies averaged over Part of Speech (POS) are analysed. The deterministic approach 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 as follows. The next section provides a detailed description of the methods used. This is followed by a section on experiments, which provides information on datasets, detailed implementation, experiment descriptions as well as discussion of results. Conclusions summarise and complete the paper.

2 Methods for ML determination

Being called "principles for ML determination", the Morpheme Order Principle and the System Morpheme Principle in reality present three of the features of CS CP (projections of complementiser) which cannot be utilised 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) 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 $x_i, .., x_j \in \mathbf{x}$ system morphemes $x_i, .., x_j \in X_{sys}$ which have grammatical relations external to their head constituent and $l_i, .., l_j \in L_1$, then L_1 is the ML and L_2 is the EL. Below detailed descriptions of the SMP method variations are presented.

2.1 System Morpheme Principle (SMP)

Compared to MOP, there are fewer issues in formulating the principle when adapting SMP 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.

2.1.1 Deterministic approach to SMP

Let's 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 data. Then the amount of influence of a morpheme x on the grammatical structure may be approximated by the morpheme frequencies $P(x)$:

$$x \in X_{sys} \approx (P(x) > \beta) \tag{1}$$

where β is threshold for determining the system morpheme set X_{sys}. This approach may include the derivation of the system morphemes from monolingual data.

For the approach to better generalise to a variety of morphemes, especially for ideographic lan-

Table 1: 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

guages, one can use morpheme frequencies averaged over its grammatical category:

$$x \in X_{sys} \approx \left(\frac{1}{|T_G(x)|} \sum_{\hat{x} \in T_G(x)} P(\hat{x}) > \beta\right) \quad (2)$$

where $\hat{x} \in T_{POS}(x)$ are the tokens of the same grammatical category G as x. Once the X_{sys} system morpheme set is obtained the ML can be predicted effortlessly using the expression from the beginning of the Section 2.

2.2 Predictive approach to SMP

Alternatively, a predictive approach to predicting ML can be defined. Two more sequences can be derived from CS CP $\mathbf{x}$: grammatical categories of morphemes $\mathbf{g} = [g_1, .., g_n]$, $g_i \in G$ and 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: $\mathbf{x} \to \mathbf{g} \to \mathbf{t}$ and $\mathbf{x} \to \mathbf{l}$, where the arrow denotes sole dependency. 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 a trained model $P_t(L|g, l, \theta_t)$ one can try to recognise the ML identity from the number of occurrences of a singular grammatical category and language combination $|(g_t, l_t)|$. Then, for a test CS dataset $D_t = [(g_1, l_1, L_1), .., (g_m, l_m, L_m)]$ one can calculate feature importance f_t for the task of ML determination:

$$f_t = \prod_{i=1}^{|D|} P_t(L = L_i|g_i, l_i, \theta) \quad (4)$$

Once calculated for all (g_t, l_t) combinations resulting in feature importances $[f_1, .., f_t] = F$ may then be used as the "content-system" morpheme scale for a specific language mix and approximate morpheme types $T_{sys} \cup T_{cont}$.

3 Experiments

In this section the efforts towards discovering the system morphemes are described. 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 reliably determine grammatical categories of morphemes. As a result the objective is to find system morphemes which are equal to whole words that act as ML markers. 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.

3.1 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., 2017) dataset is used for Mandarin, English and Spanish languages following Soto and Hirschberg 2018. The token distributions for the training, validating and testing of the model are given in Table 1. To discover system morphemes 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.

Table 2: Fleurs dataset statistics.

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

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). Finally, real CS data: SEAME and Miami is used for testing and probability estimations. Sentences that contain tokens from two languages: English/Mandarin or English/Spanish accordingly are chosen for the analysis. The statistics for the two CS datasets is given in Table 3

Table 3: CS datasets statistics.

Language	Sentence count	Token count
SEAME	57052	766525
Miami	292	3589

3.2 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. Therefore for joint POS and LID training monolingual English, Mandarin and Spanish datasets from the Universal Dependencies 2.0 are used. The statistics for the splits are given in Section 3.1. 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 F1-scores are 94% and 92%. Calculating the performance metrics on Miami gives F1 score of 80% which supports the earlier claims of relative applicability of monolingual POS systems to CS.

3.3 Data-driven discovery of system morphemes

3.3.1 Average token probabilities from monolingual

For the first experiment the method from Section 2.1.1 is applied to monolingual Fleurs data for the 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 token probabilities are estimated and average token probabilities are calculated based on the POS tag. Finally, the average probabilities are summed across the three languages and sorted to demonstrate the similarity with the conventional system morpheme set mentioned in linguistic and some NLP literature (Figure 2).

From Figure 2 it can be observed that the conventional grammatical categories that are typically represented by system morphemes 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 Table 4 which will be discussed later in more detail.

3.3.2 Average token probabilities from CS

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 3 and 4 for SEAME and in Figures 5 and 6 for Miami.

Although in the case of CS the POS which are conventionally represented by system morphemes 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 morphemes which belong to these POS as system morphemes.

3.3.3 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

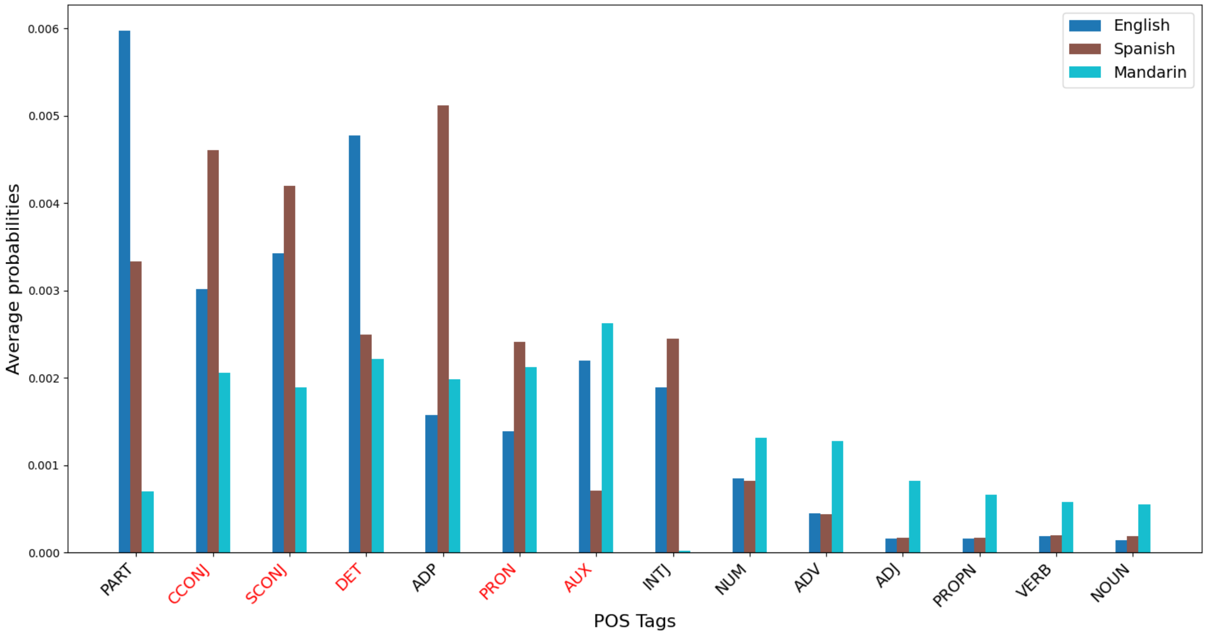

Figure 2: Average word probabilities grouped by POS and sorted by the sum of the average across languages. POS highlighted in red are the POS which are conventionally believed to be represented by system morphemes in linguistics and NLP (Myers-Scotton, 2002; Bullock et al., 2018).

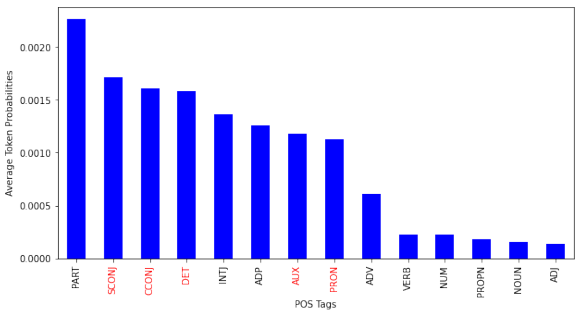

Figure 3: Average SEAME token probabilities grouped by POS for when the ML is English according to MOP.

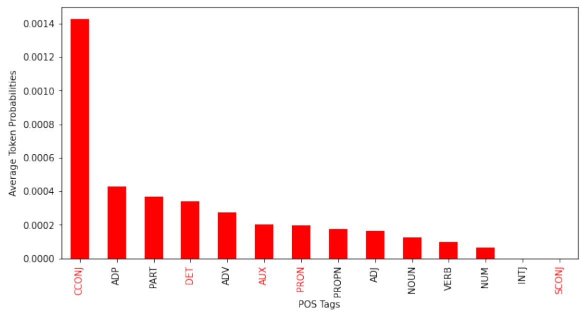

Figure 4: Average SEAME token probabilities grouped by POS for when the ML is Mandarin according to MOP.

in the deterministic SMP method (Section 2.1.1). The outcomes of the deterministic SMP method with different sets X_{sys} were compared to the baseline approach where system morphemes 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 7 for SEAME and Figure 8 for Miami where the top N selected POS varies from 1 to 14. The metric for measuring the performance 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 the figure 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

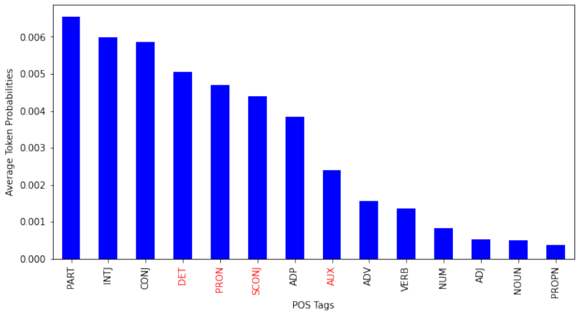

Figure 5: Average Miami token probabilities grouped by POS for when the ML is English according to MOP.

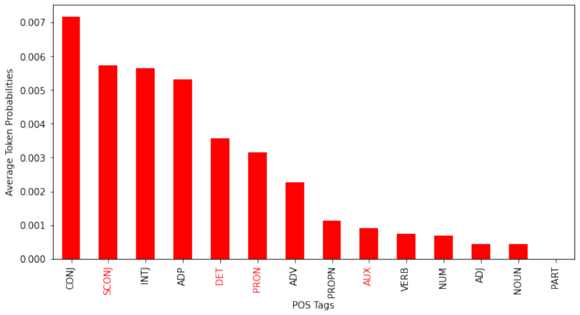

Figure 6: Average Miami token probabilities grouped by POS for when the ML is Spanish according to MOP.

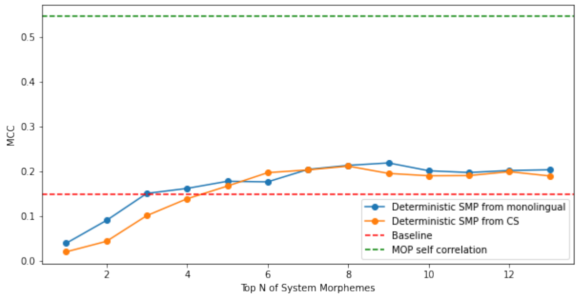

Figure 7: MCC for different SMP implementations for the SEAME dataset. The green dashed line represents the maximum MCC that could have been possible for the SMP implementation: it is not equal to 1 because MOP does not have 100% coverage. The red dashed line is the baseline implementation with 5 conventional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

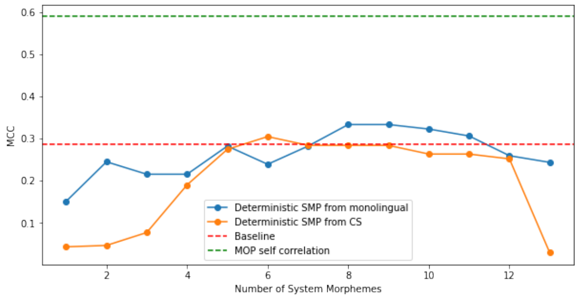

Figure 8: MCC for different SMP implementations for the Miami dataset.

best results are obtained using monolingual data to extract grammatical categories that system morphemes belong to. The best performing top N are 9 for SEAME and 8 for Miami. The ability to utilise monolingual data to estimate system morphemes provides advantages when dealing with low-resource or zero-resource data. The extracted POS which provide the system morphemes for the ML are displayed in Table 4. 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).

3.4 Model-driven discovery of system morphemes

In this section a trained approach towards SMP is described. The components are described below in detail as well as the datasets used and their construction.

There is no ML annotated dataset available, therefore a possible option is to generate a synthetic dataset following the Equivalence Constraint (EC) method described in Rizvi et al. 2021. In order to be able 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 3.2) and used as an input for the SMP ML predictor below.

The same baseline determiner as in Section 3.3 that follows the deterministic approach to SMP (Section 2.1.1) 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

Table 4: 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]

have 100% coverage due to some CS sentences containing system morphemes from both languages or not having any system morphemes 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.

3.4.1 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 morphemes are derived from monolingual data (MCC=0.22 when top 10 POS are used).

3.4.2 Feature importance analysis

While in Section 3.3 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 2.2). A trained DT-based SMP predictor is used to compute Gini importances for the (POS, lang) feature pairs of the classifier. The highest value of Gini importance is yielded by Mandarin coordinating conjunctions (CCONJ, Gini importance=0.86), while the remaining features have little or no impact (e.g. Mandarin adjectives=0.1, Mandarin numerals=0.02). This is to be expected because CCONJ are rarely aligned in dependency-aligned GALE data and therefore rarely translated following the EC-based CS simulation method. In this setup the Gini importance thus appears to tell more about the synthetic data generation process and not the actual influence of the POS tag on the ML identity decision.

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 best accuracy on simulated data (Figure 9) and the highest agreement measured in MCC on real CS data (Figure 10).

Upon looking at the accuracy values from Figure 9 one can observe the dominating role of the CCONJ for the ML prediction, in a similar fashion to the Gini importance analysis. The individual feature accuracies, unlike the Gini importances, indicate that Mandarin adjectives (ADJ) lead to almost the same amount of correctly recognised ML values as the aforementioned Mandarin CCONJ. Judging by the accuracies obtained on test CS GALE data, the most impactful English features for recognising ML are particles (PART) and adverbs (ADV).

Unlike the accuracy on synthetic GALE data, MCC values for the two ML determination approaches executed on real CS data show a different picture (Figure 10). The overall importance for each of the individual features seem 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 morpheme 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

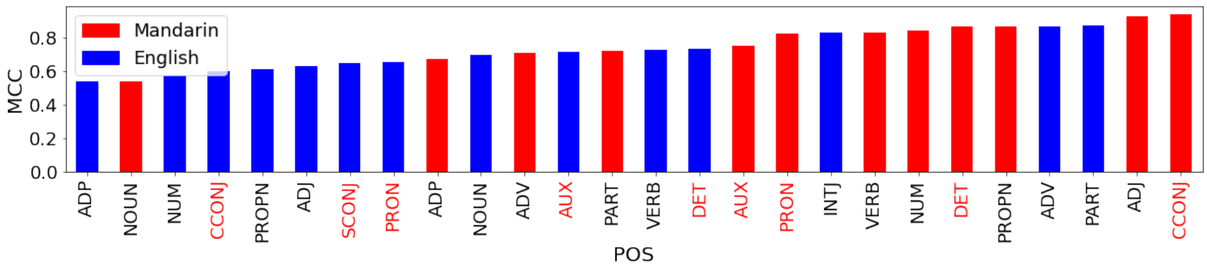

Figure 9: ML classification accuracy on the test subset of synthetic CS data. Predictive SMP uses single feature input.

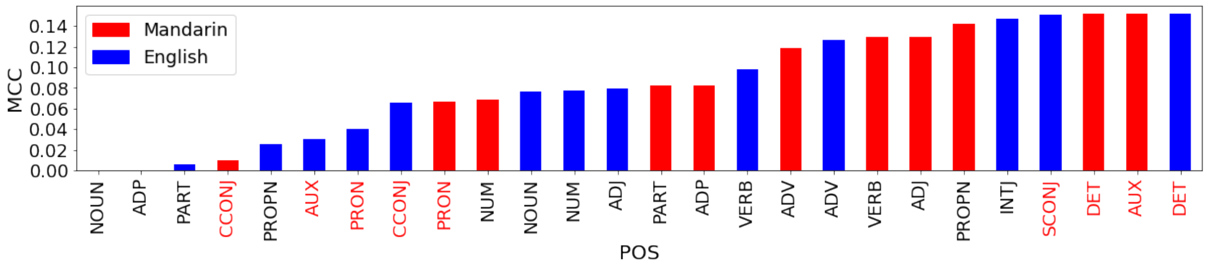

Figure 10: MCC of MOP and predictive SMP outputs on SEAME data. Predictive SMP uses single feature input.

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 Conclusion

This study introduces several novel approaches for identifying system morphemes in code-switched text based on the Matrix Language Frame (MLF) theory. Deterministic and predictive variations of the System Morpheme Principle (SMP) are developed to discover system morphemes through the task of ML determination and prediction. To assess ML determination performance across different feature sets the Morpheme Order Principle (MOP) from MLF theory is utilised.

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, emphasizing the significance of particles and adpositions. Utilizing 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 increase.

Overall, this study provides valuable insights into the relationship between token properties and their grammatical roles in code-switched 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.

Acknowledgements

This work was supported by Engineering and Physical Sciences Research Council.

5 Limitations

The main limitation of the method is related to the data availability: there is no ML-annotated CS data available to date. therefore it is problematic to assess the quality of ML classification and therefore the feature importance. ML identity can be determined in CS data using the MOP principle which has a high accuracy but the principle can only be applied in 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

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.

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.

Xinhui Hu, Qi Zhang, Lei Yang, Binbin Gu, and Xinkang Xu. 2020. Data Augmentation for Code-Switch Language Modeling by Fusing Multiple Text Generation Methods. In *Proc. Interspeech 2020*, pages 1062–1066.

Olga Iakovenko and Thomas Hain. 2024. Methods of automatic matrix language determination for code-switched speech. In *Submitted to Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

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.

C. Myers-Scotton. 1997. *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.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phng Lê Hồng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0. LINDAT/CLARIAH-CZ digital library at the

Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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.

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.