
Proceedings Paper:
Iakovenko, O. orcid.org/0000-0002-7801-6585 and Hain, T. orcid.org/0000-0003-0939-3464 (2024) Methods of automatic matrix language determination for code-switched speech. In: Al-Onaizan, Y., Bansal, M. and Chen, Y.-N., (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), 12-16 Nov 2024, Miami, Florida, USA. Association for Computational Linguistics , pp. 5791-5800. ISBN 9798891761643

https://doi.org/10.18653/v1/2024.emnlp-main.330

Methods for Automatic Matrix Language Determination of Code-Switched Speech

Olga Iakovenko and **Thomas Hain**
The University of Sheffield
{oiakovenko,t.hain}@sheffield.ac.uk

Abstract

Code-switching (CS) is the process of speakers interchanging between two or more languages which in the modern world becomes increasingly common. In order to better describe CS speech the Matrix Language Frame (MLF) theory introduces the concept of a Matrix Language, which is the language that provides the grammatical structure for a CS utterance. In this work the MLF theory was used to develop systems for Matrix Language Identity (MLID) determination. The MLID of English/Mandarin and English/Spanish CS text and speech was compared to acoustic language identity (LID), which is a typical way to identify a language in monolingual utterances. MLID predictors from audio show higher correlation with the textual principles than LID in all cases while also outperforming LID in an MLID recognition task based on F1 macro (60%) and correlation score (0.38). This novel approach has identified that non-English languages (Mandarin and Spanish) are preferred over the English language as the ML contrary to the monolingual choice of LID.

1 Introduction

Code-switching (CS) is the process of speakers switching between two or more languages in spoken or written language (Table 1). Spoken CS data is scarce, and thus models for processing CS speech 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 systems for understanding and modelling CS speech. One of the critical tasks in analyzing code-switched speech is determining the matrix language (ML), or the dominant language, which serves as the structural framework for the utterance. Accurate identification of ML is essential for various applications as well as sociolinguistic studies.

Table 1: An example of a CS utterance transcription from the SEAME dataset of colloquial Singaporean language.

毕业过后urh 你的study life 跟你的working life 有什么difference 吗

The linguistic Matrix Language Frame (MLF) theory (Myers-Scotton, 1997) provides a model for CS production and introduces the concept of a main, i.e. dominant language and a secondary, inserted language in CS utterances. These languages are ML and Embedded Language (EL), respectively. The MLF theory introduces two methods for ML determination:

1. *The Morpheme Order Principle* - ML will provide the surface morpheme order for a CS utterance if it consists of singly occurring EL lexemes and any number of ML morphemes

2. *The System Morpheme Principle* - all system morphemes which have grammatical relations external to their head constituent will come from ML

The morphemes as units within the MLF framework were first introduced by Myers-Scotton in 1997 (Myers-Scotton, 1997) and were split into **content** and **system** morphemes. Some common examples of system morphemes are quantifiers, possessives and tense/aspect determiners, while content morphemes include nouns, pronouns, adjectives, verbs and prepositions.

Matrix language identity (MLID) is the identity of the language providing the grammatical frame for the utterance and it can be defined for both monolingual and CS utterances. Moreover, the existence of ML implies a certain token distribution following the System Morpheme Principle (Myers-Scotton, 2002) which is highlighted in further Myers-Scotton works in the 4-M model. Overall, MLID provides insight into the grammatical

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5791–5800

properties of the utterance and a computational implementation of an MLID would be able to reduce the amount of manual annotation.

In this paper three MLID systems for CS text and audio were implemented. MLF theory formulates The Morpheme Order Principle and the System Morpheme Principle, which were implemented into three systems for MLID determination from text (P1.1, P1.2 and P2) and from audio ($MLID_{P1.1}$, $MLID_{P1.2}$ and $MLID_{P2}$). An extensive correlation analysis and comparison of an MLID determination system and a traditional acoustic language identities (LID) were carried out. Recognised MLID and LID from CS texts and audio were compared to ground truth ML annotation and the quality of ML recognition was measured in terms of F1 macro and Matthew's Correlation Coefficient (MCC). To conclude the findings, the distributions of textual LIDs were compared to the textual MLID distributions of the CS data.

The remainder of the paper is structured as follows. The next section reviews the relevant research presented previously on MLF and LID in CS text and speech. The third 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 a discussion of results. Conclusions summarise and complete the paper.

2 Related work

MLF theory has rarely been used to automatically analyse speech or text. Up until now it was only used for text augmentation (Bhat et al., 2016; Yilmaz et al., 2018; Lee et al., 2019) or for Language Model (LM) adaptation for code-switching. For example additional grammatical information was used during LM construction (Adel et al., 2015; Soto and Hirschberg, 2019) or a self-supervised training procedure was set up which encouraged generation of CS utterances (Chang et al., 2019). MLID classification for CS text was carried out in Bullock et al. (2018) where the ML was identified based on the token and system POS majorities.

Simultaneously in the speech processing domain a common technique to separate languages in CS is LID. LID of a whole CS utterance may be performed when CS is regarded as a separate language (Mary N J et al., 2020), in this case the component performs both LID and CS detection. A multilingual ASR system with an utterance-wise LID component as an auxiliary task was tested for CS utterances in Toshniwal et al. (2018) but the model was not able to generate CS text as a result. LID and language segmentation (LIS) systems make decisions based on similarity to the data they were trained on (Muralikrishna et al., 2021) and, to the best of our knowledge, no study was done to determine if pretrained LID/LIS are able to predict a dominant language in a CS utterance.

Neither MLID nor LID were previously used for CS analysis. Furthermore, ML determination principles were never fully implemented and compared. However, statistical methods were introduced before (Guzmán et al., 2017) which can assess the nature of CS. Among the statistical methods only the M-index (Multilingual Index) quantifies the representation of the languages in multilingual corpora. While the M-index is useful to learn about the balance of the token LIDs, it might be insufficient to learn about the utterance LID and MLID distributions.

The above indicates that theoretical methods to identify ML from text exist but previously there was only one attempt to determine MLID which was not based on the two ML determination principles. Furthermore, there are no existing MLID predictors from audio to the best of our knowledge. Therefore the objective of this study is to advance technologies for multilingual understanding and analysis by describing the implementation, comparison and performance of automatic MLID predictors from CS audio and text based on the ML determination principles from the MLF theory.

3 Principles for ML determination

The MLF theory principles mentioned above have to be implemented in order to compare the main language recognised by an LID system to the MLID. Each of them provides estimates for MLID but is conditioned by different evidence in the utterance. The Morpheme Order Principle is separated into two implementations of the 1st Principle (P1.1 and P1.2). The implementations of the principle deduction as per MLF theory are described below.

3.1 Principle 1.1: The singleton principle

The ML provides the context for singly occurring words from the EL, which will be further referred to as "singleton insertions". Although the original principle operates on the level of lexemes, the cur-

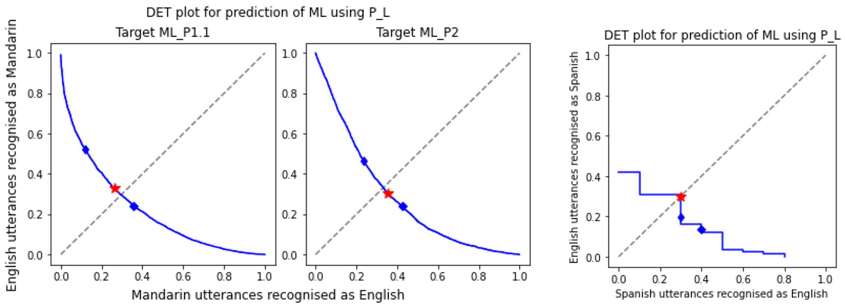

Figure 1: Detection error tradeoff (DET) curve for possible $\log \alpha$ values. Thin diamond is the default value of $\log \alpha = 0$, thick diamond - result of $\log \alpha$ estimation, red star - ground truth $\log \alpha$.

rent implementation operates on the level of words. Suppose there is a CS utterance $\mathbf{y}$ of length n, then $(\mathbf{y}', \mathbf{l}') = ((\varepsilon, l_\varepsilon), (y_1, l_1), .., (y_n, l_n), (\varepsilon, l_\varepsilon))$ are morphemes with corresponding language ID labels, ε is an empty morpheme and l_ε is an empty language morpheme tag from an empty language L_ε. If $((y_i, l_i), .., (y_j, l_j))$ constitute a word where $0 < i < j < n+1$, $\forall k$ that $i < k < j \mid l_k = L_2$ and $l_{i-1}, l_{j+1} \in L_1, L_\varepsilon$ then the language of the context L_1 is the ML while L_2 is the embedded language. For example in "哦你 post 在你的那个 blog" Mandarin will be ML since it is a context for English singleton insertions.

3.2 Principle 1.2: The token order principle

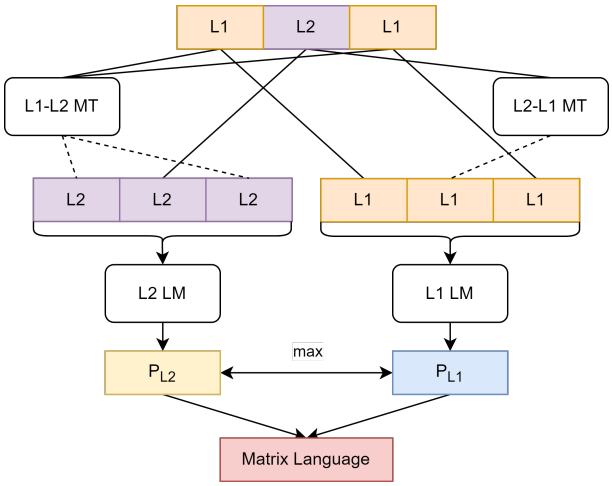

Figure 2: Pipeline of the morpheme order-based principle for ML determination P1.2.

In P1.2 the second part of the Morpheme Order Principle is implemented which postulates that the morpheme order is determined by the ML. For example, in "你觉得我们speak clear enough 吗" the English translation of the auxiliary Mandarin verb 吗will never appear at the end of an utterance in English, signifying that Mandarin is ML in this utterance. Assume languages $(L_1, L_2) \subset L$ are present in a bilingual utterance where L are all languages, then the original CS utterance $\mathbf{y}$ can be translated into two monolingual utterances $\hat{\mathbf{y}}_{L_1}$ and $\hat{\mathbf{y}}_{L_2}$. $\hat{\mathbf{y}}_{L_1}$ and $\hat{\mathbf{y}}_{L_2}$ are obtained from the original utterance $\mathbf{y}$ by a Neural Machine Translation (NMT) systems g_{L_1} and g_{L_2}. Consider an LM $P(L|\mathbf{y})$ which is used to provide a probability of the utterance belonging to a language L, then given the two languages L_1 and L_2 classification leads to:

$$\frac{P(\mathbf{y}|L_1)P(L_1)}{P(\mathbf{y})} \lessgtr \frac{P(\mathbf{y}|L_2)P(L_2)}{P(\mathbf{y})} \quad (1)$$

In the above the denominator may be eliminated. The probability $P(\mathbf{y}|L)$ may be estimated using independent monolingual LMs $P_L(\mathbf{y})$ and translation $\hat{\mathbf{y}}_L$ defined above $P(\mathbf{y}|L) \approx P_L(\hat{\mathbf{y}}_L)$ resulting in following:

$$\frac{P_{L_1}(\hat{\mathbf{y}}_{L_1})}{P_{L_2}(\hat{\mathbf{y}}_{L_2})} \lessgtr \alpha \quad (2)$$

Where α is the scaling factor for weighing the impact of the models. Taking the log 2 leads to:

$$\log P_{L_1}(\hat{\mathbf{y}}_{L_1}) - \log P_{L_2}(\hat{\mathbf{y}}_{L_2}) \lessgtr \log \alpha \quad (3)$$

Assume the difference of language log-probabilities can be expressed in terms of a factor α. This factor may be estimated by calculating the expectation of log-probabilities using utterances scored by monolingual LMs:

$$\log \alpha = \mathbb{E}\{\log P_{L_1}(\mathbf{y}_{L_1})\} - \mathbb{E}\{\log P_{L_2}(\mathbf{y}_{L_2})\} \quad (4)$$

All of the above leads to the following decision function:

Table 2: Monolingual dataset splits used for LM training in P1.2.

	Unit	Callhome English	Callhome Mandarin	Callhome Spanish
Token set size	tokens	6160	6853	3236
Train	utterances	20029	15827	19672
Valid		6030	3959	5500
Test		2609	1775	2665

$$\text{ML} = \begin{cases} L_1, \log P_{L_1}(\hat{\mathbf{y}}_{L_1}) - \log P_{L_2}(\hat{\mathbf{y}}_{L_2}) \geq \log \alpha \\ L_2, \text{otherwise} \end{cases} \tag{5}$$

A visual representation of the resulting algorithm is shown in Figure 2.

3.3 Principle 2: The system word principle

From the examples in Section 1 it is evident that there exists an overlap of content/system morphemes duality with the traditional content/function words opposition defined in linguistics, although they are not equivalent and the traditional classifications also not strictly distinguishable. Therefore in the implementation of the 2 Principle (P2) for ML determination instead of content/system morpheme duality (Myers-Scotton, 2002) a content/function Part of Speech (POS) duality is considered.

System POS are identified, namely determiners, auxiliaries, subordinating conjunctions and coordinating conjunctions, while the rest of POS are considered as content POS. ML is determined in a CS utterance if one of the participating languages provided function POS for the utterance and the other language did not. The language that provided the function POS is determined as the ML. Although CS POS taggers exist (Feng et al., 2022; Bansal et al., 2022) none of them are available in open-source and since training a POS tagger is not a goal of this work a monolingual POS tagger is used instead. For example in the utterance "im okay with **the** 蛋黄" determiner "the" is used and therefore ML is determined as English.

4 Experiments

4.1 Datasets

The experiments using the algorithms described above are carried out using monolingual Callhome subsets and 2 CS datasets: SEAME (Lyu et al., 2010) and Miami subcorpus from the Miami-Bangor corpus[1].

4.1.1 Monolingual data

Monolingual LM training for P1.2 was carried out using Callhome datasets for English[2], Mandarin[3] and Spanish[4]. Pretrained LMs were not used in this work because they do not provide likelihood scoring of morpheme units. The summary of the datasets is presented in Table 2.

4.1.2 CS data

CS spoken language corpora SEAME (120 hours) and Miami (35 hours) are used for analysis and acoustic MLID training. Agreement analysis is carried out for CS utterances from the SEAME and Miami corpora and monolingual SEAME and Miami utterances are used for estimating the scaling factor α (Table 4). The monolingual subsets of SEAME and Miami are also used for training the mapping from the LID outputs to English, Mandarin and Spanish posteriors (LID_{map}). Mandarin characters in the SEAME corpus are word segmented which is helpful when applying P1.1, a principle that operates with words. All the introduced principles require morpheme-level LID tag annotation which is available for Miami and is automatically determined for SEAME based on the script (latin vs logographic). Finally, additional MLID-annotated 91 CS Miami utterances were used to measure the quality of MLID prediction from text and audio. The annotated MLID labels were assigned to the CS utterance transcriptions on the basis of determiner-noun-adjective complexes (Parafita Couto and Gullberg, 2017).

[1] https://biling.talkbank.org/access/Bangor/Miami.html
[2] LDC97T14
[3] LDC96T16
[4] LDC96T17

Table 3: Examples of applying the principles.

Utterance	baseline	ML P1.1	ML P1.2	ML P2
i thought all trains 都是via jurongeast 去到pasirris	en	en	en	en
but 他蛮zai 的right	en	zh	en	en
but 我的parents 都没有sponsor 我	zh	zh	zh	en
还有chicken noodles	en	en	en	zh

Table 4: CS dataset splits.

	Units	SEAME	Miami
Monolingual	utterances	53086	38401
CS raw		56951	2425
CS annotated		-	91

Table 5: Perplexities calculated for the validation and test subsets of monolingual Callhome data.

	English	Mandarin	Spanish
Valid	48.97	94.98	57.76
Test	57.61	98.16	52.30

4.2 Applying ML Principles to utterance transcriptions

P1.1 ML only applies to utterance transcriptions with singleton insertions, therefore resulting in a small data coverage: only for 36% (SEAME) and 60% (Miami) of the CS data the ML is determined. In P2 POS tags are computed for constituent monolingual islands (segments) of a CS utterance using a pretrained CNN-based POS tagger (Svenstrup et al., 2017). P2 covered 31% (SEAME) and 58% (Miami) of all of the CS examples. Furthermore, a baseline MLID determiner from text is implemented which is based on the token LID count following Bullock et al. (2018). Examples of running the resulting principles implementations is presented in the Table 3.

The implementation of P1.2 includes three components: a Machine Translation (MT) system, a pseudomorpheme tokeniser and a language model (LM). CS utterances are translated word by word using Wiktionary[5] to preserve the token order. The English and Spanish LMs are trained on the tokenised English and Spanish Callhome datasets. The tokenisation was carried out using a stemmer where stem and affix would be separated. For the Mandarin Callhome dataset separate characters are regarded as morphemes. The two Transformer-based (Vaswani et al., 2017) LMs with 2 layers, 2 attention heads per layer are trained for 25 epochs with negative log-likelihood loss on one 3080 Nvidia GPU for 1 hour. Validation and test perplexities for the three languages are presented in Table 5.

[5]https://www.wiktionary.org/

Moreover, a preliminary experiment is carried out to evaluate if the trained LMs have the ability to detect the original word order (WO) among its permuted variants (up to 20 word permutations). The sequence of tokens for which the probability was the highest was chosen as the predicted original WO. Comparing the sequence with chosen WO to the original WO leads to 37% accuracy for SEAME and 60% for Miami.

Table 7: Outcomes of α estimation. "-α MCC" is the correlation measured between the MLID determined by the unscaled P1.2 approach and MLID labels from other principles (+ true MLID labels for Miami). "+α MCC" are the correlation measurements with the scaled P1.2.

	SEAME		Miami		
	P1.1	P2	P1.1	P2	true
-α MCC	0.31	**0.33**	0.36	0.08	**0.41**
+α MCC	**0.36**	0.31	**0.38**	**0.09**	0.37

Outputs of the pre-trained monolingual LMs have different probability distributions, therefore, as described in Section 3, the factor α is used to allow for scale changes. α is derived from expectations of the probabilities yielded on monolingual examples and their translations following Equation 4. As a result of α estimation the MCC of SEAME P1.1/P1.2, Miami P1.1/P1.2 and Miami P2/P1.2 has increased (Table 7). Additionally, a "true" α value is calculated using ground truth MLID for Miami and P1.1 and P2 MLID for SEAME, and they are compared to the estimated α. DET plots and highlighted thresholds in Figure 1 demonstrate that by using the estimated α the amount of False Positives (FP) and False Negatives (FN) becomes more balanced for SEAME. For Miami the α estimation does not lead to more balanced FP and

Table 6: Experimental results for SEAME. First three columns and last three rows (P1.1, P1.2 and P2) refer to the ML determination principles from text. "Coverage" row presents the percentage of all CS examples being processed. "% English" row displays the percentage of utterances recognised as "English" LID or MLID. MCC Baseline refers to the word LID majority implementation (Bullock et al., 2018). "LID" is a pretrained LID system, "LID_{map}" column is a mapping trained on monolingual utterances from SEAME. $MLID_{P1.1}$, $MLID_{P1.2}$ and $MLID_{P2}$ are trained mappings similar to LID_{map} but trained on CS data and labels generated from transcriptions by corresponding principles. $MLID_{P1.1}$, $MLID_{P1.2}$ and $MLID_{P2}$ contain correlation values with the target MLID determined from text (*italic*) and correlations with other MLID targets.

	P1.1	P1.2	P2	LID	LID_{map}	$MLID_{P1.1}$	$MLID_{P1.2}$	$MLID_{P2}$
Coverage	36%	100%	31%	100%	100%	100%	100%	100%
% English	24%	46%	49%	18%	43%	42%	44%	45%
MCC Baseline	0.99	0.28	0.69	0.33	0.33	**0.5**	0.38	0.46
MCC P1.1	1	0.36	0.83	0.41	0.5	*0.67*	0.47	**0.52**
MCC P1.2	0.36	1	0.31	0.09	0.14	**0.17**	*0.3*	0.16
MCC P2	0.83	0.31	1	0.33	0.45	**0.49**	0.4	*0.6*

FN but this improvement is not observed due to the limited test set and other reasons which will be discussed later (Section 4.5.2).

4.3 Language Identification

If one assumes that is a "dominant" language that most acoustically resembles the spoken CS utterance, then a conventional LID system can be used as an ML determiner. An ECAPA-TDNN (Desplanques et al., 2020) model pretrained on Voxlingua-107 (Valk and Alumäe, 2020) was used to automatically detect the dominant language from audio data (Table 6 and 8, column LID). The ECAPA-TDNN model was trained to recognise a large number of languages. In order to limit the models to binary task a mapping function was trained from the outputs based on a fully-connected neural network (Multi-Layer Perceptron, MLP) classifier. The mapping function is trained to map 107 language output posteriors to the binary output of the languages participating in CS. LID is a challenging task for accented data such as monolingual subsets from SEAME and Miami but still achieves 82% and 79% F1-macro respectively on cross-validation among 5 splits.

4.4 ML identification from audio

One can train an MLP mapping model using the LID posterior distribution to also predict P1.1, P1.2 and P2 from audio. Due to the different coverage rates of P1.1, P1.2 and P2 of the CS data the amount of training data would vary greatly: 16582 for P1.1, 43068 for P1.2 and 23868 for P2. The resulting systems will be further referred to as $MLID_{P1.1}$, $MLID_{P1.2}$ and $MLID_{P2}$.

4.5 Correlation analysis

The agreement between the implemented principles is measured using the MCC metric since the MLID generated by the principles are not human annotation and are automatically generated. F1-macro is computed only in cases when the human-annotated Miami subset is compared to the MLID approaches.

4.5.1 Correlation between P1.1, P1.2 and P2

Table 9: Correlation values for SEAME with unlabelled sentences given a third "unknown MLID P1.1", "unknown MLID P1.2" and "unknown MLID P2" class labels for the three principle implementations accordingly. This approach ensures 100% coverage of all utterances sacrificing the MCC.

	P1.1	P1.2	P2
MCC P1.1	1	0.03	0.09
MCC P1.2	0.03	1	0.1
MCC P2	0.09	0.1	1

P1.1, P1.2 and P2 were applied to CS text data and the agreement analysis is presented in Table 6 and Table 8 for SEAME and Miami respectively in the first three columns. P1.1 and P2 have to meet certain conditions to be applied, therefore they do not have full coverage of CS data: 36% and 31% for SEAME, 60% and 58% for Miami. Measuring the correlations only for the utterances for which the MLID is determined is performed to measure the agreement of the implementations of the linguistic

Table 8: Experimental results for Miami. "LID_{map}" column is a mapping trained on monolingual utterances from Miami. "F1-macro true" and "MCC true" are the metric values when comparing the outputs of the systems to ground truth ML annotation for Miami.

	P1.1	P1.2	P2	LID	LID_{map}	$MLID_{P1.1}$	$MLID_{P1.2}$	$MLID_{P2}$
Coverage	60%	100%	58%	100%	100%	100%	100%	100%
% English	45%	31%	31%	43%	30%	31%	42%	31%
F1-macro true	100%	67%	93%	56%	53%	56%	**60%**	56%
MCC true	1.0	0.37	0.86	0.27	0.35	0.27	0.24	**0.38**
MCC Baseline	0.99	0.28	0.67	0.59	0.81	**0.83**	0.42	0.8
MCC P1.1	1	0.38	0.81	0.45	0.42	*0.85*	0.43	**0.82**
MCC P1.2	0.38	1	0.09	0.26	**0.34**	0.35	*0.53*	0.34
MCC P2	0.81	0.09	1	0.7	0.86	**0.87**	0.51	*0.82*

principles. Higher correlation indicates that the principles in agreement may be used to generate ground truth MLID labels for downstream tasks. Calculating MCC with unknown labels leads to extremely low correlation due to a large portion of unknown labels which makes it impossible to assess the slight changes in correlation of the labeled data (Tables 9 and 10).

Among the three principles P1.1 and P2 have the greatest correlation (0.82 for SEAME and 0.81 for Miami), P1.1/P1.2 demonstrates less correlation (0.36 and 0.38), while the least correlation is observed between P1.2 and P2 (0.31 and 0.09). P1.1 and the baseline have almost identical behavior which is expected (0.99 and 1.0), whereas less correlation is observed of the baseline with P2 (0.69 and 0.67) and P1.2 (0.28 and 0.28).

Table 10: Correlation values for Miami with unlabelled sentences given the "unknown ML" class.

	P1.1	P1.2	P2
F1-macro true	55%	67%	48%
MCC true	0.44	0.37	0.56
MCC P1.1	1	0.1	0.25
MCC P1.2	0.1	1	0.16
MCC P2	0.25	0.16	1

The high correlation values for P1.1 and P2 prove that the MLF framework can reliably predict the structure and behavior of CS text. This enables to use the MLIDs generated by the rule-based principles as pseudo-labels in applications.

4.5.2 Correlation of P1.1/P1.2/P2 and the acoustic LID/MLID

The ML determined from CS text is compared to the LID computed from the corresponding audio. The procedure for the LID experiments is described in the previous subsection. Columns 4 and 5 in Tables 6 and 8 show the amount of correlation between MLID derived from text and the recognised LID classes. The same columns for Miami in Table 8 also include F1 macro and MCC for an annotated MLID subset. Training LID_{map} on the monolingual utterances seems to increase the MCC (from 0.27 to 0.35) but decrease the F1 macro (56% from 53%) for the CS Miami data.

Suppose a conventional LID system determines the dominant language in a CS audio based on the majority of time the language is spoken. Then the true annotation may be approximated by counting the textual token LIDs in a CS utterance (Baseline). However, correlation analysis shows that $MLID$ systems are better predictors of the token LID majority (columns 4-5 vs 6-8 row MCC Baseline in Tables 6 and 8).

Further experimentation comprises of comparing $MLID_{P1.1}$, $MLID_{P1.2}$ and $MLID_{P2}$ with P1.1, P1.2 and P2. Upon observing the results for SEAME data $MLID_{P1.2}$ leads to overall highest value out-of-domain MCC scores (0.47+0.4=0.87) for textual principles P1.1 and P2. A similar inspection of the Miami results shows the biggest MCC scores for $MLID_{P1.1}$ (0.35+0.87). For the annotated subset of Miami data $MLID_{P1.2}$ leads to the biggest F1 macro among all systems (60%), while $MLID_{P2}$ leads to the biggest MCC score (0.38).

Table 11: Distributions of languages in CS corpora. Utterance level LID for monolingual subsets is in the "Utterance LID" row, token level LID for CS is in the "Token LID" row and utterance level textual ML for CS are in rows P1.1/P1.2/P2.

	SEAME		Miami	
	English	Mandarin	English	Spanish
Utterance LID (mono)	54%	46%	68%	32%
Token LID (CS)	42%	58%	66%	34%
P1.1 (CS)	23%	77%	45%	55%
P1.2 (CS)	44%	56%	31%	69%
P2 (CS)	49%	51%	31%	69%

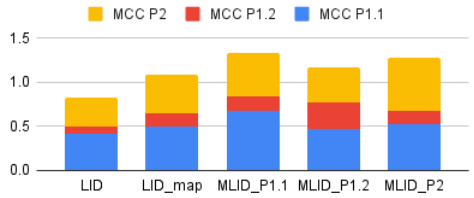

Figure 3: Correlations between acoustic *LID* and *MLID* outputs and textual P1.1, P1.2 and P2 for CS SEAME data. Each bar segment represents the amount of correlation for a LID or MLID model with textual principles, therefore the whole bar represents the sum of the correlations.

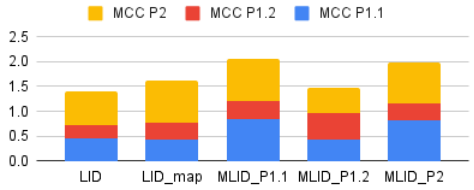

Figure 4: Correlations between acoustic *LID* and *MLID* outputs and textual P1.1, P1.2 and P2 for CS Miami data.

Lastly, the MCC scores between the textual and acoustic MLID determiners are summed up for every *LID*/*MLID* approach (Figures 3 and 4). Results show that the correlation of the proposed approaches with the MLID is higher than in *LID* systems in all occasions apart from $MLID_{P1.2}$ for CS Miami data. The latter is due to the word order of the English and Spanish languages being similar in contrast to English/Mandarin CS. This leads to the morpheme order having a better discrimination power in the case of English/Mandarin CS than in English/Spanish CS.

4.5.3 P1.1/P1.2/P2 distribution analysis

At the last step of analysis the distributions of languages are measured on utterance level LID for monolingual (Utterance LID row in Table 11), token level LID for CS (Token LID row in Table 11) and utterance level textual MLID for CS (P1.1/P1.2/P2 rows in Table 11). The numbers reveal that although the majority of the monolingual utterances are English in both corpora (54% for SEAME and 68% for Miami), it is not the preferred ML when CS occurs in the utterance for all principles. The token LID distribution also does not seem to be correlated with the choice of the ML in these corpora. In SEAME there seems to be a strong preference towards using Mandarin as an ML (77%) when EL insertions are single words (P1.1). The preference is not as strong for Spanish in the CS Miami subset (55%) but it is still a big difference in comparison to the monolingual distributions (32%). P1.2 and P2 show a similar distribution of MLIDs with the numbers indicating the preference of speakers to use the non-English language as the grammatical frame for a CS utterance.

5 Conclusion

To the best of our knowledge this is the first work that precisely carries out the Matrix Language (ML) determination of a code-switched (CS) utterance based on the Matrix Language Frame (MLF) theory and that compares Matrix Language Identity (MLID) to acoustic Language Identity (LID). Three methods for ML determination in text and audio are implemented using the ideas and the concepts of the MLF theory (Myers-Scotton, 1997). An extensive correlation analysis of the MLID systems from text and speech is carried out. A pretrained LID system *LID* is adapted to the data by training a mapping function LID_{map}, while also mapping functions $MLID_{P1.1}$, $MLID_{P1.2}$, $MLID_{P2}$ for MLID are trained. *MLID* consistently outperforms *LID* for ML determination from audio based

on Matthew's Correlation Coefficient (MCC). Comparing the results to the ground truth ML annotation shows that the trained $MLID_{P1.2}$ and $MLID_{P2}$ outperform LID in terms of F1-macro and MCC respectively. Finally, this approach reveals that despite English dominating as the utterance LID for the monolingual utterances, non-English (Mandarin or Spanish) languages set the grammatical frame for CS utterances.

The proposed approaches can be used for accurate automatic analysis of CS text and audio. It can provide insight into the nature of CS for whole datasets but also separate speakers and even utterances. Further work will explore the usefulness of the MLID implementations in Natural Language Processing and Automatic Speech Recognition (ASR) applications, namely in language and dialogue modelling and also in end-to-end multitask ASR the MLID component will be used as a part of the ASR setup. Additionally, further development of P2 is required where the system morphemes would be automatically determined from a given set of CS data rather than using a closed set of POS tags.

Limitations

The main limitation of the method is related to data availability: there is limited ML-annotated CS data openly available to date. Therefore it is problematic to assess the quality of ML classification. ML identity can be determined in CS data using the P1.1 but the principle can only be applied in case of singleton EL insertions. Since there is no ML annotation, correlation was measured for most of the experiments which is difficult to assess. Finally, although providing valuable insight into the CS data, the usefulness of the method is yet to be tested in NLP and ASR applications.

Acknowledgements

Many thanks to Maria del Carmen Parafita Couto and Diana Carter for sharing the ML annotation for the Miami-Bangor corpus. We would also like to thank Tom Pickard for the very helpful comments on the paper. This work was supported by Engineering and Physical Sciences Research Council [grant number 2676033].

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.

Srijan Bansal, Suraj Tripathi, Sumit Agarwal, Teruko Mitamura, and Eric Nyberg. 2022. PRO-CS : An instance-based prompt composition technique for code-switched tasks. In *EMPNLP 2022*, pages 10243–10255, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gayatri Bhat, Monojit Choudhury, and Kalika Bali. 2016. Grammatical constraints on intra-sentential code-switching: From theories to working models. *arXiv preprint arXiv:1612.04538*.

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*.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*. ISCA.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, et al. 2021. Multilingual and code-switching asr challenges for low resource indian languages. *arXiv preprint arXiv:2104.00235*.

Yukun Feng, Feng Li, and Philipp Koehn. 2022. Toward the limitation of code-switching in cross-lingual transfer. In *EMNLP 2022*, pages 5966–5971, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gualberto A. Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *INTERSPEECH*.

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

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*.

Metilda Sagaya Mary N J, Vishwas M. Shetty, and S. Umesh. 2020. Investigation of methods to improve the recognition performance of tamil-english code-switched data in transformer framework. *ICASSP 2020*, pages 7889–7893.

H. Muralikrishna, S. Kapoor, Dileep Aroor Dinesh, and Padmanabhan Rajan. 2021. Spoken language identification in unseen target domain using within-sample similarity loss. *ICASSP 2021*, pages 7223–7227.

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.

Maria Carmen Parafita Couto and Marianne Gullberg. 2017. Code-switching within the noun phrase: Evidence from three corpora. *International Journal of Bilingualism*, 23.

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

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

Dan Svenstrup, Jonas Meinertz Hansen, and Ole Winther. 2017. Hash embeddings for efficient word representations. *Preprint*, arXiv:1709.03933.

Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2018. Multilingual speech recognition with a single end-to-end model. *Preprint*, arXiv:1711.01694.

Jörgen Valk and Tanel Alumäe. 2020. Voxlingua107: a dataset for spoken language recognition. *Preprint*, arXiv:2011.12998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Emre Yilmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, and Thomas R. Niesler. 2018. Building a unified code-switching asr system for south african languages. *ArXiv*, abs/1807.10949.