Code-switching in Context: Investigating the Role of Discourse Topic in Bilingual Speech Production

Debasmita Bhattacharya and Anxin Yi and Siying Ding and Julia Hirschberg
Department of Computer Science
Columbia University
New York, NY, USA
debasmita.b@cs.columbia.edu, ay2616@columbia.edu, sd3609@barnard.edu, julia@cs.columbia.edu

Abstract

Code-switching (CSW) in speech is motivated by conversational factors across levels of linguistic analysis. While we know much about why speakers code-switch, there remains great scope for exploring how CSW occurs in speech, particularly within the discourse-level linguistic context. We build on prior work by asking: how are patterns of CSW influenced by different conversational contexts spanning *Academic*, *Cultural*, *Personal*, and *Professional* discourse topics? To answer this, we annotate a Mandarin-English spontaneous speech corpus, and analyze its discourse topics alongside various aspects of CSW production. We show that discourse topics interact significantly with utterance-level CSW, resulting in distinctive patterns of CSW presence, richness, language direction, and syntax that are uniquely associated with different contexts. Our work is the first to take such a context-sensitive approach to studying CSW, contributing to a broader understanding of the discourse topics that motivate speakers to code-switch in diverse ways.

1 Introduction

Code-switching (CSW) occurs when a multilingual speaker alternates between languages in speech or writing (Poplack, 1980). Speakers can code-switch between or within utterances across a variety of language pairs, producing a) syntactically simple *insertional* code-switches of single words or short phrases, or b) more syntactically complex *alternational* code-switches at grammatical clause boundaries (Muysken, 2000),[1] e.g.:

(1) a. "让 我 拿出 我 的 calculator."
["Let me get out my calculator."]
b. "我 不 懂 but the result isn't out yet."
["I don't understand but the result isn't out yet."]

[1]Insertional and alternational code-switches are known as different *strategies* of code-switching.

Prior work has examined why speakers code-switch, showing the influence of various conversational factors: speaker competency, linguistic context, affective state of the speaker, the type of information the speaker wants to convey, and listener identity, among many others (Dornic, 1978; Bell, 1984; Gardner-Chloros, 2009; Broersma, 2009; Ferreira, 2017; Bhattacharya et al., 2024b). While much work has focused on the psycho-, socio-, and paralinguistic motivations for CSW, some studies have proposed alternative explanations of CSW based on discourse-level analysis. Early discourse-functional work on code-switched speech, e.g. Blom and Gumperz (1972); Auer (1998), suggested that CSW indicates a shift in topic during spontaneous conversations. This claim has held true in more recent studies across speech settings and language pairs (see Section 2). However, little is known about the *types* of topics that tend to elicit CSW, or how different genres of topic motivate distinctive patterns of downstream code-switched language production, particularly from a quantitative perspective across large-scale datasets of conversational speech. So, while we know much about *why* speakers code-switch in speech, there remains great scope for exploring *how* CSW occurs, especially within the discourse-level linguistic context.

We begin this research by studying the extent to which the topic of bilingual Mandarin-English conversations interacts with the presence, quantity, frequency, language direction, and syntactic complexity of CSW in spontaneous speech. We do so by examining an augmented version of the SEAME corpus of code-switched speech (Lyu et al., 2010) using statistical and unsupervised learning approaches, finding not only that differences in discourse topics interact significantly with CSW, but also that these interactions result in distinctive patterns of CSW features that can be used to distinguish between conversational contexts.

Our contributions include 1) producing a man-

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 64–80

ually annotated version of the SEAME corpus for new aspects of CSW, which we share at https://tinyurl.com/3ac6jv2b; 2) building a topic classifier for automatic annotation that is robust to both monolingual and multilingual Mandarin and English speech; and 3) identifying novel and nuanced quantitative and qualitative insight into the influence of discourse on multiple aspects of CSW. Overall, we contribute to a broader understanding of the conversational contexts that motivate speakers to code-switch in diverse ways. We hope that this work will inform innovation in robust spoken language technology that is capable of both understanding and producing code-switched speech grounded in naturalistic aspects of multilingual discourse function.

2 Prior Work

The earliest work on discourse aspects of CSW focused on defining taxonomies of when and why speakers code-switch. Notably, Blom and Gumperz (1972) proposed a dichotomy between *situational* CSW that indicates the topic of conversation, and *metaphorical* CSW for signaling emphasis; the combination of these allowed for the prediction of language choice among bilingual speakers in Norway. Though the precise boundary between situational and metaphorical CSW has been the subject of debate among authors such as Auer and Wei, subsequent work has supported the claim that CSW serves as a discourse context cue that signals a semantic shift in the topic of Italian-German and Cantonese-English dialogue (Auer, 1998; Wei, 1998; Auer, 2003). Ethnographic studies of Spanish-English, e.g. Lowi (2005), have similarly shown that both intra- and inter-sentential CSW is used as a discourse feature to indicate change of topic among adult bilinguals of varying linguistic ability. These results generalize to the speech of children, in which topic and situation shift signaling is found to be a primary function of Spanish-English CSW (Reyes, 2004). Such a relationship between discourse framing and CSW has been observed in other language pairs, including Malaysian-English (Ariffin and Rafik-Galea, 2009), Bangla-English (Das, 2012), and Hindi-English (Dey and Fung, 2014; Begum et al., 2016), and across spoken and written modalities.

By primarily conducting qualitative examinations of small-scale and hand-curated speech corpora, and analyzing coarse-grained CSW characteristics (e.g. the number of code-switches in a given dialogue), the existing work described above has consistently established a link between discourse framing and CSW. However, few studies have extended this to consider finer-grained aspects of code-switched speech in relation to the nature of specific discourse topics. This is especially striking for the Mandarin-English language pair, given the vast number of global Mandarin native speakers,[2] many of whom are bilingual in English and code-switch regularly. Prior work on the high-level processing of topics in code-switched social media text and speech (Peng et al., 2014; Asnani and Pawar, 2016; Rabinovich et al., 2019) has targeted making topic modeling techniques robust to multilingual inputs, rather than identifying a deeper understanding of the types of topics that are associated with specific CSW behaviors. To address this gap, we ask **RQ:** How are specific discourse topics associated with patterns in the presence, quantity, frequency, direction, and syntactic complexity of CSW in a conversational domain?

3 Corpus

We examine the Mandarin-English Code-switching in South-East Asia (SEAME) corpus of spontaneous speech (Lyu et al., 2010). This corpus is made available by the LDC User Agreement for Non-Members. SEAME consists of 192 hours of speech and 1,074,032 transcribed words across 256 dialogues. 156 unique speakers from Singapore and Malaysia are represented in the corpus. All dialogues are in an informal register, whether they were recorded in open-domain *conversation* settings or slightly more structured *interview* settings. Recordings comprise a mix of monolingual and code-switched utterances, the latter of which are Mandarin-dominant with inter-sentential code-switches to English. The corpus-level token ratio of Mandarin to English is 1.54:1.

4 Method

Data annotation and pre-processing: aspects of CSW. We annotate SEAME for the different aspects of CSW performed by speakers. First, we inspect the 110K utterance-level transcripts and automatically label each one for whether it is code-switched or monolingual, based on the Simplified Chinese and English orthographies used in the corpus. For the code-switched utterances, we also

[2] Almost 1 billion, per Ethnologue as of early 2025.

calculate utterance-level CSW quantity using the CSW ratio and M-index metrics (Soto et al., 2018; Barnett et al., 2000) and CSW frequency using the I-index metric (Guzman et al., 2016). We provide complete definitions of these metrics in Appendix A. We then perform additional manual annotations on the code-switched utterances to distinguish between insertional (I), alternational (A), and "other" (O) forms of CSW. We define "other" CSW similarly to tag-switching (Poplack, 1980), as the strategy used in any utterance where the code-switch is a filler word at the outset or end of a sentence. All annotation is performed by the second and third authors, who are native speakers of Mandarin with first-language proficiency in English. When the annotators disagree on a label, which occurs in less than 1% of utterances, they discuss their reasoning with each other until the disagreement is resolved.

We find that 48% of utterances in the corpus are monolingual. Among the 52% of code-switched corpus utterances, 89% use insertional CSW, 12% use alternational CSW, and 7% use "other"[3] CSW.

Classifier construction: discourse topic labeling. Given the dataset's size, instead of performing a second round of manual annotation of discourse topics over the entire corpus, we use a multi-class classifier to approximate utterance-level ground truth labels of discourse topics. Rather than unsupervised topic modeling, we use classification to approximate ground truth labels since we have some prior knowledge of the discourse topics present in the corpus. To do so, we train[4] and evaluate four classifier models on a 10% sample of the corpus (11K utterances; the **ground truth** set), which we manually annotate for topic using the same protocol for resolving disagreement as above; label disagreements occur in less than 5% of utterances. Full task instructions are in Appendix A. We then apply the best-performing classifier for inference on the remainder of the corpus.

We begin with a rule-based approach and define a set of seed words in English and Mandarin as the lexicon associated with each of the following broad topic areas: *Academic*, *Cultural*, *Personal*, *Professional*.[5] We choose this particular set of topics based on those that were used to elicit speech during the collection of the corpus by Lyu et al., and our own observations of the data during our initial annotation pass, as these are most likely to reliably reflect the topics actually present in the data, and to avoid unnecessary complexity for an already time-intensive manual annotation task. We further justify this choice of topics experimentally in Appendix A. We define an additional *Other* topic to account for utterances that do not fall into any of the above topic areas. At inference time, we assign topic labels by identifying the number of exact matches with each topic's lexicon and breaking ties at random. Any utterance with zero matches is labeled as *Other*. We assess the performance of this initial classifier on both the entire ground truth set and a test-time subset of it, for consistency with subsequent models.

We then refine our rule-based approach by expanding our handcrafted lexica with lexical and conceptual synonyms, applying Havaldar et al. (2024)'s method. This involves choosing the ten most similar neighbors per seed word using a cosine similarity threshold greater than or equal to 0.9 on pre-trained GloVe embeddings (Pennington et al., 2014).[6] We incorporate these synonyms into our existing lexica, then perform inference with the expanded classifier and evaluate performance on the ground truth set and its test-time subset.

Next, we use `scikit-learn 1.6.1` to train a stacking ensemble (291K params.) combining three individually calibrated base learners: logistic regression, random forest, and gradient boosting. The ensemble takes as input utterance-level speaker gender, dialogue type, token length, presence of filler words (see list in Appendix A), presence of corpus-level frequent words, unigram tf-idf statistics, speaking rate, duration, and pause rate, in addition to weighted counts of seed words from the discourse topic lexica. We train this classifier using a logistic regression meta-learner on 80% (8.8K utterances) of the ground truth set, blending predictions with 3-fold cross-validation, and evaluate its performance on the remaining 20% of the ground truth set (2.2K utterances; the **held-out test** set).

Finally, we use a self-supervised learning approach, with a class-weighted logistic regression base model (17K params.). The input features to this model are the same as those used by the ensemble classifier, with additional bigram tf-idf statistics

[3] We retain all annotations of this CSW strategy in our augmented version of the SEAME corpus, but largely exclude this strategy from our subsequent statistical analyses for simplicity.

[4] Models are trained in about an hour on a Mac M1 chip.

[5] For detailed definitions and examples of each, as well as a complete list of seed words in each lexicon, see Appendix A.

[6] `glove-wiki-gigaword-50` accessed via Gensim.

and pre-trained sentence embeddings sourced from HuggingFace[7] (33M frozen params.). This model iteratively generates pseudo-labels for unlabeled utterances in each of up to five rounds. For each discourse topic, we select up to a fixed quota of the highest-confidence predictions, using specific thresholds tuned per class that particularly include a dynamic adjustment for the relatively sparsely represented *Cultural* class. We treat 70% (7.7K utterances) of the ground truth set as the pseudo-train set and append newly pseudo-labeled examples to it before retraining and recalibrating the model. We monitor performance on a separate 10% subset of the ground truth set (1.1K utterances), using macro F1 to determine early stopping, and evaluate final model performance on the same held-out test set as above.[8]

Statistical analysis. Once the corpus is labeled for all CSW features of interest and discourse topics, we examine the relationship between these by using chi-squared and one-way ANOVA tests.

Clustering analysis. We build on significant statistical results by using `scikit-learn 1.6.1` to perform k-means clustering (10 params.) on vectors representing utterance-level binary CSW presence, strategy, and language direction, and CSW quantity and frequency, standardized to zero mean and unit variance. We then examine the resulting clusters and compare their composition over discourse topics and each CSW feature of interest.

5 Results

5.1 ML models outperform rule-based classifiers on discourse topic labeling.

We first calculate the expected blind guessing, i.e. random, baseline accuracy on our data, given the distribution of discourse topic labels in the held-out test set: 0.33. We subsequently use this value to contextualize the performance of our models.

We find that all four of our models significantly outperform the calculated baseline over the held-out test set (Table 1).[9] As expected, the performance of our rule-based classifiers is generally inferior to that of the machine learning models, given that certain characteristics of the discourse topic cannot be captured by the raw content of an utterance alone. Somewhat unexpectedly, the expanded lexicon-based classifier performs worse than the initial rule-based one, indicating that certain synonyms effectively *dilute* associations with specific discourse topics. On the other hand, the relative performance of the two machine learning models aligns with our expectations, as the self-supervised classifier demonstrates its unique ability to leverage large quantities of unlabeled data during learning. However, despite both machine learning models' relative superior performance, we note the overall difficulty of utterance-level discourse topic labeling, reflected by the modest absolute value of all four classifiers' task accuracy.[10]

Classifier	Accuracy	F1 Score
Initial lexicon-based	0.62	0.60
Expanded lexicon-based	0.60	0.59
Ensemble: LR, RF, GB	0.68	0.62
Self-supervised LR	0.72	0.71

Table 1: Classifiers' accuracy and macro F1 score on discourse topic labeling over the held-out test set. We also report per-class performance metrics for the best-performing model in Table 14 in Appendix A.11.

Following training and evaluating the four classifiers on our sample of ground truth data, we use the self-supervised model to infer discourse topics for the remaining 90% of the corpus (99K utterances) that consists of unlabeled utterances. This results in the utterance-level distribution of discourse topics shown in Table 2, which suggests that certain conversational contexts are more popular than others.

Discourse topic	% of corpus
Academic	7.8
Cultural	0.1
Personal	28.1
Professional	2.4
Other	61.5

Table 2: Discourse topic label distribution across classes in the entire SEAME corpus. Please see Table 15 in Appendix A.12 for distributions across the ground truth and automatically-annotated subsets of the corpus.

We note that utterances on *Other* topics dominate the corpus, aligning with expectations for open-domain dialogue, and supporting our definition of this category to account for most utterances. To verify the absence of hidden clusters of topics within the *Other* category, we use exploratory LDA

[7] `sentence-transformers/all-MiniLM-L6-v2`

[8] Hyperparameter values for both the ensemble and self-supervised models are in Appendix A.

[9] For the rule-based classifiers' performance over the entire ground truth set, see Appendix A.

[10] See ablation studies in Appendix A for relative contributions of different features to topic classification performance.

and BERTopic models (Blei et al., 2003; Grootendorst, 2022).[11] Both show that *Other* utterances are typically too short[12] to clearly denote any topic, and primarily consist of function words like wh-question words and deictic pronouns, fillers (e.g. "um", "loh"), and temporal or connective markers (e.g. "then", "如果", "after"). While some functional groupings are present, we conclude that further subdividing the *Other* topic is not justified, as no additional coherent *semantic topics* emerge from this analysis.

The *Personal* topic is the next-most highly represented at the corpus-level, accounting for close to a third of utterances. This reflects how everyday conversations often revolve around personal anecdotes, thoughts, opinions, and feelings, validating our choice to examine it as a core discourse topic. The *Academic* and *Professional* topics are also present in the corpus, though their representation is relatively modest. And, while the *Cultural* topic is rare within the SEAME corpus, note that it still accounts for hundreds of unique utterances.

5.2 Discourse topics interact significantly with CSW presence, strategy, direction, quantity, and frequency.

Having a fully labeled corpus, we begin our statistical study of how discourse topics interact with aspects of CSW production in SEAME by considering differences in topic between monolingual and code-switched utterances. We control for utterances that are greater than 6 tokens in length across topics, since utterances in the *Other* topic are notably shorter than those in the other four topics.

Chi-squared tests comparing monolingual and code-switched utterances by topic all yield significant results, with clear patterns in associated odds ratios (Table 3). The odds that an utterance is about *Academic*, *Cultural*, *Personal*, or *Professional* topics, given it is code-switched, are at least two times those for a monolingual utterance. It seems that certain discourse contexts significantly lend themselves to multilingual, rather than monolingual, production, which is particularly noteworthy given their collective minority representation in the corpus overall. In contrast, the odds that an utterance is about any *Other* topic, given it is code-switched, are only about two-thirds of those for a monolingual utterance, suggesting that *Other* topics are much better expressed in a monolingual fashion.

Topic	χ^2	*p*-val.	OR	95% CI
Academic	443.3	**	1.92	[1.81, 2.05]
Cultural	13.7	**	2.48	[1.52, 4.27]
Personal	430.2	**	2.49	[2.42, 2.56]
Professional	357.4	**	3.16	[2.78, 3.59]
Other	448.1	**	0.68	[0.66, 0.70]

Table 3: Chi-squared tests and odds ratios comparing topics in monolingual and code-switched utterances. Odds ratios >1 favor CSW. Odds ratios <1 favor monolinguality. p-values less than 0.01 are denoted by **.

Honing in specifically on CSW, utterances on *Academic*, *Professional*, or *Other* topics are more likely to be insertionally code-switched than alternationally code-switched (Table 4), reflecting how specific discourse contexts intersect with the strategy of CSW that is most represented in the corpus. Each of these topics also has greater representation of insertional CSW than the corpus overall (94% on average, compared to corpus-level 89%), reinforcing the influence of topic on CSW strategy. Interestingly, in *Cultural* and *Personal* topics, insertional CSW is equally and one-third as likely as alternational CSW, respectively, indicating that the relatively more complex CSW strategy is more suited to conversation topics that may require less subject knowledge to discuss, while the opposite was true for topics that may be more difficult to speak on. These patterns suggest that speakers might attempt to achieve a balance in complexity between the nature of the discourse topic under discussion and the CSW strategy used to express it when producing CSW. Given that both topics' representation of alternational CSW is the same as the corpus overall (about 12% in all cases), these topics' relative skew towards alternational CSW in our calculated odds ratios is even more striking.

Topic	χ^2	*p*-val.	OR	95% CI
Academic	28.5	**	1.25	[1.15, 1.36]
Cultural	435.0	*	1.03	[0.61, 1.86]
Personal	2029.2	**	0.31	[0.30, 0.33]
Professional	54.7	**	1.79	[1.53, 2.11]
Other	1430.0	**	2.98	[2.81, 3.16]

Table 4: Chi-squared tests and odds ratios comparing topics in insertional and alternational CSW. Odds ratios >1 favor insertional CSW. Odds ratios <1 favor alternational CSW. p-values less than 0.05 and 0.01 are denoted by * and **, respectively.

[11] Hyperparameter details are in Appendix A.

[12] Mean and standard deviation token length for *Other* utterances are 6.5 and 6.2, respectively. Across all other utterances, these are 15.1 and 9.3, respectively.

With respect to language direction of CSW, *Academic* and *Professional* utterances are about two-thirds as likely to be code-switched from English to Mandarin, as from Mandarin to English, reflecting the frequency of insertion of English technical, jargon-like, and/or domain-specific terms into such multilingual utterances (Table 5). *Personal* and *Cultural* utterances have relatively higher odds of being code-switched from English to Mandarin, and are about equally likely to be code-switched from Mandarin to English. Only utterances on *Other* topics are significantly more likely to be code-switched from English to Mandarin at an odds ratio of 1.22, which is especially worth noting given the overall skew of the corpus towards Mandarin.

Topic	χ^2	*p*-val.	OR	95% CI
Academic	78.0	**	0.75	[0.70, 0.80]
Cultural	0.007	–	0.96	[0.61, 1.47]
Personal	0.5	–	0.99	[0.95, 1.03]
Professional	91.7	**	0.57	[0.51, 0.64]
Other	102.6	**	1.22	[1.17, 1.27]

Table 5: Chi-squared tests and odds ratios comparing topics in English-to-Mandarin (en → zh) and Mandarin-to-English (zh → en) CSW. Odds ratios >1 favor en → zh. Odds ratios <1 favor zh → en. p-values less than 0.01 are denoted by **. p-values more than 0.05 are denoted by –.

Next, we perform one-way ANOVA tests to compare CSW quantity and frequency metrics across the different discourse topics. For each of CSW ratio, M-index, and I-index, ANOVA tests show a strong and statistically significant ($p < 0.01$) association between each topic and the metric of CSW richness. This statistical significance holds even after applying Bonferroni correction to account for possible noise in discourse topic labels generated by our best-performing classifier from Section 5.1. These associations suggest that variations in discourse topic can distinguish CSW behavior in terms of both quantity and frequency of utterance-level CSW. That is, there are significant differences in CSW richness in utterances on different topics. More concretely, we find that the *Personal* and *Cultural* topics consistently rank the lowest in terms of mean CSW richness across metrics, while the *Professional* and *Academic* topics are the two most highly ranked across the board. The *Other* topic sits in the middle of the ranking in each case. These results provide further evidence of a relationship between discourse-level conversational context and various aspects of CSW behavior in SEAME.

Overall, the results of our statistical analysis reveal significant interactions between specific discourse topics and granular patterns of code-switched speech production in SEAME. Not only are utterances on *Academic*, *Cultural*, *Personal*, and *Professional* topics more likely to be expressed using CSW, but each of these topics also has a unique, typical CSW profile. Multilingual utterances on *Academic* and *Professional* topics are characterized by higher quantity and frequency of CSW, with the majority of such code-switches taking place from Mandarin to English in an insertional fashion. In contrast, *Personal* and *Cultural* utterances are characterized by fewer and less frequent alternational code-switches from Mandarin to English. Utterances on *Other* topics are overall less likely to be code-switched; when such utterances are expressed multilingually, these are less striking in their CSW quantity or frequency, but are more likely to involve insertional code-switches from English to Mandarin. These findings provide the motivation for the remainder of the work.

5.3 Unsupervised models learn many discourse-CSW relationships.

Having found multiple significant interactions between discourse topics and several fine-grained aspects of CSW behavior in SEAME, we further develop our investigation by assessing whether these relationships are salient enough to be learned by unsupervised models, and potentially in turn inform the downstream outputs of such models. Instead of methods like LDA that explicitly group datapoints by topic, we want to see if unsupervised models that do not have this specific topic-centric objective can still cluster utterances based on both topic and CSW information, as a stronger test for the validity of associated patterns. To do so, we implement k-means clustering, setting $k = 5$ to match the number of distinct discourse topic labels, with random starting points[13] and principal component analysis. We then compare resulting cluster compositions across discourse topic and CSW presence, strategy, language direction, and richness, verifying the significance of these groupings using chi-squared tests. Throughout this section, we discuss only the comparisons yielding significant p-values.

We begin by comparing cluster compositions across topics and CSW presence, and find a clear separation between Cluster 2 and the remaining

[13] We motivate this design choice further in Appendix A.

Topic	C1	C2	C3	C4	C5
Acad.	6.3%	3.6%	**13.2%**	8.6%	10.0%
Cult.	0.2%	0.1%	0.2%	0.1%	0.9%
Pers.	29.8%	18.8%	**39.0%**	30.1%	**62.2%**
Prof.	1.2%	0.7%	4.8%	2.1%	2.7%
Other	62.5%	**76.7%**	42.7%	59.1%	24.3%

Table 6: Cluster composition by discourse topic.

CSW?	C1	C2	C3	C4	C5
No	0%	**99.9%**	0%	0%	0%
Yes	100%	0.1%	100%	100%	100%

Table 7: Cluster composition by CSW presence.

four clusters in terms of multilinguality of constituent utterances; this cluster is almost entirely dominated by non-code-switched utterances (Table 7), while also representing *Other* topics most highly (Table 6), in a clear reflection of the specific association between discourse and monolingual expression we have found in Section 5.2. Clusters 1, 3, 4, and 5 are all dominated by CSW, and each represents a mix of discourse topics. Cluster 5 is most representative of *Personal* utterances while Cluster 3 contains a combined majority mix of *Academic* and *Personal* topics. Although these patterns do not exactly align with our initial hopes of obtaining five distinct clusters, each of which is uniquely dominated by one of the discourse topics, these are still interesting as they mirror many of our earlier statistical findings. We hypothesize that the absence of a clear *Professional* or *Cultural* cluster may be due to the relatively lower representation of these discourse topics in the corpus overall (Table 2).

Strategy	C1	C2	C3	C4	C5
I	58.7%	0.1%	**95.8%**	82.0%	**27.0%**
A	12.5%	0.0%	1.6%	10.5%	**3.6%**
O	28.8%	0.0%	2.6%	7.4%	69.4%
None	0.0%	99.9%	0.0%	0.0%	0.0%

Table 8: Cluster composition by CSW strategy.

Considering cluster compositions in Table 8, we again find patterns of overlap between utterance-level CSW strategies and discourse topics that align with those found in Section 5.2. For instance, Cluster 3, which we have already noted for its representation of *Academic* utterances, while simultaneously representing the greatest proportion of *Professional* utterances relative to other clusters, also contains the greatest proportion of insertional CSW. This reinforces the strength of the interaction between discourse and CSW strategy for these topics. Similarly, Cluster 5, which is dominated by the *Personal* topic and contains the greatest proportion of the *Cultural* topic relative to other clusters, has the smallest gap in representation between insertional and alternational CSW. This aligns with our statistically significant observation that these topics are less associated with insertional than alternational CSW. However, we also note the overall lower proportion of alternational CSW in each cluster, and hypothesize that this may be due to the relative infrequency of this CSW strategy in the corpus compared to insertional CSW, as noted in Section 4.

Next, we compare cluster compositions across CSW language direction (Table 9) and metrics of CSW quantity and frequency (Table 10). In the case of the latter, we transform utterance-level CSW ratio, M-index, and I-index into binary variables by denoting values less than the median of each metric of CSW richness as "low" and values greater than or equal to the median as "high".

CSW dir.	C1	C2	C3	C4	C5
en → zh	**100%**	0.1%	15.6%	0%	**64.9%**
zh → en	0%	0.0%	**84.4%**	100%	35.1%

Table 9: Cluster composition by CSW language direction: English-to-Mandarin or Mandarin-to-English.

With respect to CSW language direction, Cluster 1, which primarily consists of utterances on *Other* topics, is made up exclusively of code-switches from English to Mandarin. This is striking as we know the *Other* topic is the only one that is significantly more likely to be expressed in such a fashion. Cluster 3, whose combined majority topic representation is from the *Personal* and *Academic* topics, is dominated by Mandarin to English CSW, which also aligns with our previous finding, since the *Academic* topic in particular is more likely to be code-switched in this direction. Cluster 5's English-to-Mandarin dominance is also interesting, and is likely due to the presence of the *Personal* and *Other* topics, the former of which is equally likely to be code-switched in either direction, and the latter of which is always more likely to be code-switched from English to Mandarin; their combination likely determines the overall cluster composition in terms of CSW language direction. Recall that Cluster 2 effectively contains no CSW (Table 7) and hence does not contain CSW in either direction.

Metric	C1	C2	C3	C4	C5
R:H	90.1%	0%	**98.9%**	**90.0%**	0%
R:L	9.9%	100%	1.1%	10.0%	**100%**
M:H	95.9%	0%	**98.1%**	**87.5%**	0%
M:L	4.1%	100%	1.9%	12.5%	**100%**
I:H	90.1%	0%	**98.7%**	**90.0%**	0%
I:L	9.9%	100%	1.2%	10.0%	**100%**

Table 10: Cluster composition by metrics of CSW richness: CSW ratio (R), M-index (M), and I-index (I), binned into high (H) and low (L) values.

Finally, we examine cluster composition across metrics of CSW quantity and frequency. We find that the distribution of high vs. low values of each metric in Cluster 5 supports our previous finding that *Personal* and *Cultural* topics always contain the lowest quantity and frequency of CSW. Similarly, Cluster 3 reinforces how the *Academic* topic always has the highest values across metrics. The composition of Cluster 4 also demonstrates how the *Other* and *Academic* topics, which we know are associated with mid to high levels of CSW richness, pull metric values up within the cluster.

Overall, our clustering model is able to group utterances according to both topic and CSW characteristics, which indicates that it can learn relationships between topics and CSW patterns in a reasonable way. These results demonstrate that many of the statistical relationships we have found between discourse topics and various fine-grained aspects of CSW behavior in SEAME are significant enough to be learned by unsupervised models, and may well inform their downstream outputs, though we leave a detailed investigation of the latter claim to future work. A random baseline analysis confirms this conclusion and validates that our current clusters particularly capture topic structure beyond chance.[14] General agreement between our comparative clustering analyses and initial statistical findings lends validity to the latter, demonstrating their value in understanding *and* modeling CSW.

6 Discussion

We find that specific discourse topics have notable relationships with several fine-grained aspects of Mandarin-English CSW in SEAME. These utterance-level relationships are sufficiently striking as to produce distinct values of CSW features across dimensions of multilingual spoken behavior that effectively distinguish between the topics being discussed in those code-switched utterances.

Our exploration and subsequent findings on how discourse topics relate to the presence of CSW echo and validate prior work on code-switches functioning as signals of topic shift, e.g. Wei (1998); Auer (2003), while our study of CSW quantity, frequency, language direction, and strategy reveal novel associations with topic at a level of detail previously not attained in discourse-functional work on CSW. Specific CSW patterns that group the *Academic* and *Professional* topics, and the *Personal* and *Cultural* topics, are reminiscent of prior qualitative work on an emotional detachment effect in CSW scenarios (Ladegaard, 2018; Ferreira, 2017). We speculate that the affective properties of certain kinds of discourse topics may similarly help determine the CSW style used to express them. Using Mandarin and English emotion lexica (Mohammad and Turney, 2010, 2013), we preliminarily find that utterances on *Personal* and *Cultural* topics have significantly greater emotional intensity than those on *Academic* and *Professional* topics (details in Appendix A). This aligns with our results on the association between discourse topic complexity and CSW strategy, and suggests that affect may modulate this interaction, though further work is required to confirm this hypothesis. Separately, we show that more formal topics (i.e. *Academic*, *Professional*) can involve CSW in speech, unlike prior studies that have primarily noted CSW in informal contexts, e.g. Bhattacharya et al. (2023). Finally, we show that many of the relationships we find can even be learned and applied by a simple unsupervised clustering model, lending validity to our statistical findings in a clearly interpretable manner.

7 Conclusion

We extensively examine the relationship between discourse topic and patterns of spontaneous CSW in the SEAME corpus. We find that (1a) certain discourse topics are much more likely to be expressed in code-switched utterances than monolingual ones; (1b) those discourse topics have significant associations with multilingual language production across previously unexamined patterns of CSW strategy, language direction, quantity, and frequency; (2) these associations lend themselves towards the inference of unique CSW profiles linked to specific

[14] Cramer's V measures show that the strength of association between topic labels and current clusters is several orders of magnitude greater than between topic labels and random, size-matched clusters (0.168 vs. 0.005; both p-values < 0.01).

(groups of) topics; (3) the statistical relationships found in (1) and (2) are salient enough to be learned and applied in part by unsupervised clustering models. We conclude that the nature of the discourse topic in conversation contributes meaningfully towards motivating diverse patterns of Mandarin-English CSW in speech. Our work's novelty is based in its context-sensitive approach towards understanding a dataset that we augment with new annotations across features of discourse and CSW. We hope this work will serve as a first step towards building improved models of CSW comprehension and informing the generation of authentic and discourse-informed multilingual speech.

Limitations

Our work focuses on a single language pair in a single corpus of CSW, which is somewhat skewed towards Mandarin relative to English. Both languages are represented only in the forms in which they are typically spoken in Singapore and Malaysia, in contrast to the majority of Mandarin-English code-switched corpora that are sourced from Mainland Chinese speakers. We acknowledge the need to extend our methods to the same language pair within different cultural contexts, and to additional language pairs with varying levels of typological distance, to test the robustness of our findings. We plan to do so in future work. Due to lack of access to CSW datasets, particularly those containing highly time-intensive manual discourse-level annotations and/or less discourse topic sparsity than in SEAME, our work makes use of the best currently available resources and serves as a reasonable first step towards understanding the role of discourse topics on code-switched speech production. For the *Cultural* topic in particular, we acknowledge that the relative corpus-level representation of this discourse topic in SEAME makes the associated findings, though novel and insightful, difficult to generalize. We are very interested in ultimately replicating our analyses on other CSW datasets, but also note that direct comparisons may be difficult since the categories and distributions of topics may differ across datasets.

With respect to our discourse topic classifiers, we note the inherent limitation of a single utterance receiving only a single label in our multi-class setup. By definition, this model design choice ignores the possibility of certain utterances dealing with multiple topics at a time by collapsing predictions into a single output label. Given the number of discourse topics we examine in this work, we believe this was nonetheless a reasonable design choice that prevented subsequent analyses from becoming overly complex.

Relatedly, it could have been helpful to incorporate additional features, such as Linguistic Inquiry and Word Count (LIWC) labels (Boyd et al., 2022), into our machine learning discourse topic classifiers. We speculate that such features covering psychological processes and personal concerns could have augmented the performance of our supervised models. However, it is difficult to reliably extract LIWC features from code-switched language, as this framework was originally developed for use in monolingual settings, and we leave this methodological extension to future work.

Finally, while our best-performing classifier achieves an accuracy of 72%, which is well above baseline performance, there remains 28% error in subsequently inferred discourse topic labels. This residual noise in the data could impact downstream statistical analyses. We handle this using error aware correction in our one-way ANOVA tests, and preliminarily find in Appendix A.12 that any remaining noise effectively has no impact on our current results. However, a fruitful direction for future work would be to replicate these downstream results by exploring alternative methods for deriving discourse topic labels, such as pre-trained multilingual transformer models and LLMs, e.g. mBERT or zero-shot GPT. We chose not to use these in the present work primarily in order to avoid issues arising from domain mismatch in pre-training data, which may not be sufficiently mitigated through fine-tuning due to a scarcity of appropriate code-switched data, as well as the relatively lower inherent transparency, interpretability, and modularity of these methods in comparison to each of our four classifiers. However, we acknowledge that in future work it may be worth trading off the drawbacks of these methods, as well as relevant cost and feasibility concerns, in favor of their potential to boost classification performance, which would increase the reliability of downstream analyses. Our deliberate design choice to avoid such models in the present work is particularly relevant since our main contribution is not to provide state-of-the-art model performance, but rather to leverage our current custom models to augment data and provide nuanced insights on that data.

Ethical considerations

This study was conducted exclusively on secondary data, and did not require human experiments. We did not access any information that could uniquely identify individual users within the corpus, as its original authors de-identified all speakers as outlined in the documentation of the dataset. Though we did not collect the data used in this work, we note that all participants in the original corpus had consented to sharing the data that we analyze in our study.

Acknowledgments

We thank Nicholas Deas, Lin Ai, and Chhavi Dixit for helpful discussions and feedback. This work was supported in part by the National Science Foundation under Grant IIS 2418307.

References

Kamisah Ariffin and Shameem Rafik-Galea. 2009. Code-switching as a communication device in conversation. *Language & Society Newsletter*, 5(9):1–19.

Kavita Asnani and Jyoti D Pawar. 2016. Use of semantic knowledge base for enhancement of coherence of code-mixed topic-based aspect clusters. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 259–266, Varanasi, India. NLP Association of India.

JC Peter Auer. 2003. A conversation analytic approach to code-switching and transfer. In *The Bilingualism Reader*, pages 167–187. Routledge.

Peter Auer. 1984. On the meaning of conversational code-switching.

Peter Auer. 1998. *Bilingual Conversation Revisited*, pages 1–24. Routledge, London, UK.

Ruthanna Barnett, Eva Codó, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, Marianne Starren, and Sietse Wensing. 2000. The LIDES Coding Manual: A document for preparing and analyzing language interaction data version 1.1—July, 1999. *International Journal of Bilingualism*, 4(2):131–270.

Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, and Niloy Ganguly. 2016. Functions of code-switching in tweets: An annotation framework and some initial experiments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1644–1650, Portorož, Slovenia. European Language Resources Association (ELRA).

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13(2):145–204.

Debasmita Bhattacharya, Jie Chi, Julia Hirschberg, and Peter Bell. 2023. Capturing formality in speech across domains and languages. In *Interspeech 2023*, pages 1030–1034.

Debasmita Bhattacharya, Siying Ding, Alayna Nguyen, and Julia Hirschberg. 2024a. Measuring entrainment in spontaneous code-switched speech. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2865–2876, Mexico City, Mexico. Association for Computational Linguistics.

Debasmita Bhattacharya, Eleanor Lin, Run Chen, and Julia Hirschberg. 2024b. Switching tongues, sharing hearts: Identifying the relationship between empathy and code-switching in speech. In *Interspeech 2024*, pages 492–496.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Jan-Petter Blom and John J. Gumperz. 1972. *Social meaning in linguistic structure: code-switching in Norway*, pages 75–96. Routledge.

Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. The development and psychometric properties of LIWC-22.

Mirjam Broersma. 2009. Triggered codeswitching between cognate languages. *Bilingualism: Language and Cognition*, 12(4):447–462.

Basudha Das. 2012. Code-switching as a communicative strategy in conversation. *Global Media Journal–Indian Edition*, 3(2):1–20.

Anik Dey and Pascale Fung. 2014. A Hindi-English code-switching corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Stanislav Dornic. 1978. *The Bilingual's Performance: Language Dominance, Stress, and Individual Differences*, pages 259–271. Springer US, Boston, MA.

A. Virginia Acuña Ferreira. 2017. Code-switching and emotions display in Spanish/Galician bilingual conversation. *Text & Talk*, 37:47 – 69.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge University Press.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 12–20, Austin, Texas. Association for Computational Linguistics.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. Building knowledge-guided lexica to model cultural variation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 211–226, Mexico City, Mexico. Association for Computational Linguistics.

Hans J. Ladegaard. 2018. Codeswitching and emotional alignment: Talking about abuse in domestic migrant-worker returnee narratives. *Language in Society*, 47(5):693–714.

Rosamina Lowi. 2005. Code switching: An examination of naturally occurring conversation. In *Proceedings of the 4th International Symposium on Bilingualism*, pages 1393–1406. Cascadilla Press Somerville, MA.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Chng, and Haizhou Li. 2010. Mandarin–English code-switching speech corpus in South-East Asia: SEAME. In *Language Resources and Evaluation*, volume 49, pages 1986–1989.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Pieter Muysken. 2000. *Bilingual speech: a typology of code-mixing*. Cambridge University Press.

Nanyun Peng, Yiming Wang, and Mark Dredze. 2014. Learning polylingual topic models from code-switched social media documents. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 674–679, Baltimore, Maryland. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. *Linguistics*, 18:581–618.

Ella Rabinovich, Masih Sultani, and Suzanne Stevenson. 2019. CodeSwitch-Reddit: Exploration of written multilingual discourse in online discussion forums. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4776–4786, Hong Kong, China. Association for Computational Linguistics.

Iliana Reyes. 2004. Functions of code switching in schoolchildren's conversations. *Bilingual Research Journal*, 28(1):77–98.

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The Role of Cognate words, POS Tags and Entrainment in Code-Switching. In *Interspeech 2018*, pages 1938–1942.

Li Wei. 1998. *The 'Why' and 'How' Questions in the Analysis of Conversational Code-Switching*, pages 156–176. Routledge, London, UK.

A Appendix

A.1 CSW quantity and frequency metric definitions.

CSW ratio measures the number of code-switches normalized by the token length of the utterance. This differs from M-index, which incorporates information about the utterance-level balance between language varieties present. Different from both metrics of CSW quantity, I-index reflects CSW frequency via the number of potential switch points in an utterance. All three metrics of CSW richness have a minimum value of 0, associated with monolingual utterances. The maximum value of CSW ratio approaches but does not equal to 1, while both M- and I-indices can achieve maximum values of 1, associated with a code-switched utterance evenly mixed between languages.

A.2 Discourse topic definitions and examples.

Academic: utterances discussing education (at primary, secondary, or university levels), research, studying, or coursework.

Example 1: 那种 他 exam. ["The kind he takes on an exam."]

Example 2: 他 是 full time tuition teacher. ["He is a full time tuition teacher."]

Example 3: 这样 他 还 可以 graduate in four years 还 蛮 强 的. ["This way he can graduate in four years, which is pretty impressive."]

Cultural: utterances discussing traditions, customs, festivals, cultural practices, holidays, or religious celebrations.

Example 1: 然后 你 会 一直 听 到 有 人 在 唱 Christmas carols 然后 很好 玩. ["Then you'll keep hearing people singing Christmas carols and it's fun."]

Example 2: Oh Chinese New Year 每年 都是 一样 的. ["Oh Chinese New Year is the same every year."]

Example 3: But 最少 他 还 preserve 他的 那个 culture 他 还是 会 Hokkien. ["But at least he still preserves his culture and he still knows Hokkien.]

Personal: utterances discussing hobbies, day-to-day/habitual experiences, opinions, feelings, preferences, family, friends, or other relationships.

Example 1: 可是 她 的 boyfriend 是 like 做- ["But her boyfriend likes to do - "]

Example 2: 他 就 觉得 like 就 跟 那个 the feeling is not right anymore because 我 明明 知道 你 背叛过 我 这样. ["He feels like the feeling is not right anymore because I clearly know you betrayed me.]

Example 3: I think when you break up 跟 你们 可以 做 继续 做 friend 是 因为 at that time 你们 真的 没有 很 在一起 那种 感觉. ["I think when you break up you can continue to be friends because at that time you didn't really feel like you were together.]

Professional: utterances discussing work life and technical aspects of a job, including the use of technology or programming.

Example 1: Actually what I did was to source for 他们 客户 的 data. ["Actually what I did was to source for their clients' data."]

Example 2: start 你的 business 你 可以. ["You can start your business."]

Example 3: 就 比如 我 interested in supply chain 如果. ["For example, if I am interested in the supply chain."]

A.3 Instructions for discourse topic annotation task.

Please label the following utterances for topic of conversation discussed: Academic*,* Professional*,* Personal*,* Cultural*, or* Other*. Below are guidelines to help you distinguish between topics.*

- Academic: *Topics related to education, research, school, university, study, or coursework.*
  - *Examples: discussions about classes, research projects, GPA, or teachers.*
- Professional: *Topics related to work, technology, programming, or aspects of a job, especially technical ones.*
  - *Examples: discussions about work, company, office, salary, etc.*
- Personal: *Topics focused on personal life, hobbies, family, friends, feelings, or relationships.*
  - *Examples: conversations about family members, personal emotions, thoughts, preferences, or day-to-day/habitual experiences.*
- Cultural: *Topics related to traditions, festivals, cultural practices, holidays, or religious celebrations.*
  - *Examples: discussions about cultural holidays*/events, or traditional customs. *Not just mentions of vacations or trips.*

- Other: *If the utterance does not clearly fit into any of the above categories.*
  - *Examples: Short utterances will tend to fit into this topic.*

A.4 Validating the current set of core discourse topics.

We believe the current granularity of discourse topics studied represents a reasonable starting point for this work. To validate this, we perform additional exploratory topic modeling, with an automatic number of topics from a BERTopic model, on each predefined topic in the corpus to assess whether additional, finer-grained topics emerge from any (see Appendix A.13 for implementation details).

While each of the *Personal*, *Academic*, and *Professional* topics demonstrate related subdivisions, (discussing relationships with friends/family vs. thoughts/feelings/preferences; studying for exams vs. specific school subjects; technical jargon vs. professional roles and responsibilities), none of these is distinct enough from its parent topic to warrant defining a distinct new topic; see Appendix A.5 for detailed examples. Thus, we confirm the appropriate granularity of the set of topics we choose to study.

A.5 Examples of related but non-distinctive sub-topic clusters within *Academic*, *Personal*, and *Professional* topics.

A.5.1 *Academic*.

- Sub-topic 1 (discussing exam preparation): 学校, exam, study, 备考, studying.
- Sub-topic 2 (discussing specific school subjects): school, lecture, maths, science, german.

Other *Academic* sub-topics are defined mainly by functional words, indicating that more granular topics do not exist within the *Academic* topic, e.g. Sub-topic 3: 我们, then, 没有, 那些, 东西.

A.5.2 *Personal*.

- Sub-topic 3 (discussing thoughts and preferences): 觉得, 这样, 比较, like, think.
- Sub-topic 5 (discussing interpersonal relationships): 他们, my, friend, 我的, parents.

Other *Personal* subtopics are also defined mainly by functional words, indicating that more granular topics do not exist within the *Personal* topic, e.g. Sub-topic 1: that, is, it, and, then.

A.5.3 *Professional*.

- Sub-topic 3 (discussing professional roles and responsibilities): lead, business, project, manager, fulltime.
- Sub-topic 5 (discussing jargon-like aspects): processing, 一个, job, software, data.

Other *Professional* subtopics are defined with related words, but in a less clearly cohesive way, indicating that more granular topics do not exist within the *Professional* topic, e.g. Sub-topic 4: part, time, 我们, 现在, 做工.

A.6 Seed words used for initial rule-based classifier.

- *Academic*: course, class, unit, lesson, lecture, batch, final, review, conference, presentation, reference, archive, result, semester, sem, academic, student, admit, scholar, teacher, prof, report, learn, uni, school, book, chapter, syllabus, read, paper, essay, econ, math, physics, chem, bio, science, psychology, grade, point, score, credit, fail, committee, major, master, phd, thesis, module, subject, average, analyze, analyse, honours, junior, question, lab, diploma, percent, quiz, exam, tutor, tuition, enroll, prove, uniform, graduate, orientation, levels, recess, homework, primary, secondary, year, study, engineering, gpa, studies, college, research, education, edu, pre, 课程, 班级, 单元, 课程, 讲座, 批次, 期末, 复习, 会议, 演示, 参考, 存档, 结果, 学期, 学期, 学术, 学生, 录取, 学者, 老师, 教授, 报告, 学习, 大学, 学校, 书, 章节, 教学大纲, 阅读, 论文, 文章, 经济, 数学, 物理, 化学, 生物, 科学, 心理学, 成绩, 分数, 得分, 学分, 不及格, 委员会, 专业, 硕士, 博士, 论文, 模块, 科目, 平均, 分析, 分析, 荣誉, 初级, 问题, 实验室, 文凭, 百分比, 测验, 考试, 导师, 学费, 注册, 证明, 制服, 毕业, 迎新, 水平, 休息, 家庭作业, 小学, 中学, 年级,学习,班,工程,课,章节,章,学院,研究,科研,项目,教育.
- *Cultural*: christmas, carol, halloween, new year, typical, traditional, pray, buddha, jesus, islam, church, god, goddess, red envelope, gathering, taoist, bible, scripture, christian, orthodox, holiday, holidays, religion, lantern, china, mid-autumn, 圣诞节,圣诞,万圣, 颂歌, 万圣节, 新年, 典型, 传统, 祈祷, 佛, 耶稣, 伊斯兰教, 教堂, 神, 女神, 红包, 聚会, 道教, 圣经, 经文, 基督教, 正统, 假日,节

假日,元宵,元宵节,放假,宗教,红包,拜年,鞭炮,花灯,灯谜,中国,春节,过年.

- *Personal*: think, feel, feeling, understand, believe, trust, like, know, prefer, want, miss, worry, regret, stress, remember, happy, sad, afraid, miserable, excite, mad, anger, angry, friend, cousin, mother, father, mom, dad, parent, child, brother, sister, sibling, uncle, aunt, daughter, son, family, relative, husband, wife, boyfriend, girlfriend, fun, together, individual, personal, relationship, play, piano, rugby, tennis, badminton, soccer, football, basketball, hobby, usual, often, home, house, room, live, birthday, community, facebook, show, name, person, people, identity, favourite, favorite, swim, self, 思考, 感觉, 感受, 理解, 相信, 信任, 喜欢, 知道, 更喜欢, 想要, 想念, 担心, 后悔, 压力, 记住, 快乐, 悲伤, 害怕, 痛苦, 兴奋, 生气, 愤怒, 朋友, 表亲, 母亲, 父亲, 妈妈, 爸爸, 父母, 孩子, 兄弟, 姐妹, 兄弟姐妹, 叔叔, 阿姨, 女儿, 儿子, 家庭, 亲戚, 丈夫, 妻子, 男朋友, 女朋友, 乐趣, 一起, 个人, 私人, 关系, 玩, 钢琴, 橄榄球, 网球, 羽毛球, 足球, 篮球, 爱好, 平常, 经常, 家, 房子, 房间, 生活, 生日, 社区, 脸书, 表演, 姓名, 人物, 人物, 身份, 最喜欢, 游泳, 哥哥,哥,姐姐,姐,妹妹,妹,弟弟,弟,爸爸,爸,妈妈,妈,爷爷,奶奶,外公,外婆,祖父,祖母,自己,目前,开心,伤心,难过,悲伤,悲哀,恐怖.

- *Professional*: freelance, position, job, part-time, occupation, apply, application, work, interview, dollar, cent, salary, technology, program, boss, colleague, staff, regression, model, correlation, correlate, download, sensor, email, database, internet, website, system, algorithm, bug, update, server, warranty, business, service, user, experience, audit, consult, career, manage, stock, portfolio, project, system, procedure, develop, design, quality, team, equipment, lead, produce, function, tool, skill, consumer, customer, employee, contract, information, solve, solution, profit, design, machine, paperwork, training, zoom, company, firm, hierarchy, maintain, chip, weld, manufacture, manufacturing, property, properties, special, identify, admin, bank, software, tech, troubleshoot, industry, data, recruit, hire, offer, trademark, market, competition, government, capital, promotion, reboot, protocol, profit, commission, downstream, commercial, industrial, stage, tutorial, manage, manager, interface, 自由职业, 职位, 工作, 兼职, 职业, 申请, 申请, 工作, 面试, 美元, 分, 薪水, 技术, 程序, 老板, 同事, 员工, 回归, 模型, 相关性, 相关, 下载, 传感器, 电子邮件, 数据库, 互联网, 网站, 系统, 算法, 错误, 更新, 服务器, 保修, 业务, 服务, 用户, 经验, 审计, 咨询, 职业, 管理, 股票, 投资组合, 项目, 系统, 程序, 开发, 设计, 质量, 团队, 设备, 领导, 生产, 功能, 工具, 技能, 消费者, 客户, 员工, 合同, 信息, 解决, 解决方案, 利润, 设计, 机器, 文书工作, 培训, 缩放, 公司, 公司, 层次结构, 维护, 芯片, 焊接, 制造, 制造业, 财产, 属性, 特殊, 识别, 管理员, 银行, 软件, 技术, 故障排除, 行业, 数据, 招聘, 雇用, 提供, 商标, 市场, 竞争, 政府, 资本, 促销, 重新启动, 协议, 利润, 佣金, 下游, 商业, 工业, 阶段,做工,做,工,教程,经理,界面.

A.7 Performance of rule-based classifiers on full ground truth set.

Classifier	Accuracy	F1 Score
Initial lexicon-based	0.61	0.54
Expanded lexicon-based	0.59	0.53

Table 11: Rule-based classifiers' accuracy and macro F1 score on discourse topic labeling over the entire ground truth set.

A.8 Filler words used in ensemble and self-supervised classifiers.

Um, ah, uh, eh, er, hmm, mm, mmm, umm, ar, hm, 啊, 呃, 呃, 啊, 呃, 嗯, 嗯,哼.

A.9 Classifier models' hyperparameter settings.

In the **ensemble classifier**, the `TfidfVectorizer` used has its maximum features set to 2000 and its analyzer set to 'word'. All other parameters are left at default values. All numeric features used in this model are normalized to zero mean and unit variance using `StandardScaler`. The ensemble architecture consists of three base models and one meta model. The first base model is a logistic regression model with the following hyperparameters: $C = 0.9$, class weight = 'balanced', maximum iterations = 3000, and solver = 'saga'. The second base model is a random forest classifier with the following hyperparamters: number of estimators = 300, maximum depth = 20, class weight = 'balanced', and random state = 57. The

third base model is a gradient boosting classifier with the following hyperparameters: number of estimators = 250, learning rate = 0.03, maximum depth = 6, subsample = 0.8, and random state = 57. All base models are calibrated using `CalibratedClassifierCV` with cv = 3. Calibration of output probabilities uses Platt scaling. The meta model is a logisic regression model with default hyperparameter settings. The train/test split for the ensemble is determined using random state = 57.

In the **self-supervised classifier**, the `TfidfVectorizer` used has its maximum features set to 3000, ngram range set to (1, 2), and minimum df set to 3. As with the ensemble model, the numeric features used in this classifier are normalized to zero mean and unit variance using `StandardScaler`. The sentence embeddings used have output dimension = 384. These are converted to sparse CSR and stacked horizontally. The base model for the self-supervised classifier is a logistic regression with solver = 'liblinear', class weight = 'balanced', and maximum iterations = 3000 (2000 in loop iterations). This base model is calibrated using `CalibratedClassifierCV` with cv = 5 and method = 'sigmoid'. Output probabilities are also calibrated using Platt scaling. As with the ensemble classifier, data splits for training and evaluation are determined with random state set to 57. In addition, the stratify parameter is set to y. The self-training loop has its number of iterations set to 5, with base per-class confidence thresholds as follows: *Academic* = 0.6, *Cultural* = 0.3, *Other* = 0.9, *Personal* = 0.8, *Professional* = 0.9. The maximum pseudo-labeled samples per class per iteration are as follows: *Academic* = 100, *Cultural* = 200, *Other* = 100, *Personal* = 100, *Professional* = 100. For the *Cultural* class in particular, we implement dynamic thresholding, which we adjust using the 90th percentile probabilities for the class. This can be lowered slightly if insufficient samples are added, using dynamic adjustment = 0.05. We also implement hard negative mining for the *Cultural* class. After self-training, misclassified *Cultural* samples in the validation set are oversampled threefold and added back into the training set. Re-training occurs with these hard examples.

For both of the above models, we manually tuned hyperparameter settings until we found a good set of values that produced reasonable per-topic performance, especially on minority classes in the data.

A.10 Ablation studies on ensemble and self-supervised classifiers.

For the ensemble classifier, sentence embeddings, tf-idf, and lexicon count features contribute slightly positively to model performance and to roughly equal degrees, as demonstrated by the small decreases in accuracy resulting from each of their exclusion from the model pipeline (Table 12). In contrast, the exclusion of acoustic-prosodic and other lexical features from the model improves model performance, suggesting that these features are detrimental to accurate classification decisions.

Excluded feature group	Accuracy	F1 score
–	0.68	0.62
tf-idf	0.67	0.67
Lexicon seed word counts	0.67	0.67
Lexical	0.69	0.69
Acoustic-prosodic	0.71	0.70
Sentence embeddings	0.67	0.67

Table 12: Comparing ensemble classifier accuracy and macro F1 score across subsets of the entire feature set. The first row, where no features are excluded, denotes the performance of the model on the entire feature set, as originally shown in Table 1.

Excluded feature group	Accuracy	F1 score
–	0.72	0.71
tf-idf	0.70	0.69
Lexicon seed word counts	0.70	0.69
Lexical	0.71	0.71
Acoustic-prosodic	0.72	0.71
Sentence embeddings	0.68	0.67

Table 13: Comparing self-supervised classifier accuracy and macro F1 score across subsets of the entire feature set. The first row, where no features are excluded, denotes the performance of the model on the entire feature set, as originally shown in Table 1.

For the self-supervised classifier, sentence embedding features are the single biggest positive contributor to model performance, as demonstrated by the drop in accuracy from its exclusion (Table 13). Tf-idf and lexicon count features also contribute positively and roughly equally. On the other hand, the exclusion of acoustic-prosodic and other lexical features from the model does not affect performance, indicating that these may act as a source of noise instead of a model signal. These patterns are generally consistent with those from ablations over the ensemble classifier's features.

A.11 Per-class performance of our best-performing (self-supervised) topic classifier.

We report per-class performance metrics and corresponding confusion matrix statistics for our best-performing topic classifier to derive further insight into which classes drive the overall performance of this self-supervised model (Table 14 and Figure 1).

Class	Precision	Recall	F1	Support
Academic	0.78	0.67	0.72	227
Cultural	0.85	0.71	0.77	24
Personal	0.67	0.66	0.67	685
Professional	0.74	0.53	0.62	189
Other	0.73	0.81	0.77	914

Table 14: Self-supervised classifier performance over the held-out test set, stratified by class, i.e. discourse topic. Recall that our best-performing topic classifier achieves an overall accuracy of 0.72, corresponding to a macro F1 score of 0.71 (Table 1). Given the respective class-level support values, we calculate the contribution to model accuracy of each class, in order: 0.07 (*Academic*), 0.01 (*Cultural*), 0.22 *Personal*, 0.05 (*Professional*), and 0.36 (*Other*). Thus, it appears that the overall performance of the model is primarily driven by the *Other* class (corresponding to 36.3% of correct predictions) and the *Personal* class (corresponding to 22.2% of correct predictions), while the other three discourse topic classes contribute relatively little.

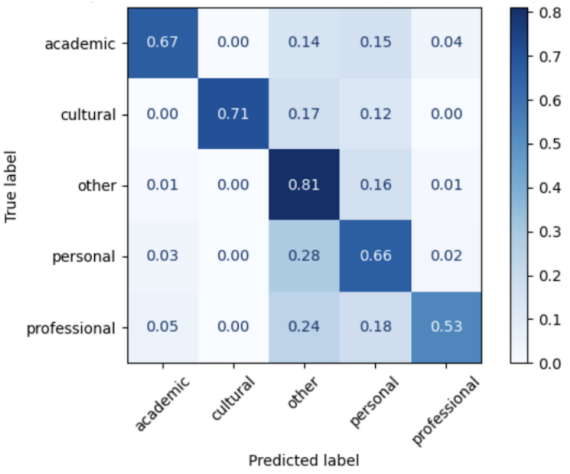

Figure 1: Row-normalized confusion matrix associated with our best-performing topic classifier. This provides additional support for the *Other* topic driving the major part of this self-supervised model's performance.

A.12 Discourse topic label distribution in subsets of the corpus.

Discourse topic	% of corpus	% of GT	% of AA
Academic	7.8	11.0	7.5
Cultural	0.1	1.2	0.1
Personal	28.1	34.1	27.5
Professional	2.4	9.1	1.7
Other	61.5	44.6	63.2

Table 15: Discourse topic label distribution across classes in the entire SEAME corpus, the ground truth (GT) subset of the corpus, and the automatically-annotated (AA) subset of the corpus. This breakdown of results helps to diagnose a potential source of class imbalance in the fully-annotated corpus and may point to some distributional shift induced by our best-performing classifier. However, in spite of this, note that we replicate all the patterns described in Section 5.2 when we run the same analyses on 1) data from the ground truth set only and 2) data from high confidence subsets ($P >= 0.7$, $P >= 0.8$, $P >= 0.9$) of the automatically-annotated portion of the corpus; the subsequent results remain statistically significant with the same general trends by discourse topic for CSW presence, strategy, direction, quantity, and frequency, though the exact values of odds ratios are slightly different. Similarly, we also replicate the majority of clustering patterns described in Section 5.3 using both 1) the ground truth subset and 2) the high-confidence subsets of the automatically-annotated portion of the corpus; cluster means in each case closely match those corresponding to clusters in Section 5.3. Combined, these replications reinforce the reliability of our current findings.

A.13 Exploratory LDA and BERTopic model hyperparameter settings.

For our LDA analysis on the *Other* topic reported in Section 5.1, we defined a custom list of English and Mandarin filler words (see Appendix A.8) based on prior work (Bhattacharya et al., 2024a). We treated these as stopwords. When building the model vocabulary, we ignored terms that had a document frequency higher than 0.9, and used a cut-off value of 50. We used 5 as the number of topics, which agreed with the automatic number of topics yielded by the BERTopic model with `all-MiniLM-L6-v2` embedding for the same analysis. We used the default 'batch' learning method for LDA. For the exploratory sub-topic analyses on the *Academic*, *Personal*, and *Professional* topics (see Appendix A.4), we use the same model implementation details.

A.14 Unsupervised clustering initialization.

We select starting points at random to ensure that final k-means clustering results in Section 5.3 are not unduly influenced by initial points. We verify the validity of the conclusions that directly follow this design choice by re-running our current clustering implementation across 30 additional random seeds using different random starting points. We also try an initialization setting that uses one utterance from each topic as starting points for clustering. In each case, the resulting clusters retain the overall trends in cluster compositions that we report in Section 5.3, though the exact proportions in each cluster change slightly. This demonstrates that that our qualitative conclusions are robust to initialization and validates our design choice.

A.15 Investigating the interplay of affect and discourse topic.

We conduct exploratory analysis to follow up on our hypothesis of a relationship between affect and discourse topic that goes on to shape the CSW patterns used to express those topics. To begin to verify the extent to which each discourse topic uses affective language, we combine English and Mandarin emotion lexica from Mohammad and Turney (2010, 2013) and calculate utterance-level normalized emotional intensity scores across eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, disgust) and two categories of sentiment (positive, negative). We bin utterances into high- or low-emotional intensity, based on the corpus-level median normalized emotional intensity score. We then perform chi-squared tests to compare emotional intensity across topics (Table 16). On average, the *Personal* and *Cultural* topics have greater odds of expressing high emotional intensity than low emotional intensity, relative to the *Academic* and *Professional* topics. Based on these results, we posit that greater affect is linked to discourse topics that require less up-front subject knowledge to discuss, i.e. less complex topics, which in turn allows for their expression using more structured and complex CSW strategies.

Topic	χ^2	***p*-val.**	**OR**	**95% CI**
Academic	185.8	**	1.36	[1.30, 1.42]
Cultural	4.0	*	1.39	[1.01, 1.91]
Personal	3950.7	**	2.34	[2.28, 2.41]
Professional	231.8	**	1.80	[1.67, 1.95]
Other	4982.2	**	0.41	[0.40, 0.42]

Table 16: Chi-squared tests and odds ratios comparing topics in high and low emotional intensity utterances. p-values less than 0.05 are denoted by *. p-values less than 0.01 are denoted by **.