Discourse-Driven Code-Switching: Analyzing the Role of Content and Communicative Function in Spanish-English Bilingual Speech

Debasmita Bhattacharya and Juan Junco and Divya Tadimeti and Julia Hirschberg
Department of Computer Science
Columbia University
New York, NY, USA
debasmita.b@cs.columbia.edu, jej2162@columbia.edu, dt2760@columbia.edu, julia@cs.columbia.edu

Abstract

Code-switching (CSW) is commonly observed among bilingual speakers, and is motivated by various paralinguistic, syntactic, and morphological aspects of conversation. We build on prior work by asking: how do discourse-level aspects of dialogue – i.e. the content and function of speech – influence patterns of CSW? To answer this, we analyze the named entities and dialogue acts present in a Spanish-English spontaneous speech corpus, and build a predictive model of CSW based on our statistical findings. We show that discourse content and function interact with patterns of CSW to varying degrees, with a stronger influence from function overall. Our work is the first to take a discourse-sensitive approach to understanding the pragmatic and referential cues of bilingual speech and has potential applications in improving the prediction, recognition, and synthesis of code-switched speech that is grounded in authentic aspects of multilingual discourse.

1 Introduction

Code-switching (CSW) occurs when a speaker alternates between languages (Poplack, 1980), and may be performed between or within utterances in various language pairs. CSW is common among bilingual speakers who may produce 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. "*Pero mi* printer *no funciona.*"
["But my printer doesn't work."]
b. "*No la puedes hacer* because you can't check it."
["You can't do it because you can't check it."]

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

Prior paralinguistic work has shown that a range of speaker and listener attributes affect or correlate with the prevalence of CSW during conversation. These include interlocutor gender (Gardner-Chloros and Edwards, 2004; Finnis, 2014), linguistic competency (Dornic, 1978; Bhattacharya et al., 2025), and affective state (Ferreira, 2017; Bhattacharya et al., 2024b), among many others. Previous work taking a more syntactic and/or morphological approach to understanding how and why speakers code-switch has proposed a number of competing grammatical constraints governing CSW to differing extents (Joshi, 1982; Myers-Scotton, 1993; Poplack, 1978; MacSwan, 2000; Tsoukala et al., 2019).

While many levels of linguistic analysis are well-represented in prior studies of CSW, much scope remains for exploring how specific discourse-level aspects of dialogue influence CSW in speech. In particular, little is currently known at scale about the independent and interactive relationships between the content and intended function of bilingual speech and their downstream impact on spontaneous, conversational CSW production. As CSW becomes an increasingly common phenomenon in a globalized world, it is important to develop and apply such insight into bilinguals' motivation for CSW in implementing downstream applications to serve this growing community of speakers. We believe the best way to do so is by building a holistic understanding of CSW based on multiple dimensions of linguistic analysis, particularly those that have historically received less attention in the field.

In this work, we explore a Spanish-English spoken corpus and examine the links between CSW and speech **content** and **function**, as encoded by **named entities** and **dialogue acts** expressed during conversation. Our extensive analyses show significant associations of both discourse content and function with various aspects of CSW, including CSW quantity and syntactic structure. The varying

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23416–23431
November 4-9, 2025

degree of these associations can be leveraged to inform downstream predictions of transition models, which offer preliminary insight into the interactive effect of discourse features on bilingual speech.

Our contributions include 1) creating new annotations on multiple dimensions of a CSW corpus, which we share at https://tinyurl.com/y6zfb86w; 2) applying rigorous statistical testing to identify novel and nuanced quantitative and qualitative insights into the role of discourse in shaping CSW; and 3) incorporating these insights into building discourse-informed predictive models of CSW whose performance corroborates statistical testing. Our work is the first to take such a thorough, discourse-sensitive approach to understanding pragmatic and referential cues of bilingual speech and has potential for improving the recognition and synthesis of naturalistic CSW that is grounded in authentic multilingual discourse patterns.

2 Related Work

Discourse-level analysis of CSW. Early discourse-functional work on CSW has highlighted the challenge of attributing specific discourse meanings to particular code-switches, as a single code-switch can simultaneously perform multiple functions (Stroud, 1992). Several studies have since proposed various taxonomies for understanding when and why speakers code-switch: to signal changes in speech setting, listener, semantic topic, and speaker affect, as well as contrasts between direct and reported speech, and emphasized and parenthetical speech (Blom and Gumperz, 1972; Auer, 2007; Lowi, 2005). Such facets of discourse framing have been observed across language pairs, modalities, speaker ages, and both native-level and language-learning CSW settings (Auer, 2003; Wei, 1998; Reyes, 2004; Ariffin and Rafik-Galea, 2009; Das, 2012; Dey and Fung, 2014; Begum et al., 2016; Liebscher and Dailey-O'Cain, 2005; Almusallam, 2024). However, these taxonomies were often derived from the idiosyncratic characteristics of the specific dataset under investigation in each case, making their discourse-functional explanations of CSW difficult to unify and extend to other corpora. Also, while CSW is a key feature of the corpora examined in such studies, few examples of prior work have studied its interaction with discourse beyond the raw number of code-switches present in an utterance. Though Hartmann et al. (2018) attempted to address some of these limits, aspects of CSW outside of the context of automatic identification or derivation of code-switched discourse-structuring functions remain unexplored.

Named entities in CSW. As alluded to in Ahn et al. (2020), named entities (NEs) often feature in code-switched contexts, particularly when speakers employ insertional CSW and refer to names of people, places, and organizations in conversation. However, most related work has either focused on improving the performance of machine learning models on the task of NE recognition on code-switched datasets, e.g. Aguilar et al. (2018, 2020); Whitehouse et al. (2022); Sterner (2024), or studied only broad lexical classes and categories of entities which do not necessarily include proper NEs, e.g. Parekh et al. (2020); the question of the discourse-structuring influence of NEs on CSW, or vice versa, has largely been ignored. One question this work will address is **RQ1:** How are patterns of CSW influenced by the content of the utterance, in terms of the presence, type, and distribution of NEs?

Dialogue acts and CSW. Dialogue acts (DAs) are an inherent aspect of discourse structure and are particularly salient in spontaneous, unplanned conversation (Stolcke et al., 2000). Understanding the type of dialogue is essential for inferring speaker intent and communicative function within the context of a given utterance or conversation (Duran and Battle, 2018). However, there is as yet no work examining bilingual speakers' conversational intent in code-switched speech from the perspective of a unified set of DAs, such as the Switchboard Dialogue Acts (SwDA) tag set (Jurafsky et al., 1997). While some recent work has used characteristics of CSW to classify a mix of discourse functions and multilingual speech properties (Belani and Flanigan, 2023), none has asked the question we pose in **RQ2:** How is CSW influenced by the function of the utterance, as represented by DAs from SwDA, and how does this interact with patterns stemming from the content of code-switched speech?

3 Method

Corpus. We examine the Bangor Miami (BM) corpus of spontaneous, informal conversations (Deuchar, 2011).[2] BM consists of 35 hours of recorded conversation and 46.9k transcribed utterances across 56 dialogues, i.e. 837.5 utterances per conversation on average. 84 unique speakers from

[2]This corpus is made available under the GNU General Public License version 3 or later.

Miami are represented in the corpus. The recorded dialogues comprise a mix of monolingual English, monolingual Spanish, and code-switched Spanish-English utterances, all of which were manually annotated with token-level language identification (LID) tags upon transcription by Deuchar.

Data annotation and pre-processing. A fluent Spanish/English bilingual annotates the BM corpus for *NEs*, noting their utterance-level transcript text, position, language, and type. We subdivide types of NEs into the following categories: *N*: names of people; *O*: names of organizations, institutions, and companies; *P*: names of places, including cities, countries, and districts, as well as demonyms; *B*: brand names, products, and commercial labels, including movie and television program titles; *T*: temporal expressions including days of the week, holidays, and other calendar events; *R*: references to religious figures, texts, and expressions; and *U*: all other NEs that do not fall into any of the previous categories.[3] A native Spanish/English bilingual[4] adds utterance-level annotations of *DAs* to the BM corpus, applying the SwDA tag set for conversational function and intent (see Appendix A.2 for the complete list of DA labels used). Finally, for only the code-switched BM utterances, we use the provided LID tags to calculate utterance-level CSW quantity based on 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), in order to enrich prior annotations of the CSW strategies present in BM, of which 72% are insertional and 13% are alternational (Bhattacharya et al., 2024a).[5] We also use the LID tags to determine language predominance within code-switched utterances, comparing counts of content tokens spoken in each language.

Statistical analyses. Following data annotation, we perform statistical analyses of *NEs* and *DAs* across and within the monolingual and code-switched subsets of the corpus, examining their relationship with various aspects of CSW. In the portions of analysis that examine proximity of NEs to code-switches, for simplicity, we consider only those transcripts containing a single NE. When using dependency distance, i.e. the number of words intervening between two syntactically related words, as a measure of proximity, we first perform automatic translation of the code-switched transcripts to English using the Google Translate API, to simplify the calculation of utterance-level dependency relations.[6] We use `SpaCy 3.8.5` to extract dependency parses of the translated transcripts and compute dependency distance between NEs and (translated) code-switched segments.[7]

Predictive modeling. We incorporate the results of our statistical analyses into two transition-based predictive models of CSW, one using logistic regression (12 params.), and the other using a supervised two-state hidden Markov model (HMM; 50 params.).[8] We choose the former to model local cues at the utterance-level, and the latter to model global dynamics between monolingual and code-switched states across an entire dialogue; details on why each model is particularly suited to its task, and both models' hyperparameter settings, are in Appendix A.3. We randomly divide the BM corpus into an 80%-20% train-test split, at the utterance- and conversation-level for the logistic regression and HMM base models, respectively. We mitigate class imbalance between the majority class of monolingual utterances and minority code-switched class by implementing the Synthetic Minority Over-sampling Technique (SMOTE) (Galli, 2023) on the training set for the logistic regression base model, generating synthetic code-switched examples until we achieve parity with the monolingual class. For the HMM training set, we apply a weighted-transition scheme inspired by W-Trans (Khan and Siddiqi, 2020), multiplying every transition involving the CSW state by a factor $w = 5$. We visualize this scheme in Appendix A.3.

[3] We choose these categories of NEs based on observations made during an initial pass over the corpus, with the goal of minimizing U-type entities. Further annotation-related details are in Appendix A.1.

[4] We discuss the implications of our annotator demographic profiles on data curation in detail under Limitations.

[5] 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 distribution of language varieties present. All three metrics 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.

[6] We choose this method based on English and Spanish generally sharing the same SVO word order. For a more detailed discussion of how translation may impact dependency distance, see the Limitations section.

[7] We also perform this analysis using `stanza 1.10.1`. Our subsequent results replicate those obtained from using `SpaCy`, so we report only the `SpaCy` results in the paper.

[8] Models are trained in under an hour on a Mac M1 chip.

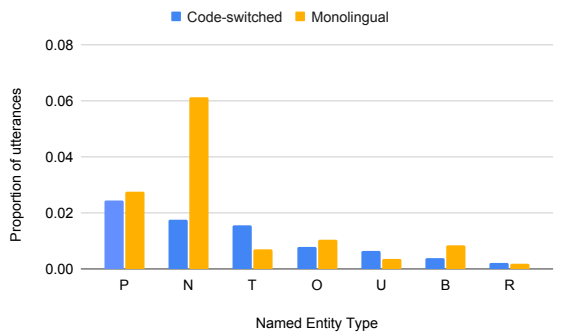

Figure 1: Distribution of named entity types across code-switched and monolingual contexts.

4 Results

4.1 How is CSW influenced by its content?

Named entities differ between code-switched and monolingual contexts. We begin by studying differences in the presence and type of NEs between monolingual and CSW contexts, and whether these might indicate a prompting effect of NEs on CSW, relative to monolingual speech. First, we find that only 9.8% of BM utterances contain NEs; 10.0% and 6.1% of monolingual and code-switched utterances respectively contain NEs, indicating that NE use is generally quite rare in the corpus, and particularly so in code-switched contexts. According to a two-sample z-test of proportions, the proportion of utterances containing NEs is significantly different ($p < 0.001$) between monolingual and code-switched contexts.

NEs also vary widely by type across code-switched and monolingual contexts (Figure 1). The most commonly used NEs across the entire BM corpus are references to places (P; e.g. "*Inglaterra*"), people (N; e.g. "*Marta*"), and temporal expressions (T; e.g. "Friday"). References to temporal expressions (T) and religious figures, texts and other expressions (R; e.g. "*Dios*") are more frequently associated with code-switched utterances, potentially reflecting specific usage patterns that are unique to bilingual discourse. Conversely, NEs referring to names of places (P), people (N), and brand names, products, and commercial labels (B; e.g. "Toyota") appear more frequently in monolingual utterances, pointing again to unique contextual dependencies that may be determined by the linguistic setting.

To verify the statistical significance of these trends, we perform chi-squared tests and odds ratio calculations to determine whether utterances containing a given type of NE are more or less likely to also involve CSW (Table 1). Across types of NEs, statistical testing reveals that utterances containing a NE are generally *less* likely to contain code-switches compared to those without NEs. This reflects the slightly greater representation of NEs in monolingual utterances compared to code-switched ones in the corpus, and is statistically significant for N- and B-type NEs in particular. The odds that an utterance contains an N- or B-type NE, given that it is code-switched, are less than half of those for a monolingual utterance. This aligns with the relative distribution of these NEs across contexts, as shown in Figure 1. The only type of NE for which the opposite trend is statistically significant is T-type NEs; the odds that an utterance contains a T-type NE, given that it is a code-switched utterance, are more than twice those of a monolingual utterance. This, too, reinforces the relative distributions across contexts in Figure 1, and presents evidence of NE type being an influential factor that helps to distinguish CSW from monolingual speech.[9] The differences we find may be due to phonotactic (in)compatibility of NEs with surrounding context, relative ease of cognitive access to NEs across languages, or other cultural or sociopragmatic factors whose study is presently out of scope, and hence left to future work.

Overall, the difference in NE distribution between code-switched and monolingual utterances in BM suggests that NEs may serve as linguistic *anchors* that effectively discourage CSW, rather than triggers that prompt CSW in this corpus.

Entity Type	χ^2	*p*-val.	OR	95% CI
P	0.126	–	0.93	[0.66, 1.28]
N	45.99	*	0.32	[0.22, 0.45]
T	18.06	*	2.44	[1.55, 3.69]
O	2.460	–	0.57	[0.26, 1.09]
U	0.999	–	1.56	[0.66, 3.21]
B	8.675	**	0.20	[0.04, 0.59]
R	0.0	–	0.92	[0.18, 2.83]

Table 1: Summary of chi-squared tests and odds ratios comparing NE presence in monolingual and code-switched utterances. p-values below 0.05 and 0.01 are denoted by * and **, respectively. p-values above 0.05 are denoted by –.

Named entities vary with aspects of CSW. We now explore the language and syntactic character-

[9]We rule out linguistic accommodation as a potential confound to these findings by adapting the method used in Bhattacharya et al. (2024a), and finding no statistically significant evidence of conversation- or turn-level convergence on NE usage in the BM corpus.

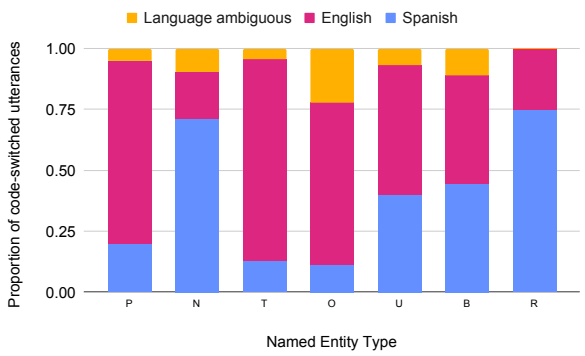

Figure 2: Distribution of named entity types by language of production in code-switched utterances.

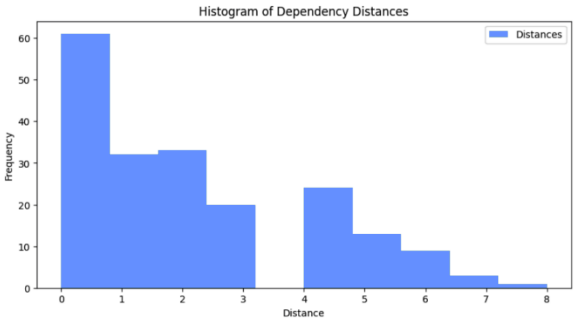

Figure 3: Distribution of dependency distances.

istics of NEs in code-switched utterances and how these specifically relate to patterns of CSW in BM.

Across NE types, the language in which entities are produced varies, although most are predominantly produced in English (Figure 2). Notable exceptions are N- and R-type entities, which are predominantly produced in Spanish (e.g. "*Pedro*", "*Jesús*"). These language-specific patterns suggest linguistic preferences or contextual cues that might influence language choice in NE production in a code-switched setting. We further explore a potential relationship between NEs and language choice through chi-squared tests to determine whether NE types differ significantly when code-switches occur from English to Spanish compared to those occurring from Spanish to English. However, initial tests do not yield significant results, suggesting that more sophisticated tests with interaction features may be required to model this relationship.

Next, we find that, on average, the raw token distance between a NE and the subsequent code-switched segment within the utterance ranges from 1.0 to 3.5, depending on the type of entity (Table 5 in Appendix A.1), suggesting possible positional relationships between NEs and code-switches. Analysis of dependency relations between NEs and code-switched segments shows a generally negative relationship between them (Figure 3), indicating that NEs are most often directly involved in or immediately adjacent to CSW events, with few intervening syntactic elements. As dependency distance grows, NEs become much less frequent; the further the syntactic separation between a NE and the nearest code-switch, the less likely they are to be directly related. Finer-grained analysis of NEs by type also supports this (Table 6 in Appendix A.1): most types of NEs are separated from code-switched segments by dependency distances under 2.0. Overall, we find that most CSW interactions involving NEs tend to occur with close syntactic proximity, with a clear tendency for these interactions to diminish as distance increases, pointing to the influence of NEs on CSW in speech.

Finally, we compare NE production in insertionally code-switched utterances to that in alternationally code-switched ones. Odds ratio calculations show that the former are twice as likely to contain NEs than the latter. These results replicate when considering each type of NE independently and align with intuition on how speakers typically incorporate NEs in code-switched utterances in an insertional fashion, signaling the relatively low syntactic complexity of code-switches involving NEs.

In sum, when restricting our analysis of NE production to a code-switched context, we find evidence of NEs playing a role in shaping certain aspects of spoken CSW in BM. Though rare overall, when NEs are present in code-switched BM utterances, these appear to have subtle relationships with language choice and the syntax of code-switched segments in BM.

4.2 How is CSW influenced by its intended function?

Dialogue acts differ between code-switched and monolingual contexts. We now study DAs across monolingual and code-switched utterances, to determine how the expression of communicative intent relates to linguistic context. Unlike NEs, each utterance in the corpus has an intended discourse function, and thus an associated DA. These appear generally similar in occurrence by DA type across code-switched and monolingual contexts (Figure 4). The most commonly used DAs in the BM corpus, in both linguistic contexts, are statements of opinion (sv-fx; e.g. "oh no, *qué estúpida.*") and non-opinion (sd; e.g. "*creo que* it's

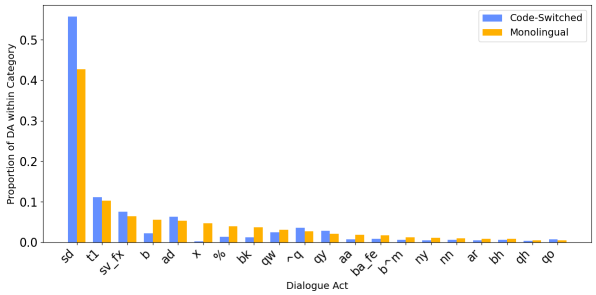

Figure 4: Distribution of dialogue act types across code-switched and monolingual contexts. Proportions are calculated within each context type.

about 350 square feet.") and self-talk (t1; e.g. "*a ver*, let's see..."). It seems like distinct discourse functions may be independent of bilingual mode of expression; we perform statistical tests on the ten most frequent DAs in the corpus to verify this hypothesis. Chi-squared tests with odds ratio calculations reveal that utterances expressing half of the top ten discourse functions, including acknowledgment responses and backchannels (bk, b; e.g. "uh huh"), questions (qw; e.g. "who told you to do so?"), and non-verbal acts such as laughter (x, %), are generally less likely to contain CSW (Table 2). However, the three most frequent DAs in the corpus follow the opposite trend: the odds that an utterance functions as a statement or self-directed comment, given that it is code-switched, are about 1.3 times as high as those for a monolingual utterance. This indicates that discourse functions conveying new information are significantly better expressed in a multilingual fashion, while those that advance a conversation are better expressed monolingually.

DA Type	χ^2	*p*-val.	OR	95% CI
sd	119.2	**	1.59	[1.46, 1.73]
t1	1.13	–	1.08	[0.94, 1.23]
sv-fx	2.08	–	1.13	[0.96, 1.32]
b	39.1	**	0.41	[0.30, 0.55]
ad	1.86	–	1.13	[0.95, 1.35]
x	28.3	**	0.11	[0.03, 0.28]
%	38.7	**	0.34	[0.23, 0.49]
bk	35.8	**	0.33	[0.22, 0.48]
qw	4.48	*	0.74	[0.55, 0.97]
^q	4.81	*	1.30	[0.99, 1.68]

Table 2: Summary of chi-squared tests and odds ratios comparing DA distribution in monolingual and code-switched utterances. p-values less than 0.05 and 0.01 are denoted by * and **, respectively. p-values greater than 0.05 are denoted by –.

To further investigate differences in discourse functions between monolingual and code-switched utterances, we use `scikit-learn 1.6.1` to perform k-means clustering on one-hot encoded DAs, reducing dimensionality using principal component analysis. We set $k = 3$; motivating this design choice is the need for an appropriate k-value, given the moderate dataset size, that can capture meaningful structure without being too noisy or overfitting.[10] The resulting monolingual and code-switched clusters appear similar (Figure 5 in Appendix A.2), reflecting the general distribution patterns we saw previously.

A detailed inspection shows one code-switched cluster (0 in Table 3 – left) is dominated by statements of non-opinion (sd), quotations (^q; e.g. "*mi mamá dijo que* she didn't know."), and yes-or-no questions (qy; e.g. "*entonces,* does she work?"), which mirrors our earlier findings on the DAs that are more likely to be expressed multilingually than monolingually. The primary presence of these DAs in a code-switched context suggests goal-driven and structured bilingual exchanges that focus on conveying factual information, providing explanations, or confirming understanding. CSW may thus align more closely with pragmatic and task-oriented purposes, where clarity and structure are prioritized. The other code-switched clusters (1 and 2 in Table 3 – left) are each entirely composed of a single DA: rhetorical questions (qh; e.g. "yeah but *tú sabes*?") and self-talk (t1), respectively. Cluster 1 is particularly striking as it captures a niche conversational strategy that is poorly represented in BM, possibly highlighting the role of stylistic or rhetorical emphasis in code-switched contexts.

Among the monolingual clusters, Cluster 0 (Table 3 – right) overlaps on two of three DAs represented in its code-switched counterpart. The third DA in its composition represents statements of opinion (sv-fx); the addition of this more exploratory function suggests greater conversational flexibility in monolingual contexts. Monolingual conversations may support broader thematic exploration, in contrast to code-switched ones which maintain more focused and structured discourse goals. The other monolingual clusters are each entirely composed of a single DA, too, representing acknowledgements and backchannels (b) and self-talk (t1). The inclusion of the former suggests greater collaborative dialogue with close dyadic interaction in monolingual utterances. We note that

[10] We also conduct sensitivity analyses on other k-values to justify using $k = 3$; see Appendix A.2 for details.

the latter DA is also included in the code-switched clusters, indicating that introspective utterances may be linguistically universal and independent of multilingual conversational dynamics.

DA	Cluster #		
	0	1	2
sd	**0.52**	0	0
^q	**0.18**	0	0
qy	**0.03**	0	0
qh	0	**1**	0
t1	0	0	**1**

DA	Cluster #		
	0	1	2
sd	**0.40**	0	0
^q	**0.15**	0	0
sv-fx	**0.07**	0	0
t1	0	0	**1**
b	0	**1**	0

Table 3: Cluster composition by proportion: code-switched (left) and monolingual (right) utterances.

Overall, we find subtle but important differences in DAs and associated discourse functions between code-switched and monolingual utterances, suggesting that certain conversational functions have preferred modes of expression in this corpus. CSW in BM leans toward task-oriented, structured exchanges, suggesting a focus on clarity and goal completion, while monolingual speech is more varied, incorporating feedback and exploratory acts.

Dialogue acts vary with patterns of CSW. We now explore the interaction between DAs and various patterns of bilingual speech within the code-switched subset of BM. First, we explore a potential relationship between DAs and CSW language direction. We find that the ten most frequent DAs in BM are about equally likely to be code-switched from English to Spanish as from Spanish to English, and none of the supporting chi-squared test results are statistically significant. So, our results fail to support a relationship between multilingual discourse function and CSW language direction.

Next, we consider the richness of utterance-level CSW in relation to DAs. One-way ANOVA tests show significant differences in quantity of CSW produced between DAs with differing functions. This is true for both metrics of CSW quantity, and is more notable for CSW ratio ($F = 6.45$, $p < 0.01$) than for M-index ($F = 2.27$, $p = 0.01$). Acknowledging responses (bk), backchannels (b), and action directives (ad; e.g. "okay, *ponlo ahí.*") seem to require greater quantities of CSW (mean CSW ratio: 0.26, SD CSW ratio: 0.14), compared to statements (sd, sv-fx), self-talk (t1), questions (qw), and quotations (^q) (mean CSW ratio: 0.19, SD CSW ratio: 0.11). The former group roughly corresponds to the set of DAs that were more likely to be expressed in monolingual utterances, while the latter group corresponds to those that were more likely to contain CSW. Given this prior result, it is striking that the former group exhibits greater CSW richness than the latter, suggesting a potential compensatory multilingual mechanism at play.

Finally, we examine how CSW strategy relates to DAs. The distribution of multilingual DAs expressed via syntactically simpler and shorter utterances (mean: 10.1 tokens, SD: 6.8 tokens) of insertional CSW is notably different from that expressed via relatively complex and longer utterances (mean: 14.9 tokens, SD: 6.3 tokens) of alternational CSW ($p < 0.001$). For example, self-talk (t1) is 9.6 times more likely to be produced as insertional CSW than alternational CSW, given that the DA is code-switched. This pattern is especially significant for English-predominant code-switches involving Spanish insertions, e.g. "*pues,* what was I saying...". Statements of opinion (sv-fx) are also more likely to be insertionally code-switched, with an odds ratio of 2.3 times, but more prominently so in Spanish-predominant code-switches with English insertions, e.g. "*para mí, es igual que cuando uno manda los* checks *al* IRS.". These results suggest that introspective and information-conveying functions of dialogue are more effectively expressed through simple code-switches. In contrast, quotations (^q) are always alternationally code-switched, and wh-questions (qw) are twice as likely to be alternationally than insertionally code-switched, reflecting the syntactic structure of how such discourse acts tend to be expressed in separate but connected complete grammatical clauses, e.g. "*cuándo vas a ver el apartamento* and how many bedrooms does it have?". This is striking given the lower representation of this CSW strategy in the corpus overall. These code-switched DAs exemplify how speech structure informs function, and vice versa.

Overall, when considering only bilingual BM utterances, we find compelling indications of a relationship between DAs and multiple aspects of code-switched speech production. Expression of specific discourse functions supports CSW of varying quantity, syntactic structure, and complexity.

4.3 Can transition models leverage discourse patterns to improve CSW predictions?

Given our statistical findings on notable interactions between each of discourse content and function with aspects of CSW behavior, we explore whether incorporating discourse information improves local and global transition modeling of code-

Base model	Metric	Baseline	Baseline + NE	Baseline + DA	Baseline + NE + DA
Logistic regression	Accuracy	0.633	0.613	0.640	0.638
	F1: ML	0.765	0.752	0.761	0.762
	F1: CSW	0.122	0.107	0.122	0.122
	p-val.	–	0.050	0.0001	0.020
HMM	Accuracy	0.990	0.984	0.991	0.984
	F1: ML	0.995	0.990	0.995	0.990
	F1: CSW	0.909	0.860	0.909	0.860
	p-val.	–	0.001	0.002	0.001

Table 4: Transition model performance across ablation settings. Features in use include a baseline feature set, NE features, and DA labels, effectively isolating the marginal contribution of discourse-pragmatic and referential cues when predicting forthcoming CSW. We report overall model accuracy and F1 scores on the monolingual (ML) and code-switching (CSW) subsets of the corpus, along with p-values from z-tests of proportions relative to the baseline.

switched dialogue. Specifically, we train two transition models to predict whether the next utterance is code-switched, given preceding conversational discourse context and varying access to content and function information. We are also interested in how model behavior across such settings might provide insight into potential interactions between discourse content and function in multilingual speech.

A comparison of both models' performance across ablation settings in which we use a set of baseline features,[11] NE features as described in Section 4.1, and DA labels reveals that discourse *content* information generally *worsens* transition model performance relative to the baseline (Table 4). In contrast, the inclusion of discourse *function* information as a feature *improves* upon the baseline in both cases. Combining feature sets also leads to the logistic regression-based model outperforming the baseline, whereas DA labels are unable to compensate for the detrimental presence of NE features in the combined setting of the hidden Markov model-based one. This suggests that the potential interactive influence of discourse content and function on CSW may be more complex than anticipated, potentially depending on local vs. global conversational dynamics. We perform z-tests of proportions on the test samples in each ablation setting to compare model accuracy to the baseline, all of which yield statistically significant p-values. This shows that although changes in model performance across ablation settings are small in absolute value, these are nonetheless meaningful.

Overall, these performance results indicate that the statistical relationships we have found between NEs, DAs, and CSW in BM are strong and salient enough to be leveraged by models that output predictions for a related task. The direction of change in performance under varying feature settings points to a weaker overall relationship between CSW and discourse content as encoded by NEs, compared to that between CSW and discourse function as encoded by DAs. This generally agrees with and lends validity to our statistical findings, and demonstrates their value in not only understanding but also predicting CSW. Further statistical modeling is required to disentangle the individual contributions of content and function towards influencing CSW in tandem in this corpus.

[11] These are: count of personal pronouns, fillers, affirmative cues, and high-frequency words; CSW ratio; M-index; I-index; CSW strategy; utterance length; language-predominance indicator; and speaker gender. We select these for their documented predictive value in prior work (Bullock et al., 2018).

5 Conclusion

We extensively analyze the relationships between discourse content and function and spontaneous CSW in the BM corpus. We find that (1) CSW patterns are somewhat influenced by discourse content expressed via NEs; (2) discourse function encoded by DAs has a relatively greater influence on CSW patterns; (3) the statistical relationships discovered in (1) and (2) are salient enough to be learned and applied by transition models that predict CSW, which (4) additionally point to the two discourse aspects interacting in modeling contexts, though further work is required to uncover their specific joint effect on CSW. Our novel discourse-centric exploration of the pragmatic and referential cues of multilingual speech enables us to conclude that discourse-structuring aspects contribute importantly towards shaping spoken Spanish-English CSW. We hope this work will serve as a first step towards building improved models of CSW prediction and informing the authentic generation of discourse-motivated bilingual speech.

Limitations

Our work focuses on a single language pair in a single corpus of CSW, which is somewhat skewed towards English relative to Spanish. Both languages are also represented only in the forms in which they are typically spoken in Miami, Florida, in the United States. 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 are very interested in ultimately replicating our analyses on other CSW datasets, and we plan to examine whether our findings generalize in other code-switched settings in future work. Due to lack of access to CSW datasets, particularly those containing highly time-intensive manual discourse-level annotations, our work makes use of the best currently available resources and serves as a reasonable first step towards understanding the role of discourse content and function on code-switched speech production. We hypothesize that the greatest overlap in findings between the present work and future studies of other language pairs might be found in CSW between other dialects of Spanish and English, followed by X-English CSW, where X is a language that is code-switched in similar cultural contexts to Spanish-English in the United States and is typologically similar to Spanish. We plan to carry out such investigation in the future.

Relatedly, in selecting volunteer annotators to label the BM corpus, we chose a native speaker of both Spanish and English (the second author; Annotator A) in order to produce the most reliable and context-appropriate labels of DAs, which are heavily dependent on cultural context and understanding. We relaxed this requirement for the annotation of NEs, which require less contextual understanding to identify, and chose a fluent, but not native-level, speaker of both languages (Annotator B).[12] We recognize that this could have led to slight inconsistencies in the precision of annotation, but we believe that any such effect should be negligible due to the inherent differences in the two labeling tasks, which effectively account for varying levels of annotator linguistic proficiency.

We also acknowledge that having a single annotator for each task was not ideal. The size of the corpus and lack of availability of suitable expert annotators restricted our gathering of further labels on which to assess corpus-level inter-annotator agreement. We attempted to offset this drawback in two ways. First, a fluent Spanish-English speaker other than Annotators A and B (the first author; Annotator C) conducted spot checks of a small proportion of each set of labels to ensure that assigned labels were reasonable. Annotator C then provided an additional set of NE type annotations for a randomly-selected 10% sample of these labels (=457 instances), and an additional set of DA annotations for a randomly-selected 5% sample of these labels (=2345 instances). Agreement on these additional labels was $\kappa = 0.81$ and $\kappa = 0.69$, respectively. While small samples are certainly not ideal, we believe that these substantial agreements help to validate the quality of the corpus overall, and partially offset data reliability concerns.

Separately, in our analysis of dependency distance between NEs and CSW, we relied on automatic translation to English. First, while such translations may not be perfect, we believe this was still a reasonable choice, given the lack of currently available methods for extracting dependency parses from multilingual utterances, and the general reliability of dependency parsing models for English, relative to Spanish. Second, since English follows a fixed SVO word order, which is also generally the norm in Spanish, we believe that any impact on dependency distance calculations arising from translation is minimal. Based on a manual inspection of a small subset of the parse trees, when translation does have an effect on such calculations, these are primarily related to differences in the position of noun modifiers, subject/verb inversion in questions, and the optional presence of subject pronouns in Spanish, which would effectively result in dependency distances being off by 1 in each case. Thus, our dependency distance calculations may be inflated due to translation, implying that true dependency distance values should be smaller than what we report; this would provide even stronger support for our claim that NEs' interaction with code-switches depends on close proximity. Not only is the impact of translation on this part of the analysis minimal, it also serves as a stronger test for the validity of our finding. We plan to explore alternatives to this method that would easily extend to other CSW language pairs in future work.

Regarding our analysis of DAs, while many alternative taxonomies of discourse function and/or

[12] Annotator B was a local high school student who reached out to the last author to volunteer for the labeling task, and was briefed on how the annotations would be used prior to starting.

speech acts exist in the literature, we chose to use the unified set of SwDA labels as these can be generalized to any corpus. Using these tags also makes it easier to compare our study to the vast quantity of work on monolingual speech that has already been done using the same framework of DAs.

We only briefly consider the interactive effect between discourse content and function in Section 4.3 of our work, which yields interesting preliminary results. We plan to follow up on this interaction in more detail in future work, potentially experimenting with complex interaction terms and/or multimodal architectures to capture synergistic effects between content and function.

Finally, while there is some debate around the extent to which we can uncover bilinguals' motivation for CSW, we believe our work provides consistent empirical evidence supporting insight into the specific ways in which discourse aspects have the potential to influence various patterns of code-switched language production.

Ethics Statement

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

Acknowledgments

We thank Nicholas Deas and Ziwei Gong for helpful discussions and feedback, and Margot Story for providing NE annotations. This work was supported in part by the National Science Foundation under Grant IIS 2418307.

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan W Black. 2020. What Code-Switching Strategies are Effective in Dialog Systems? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 254–264, New York, New York. Association for Computational Linguistics.

Inas Almusallam. 2024. Code-Switching in Speech Acts: A Focus on Offer Interactions by Saudi EFL Female Bilinguals. , 25:55–83.

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

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

Peter Auer. 2007. *The pragmatics of code-switching: A sequential approach*, pages 123–138. Routledge.

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).

Ritu Belani and Jeffrey Flanigan. 2023. Automatic Identification of Code-Switching Functions in Speech Transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7438–7448, Toronto, Canada. Association for Computational Linguistics.

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.

Debasmita Bhattacharya, Aanya Tolat, and Julia Hirschberg. 2025. From Context to Code-switching: Examining the Interplay of Language Proficiency and Multilingualism in Speech. In *Interspeech 2025*, pages 4528–4532.

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

D. Bullock, Ö. Çetinoğlu, and S. Kübler. 2018. Should code-switching models be asymmetric? In *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech 2018)*, pages 3433–3437. Accessed: 2025-04-24.

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

Margaret Deuchar. 2011. Miami corpus: Preliminary documentation - bangortalk.

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.

Nathan Duran and Steve Battle. 2018. Probabilistic Word Association for Dialogue Act Classification with Recurrent Neural Networks. In *Engineering Applications of Neural Networks*, pages 229–239, Cham. Springer International Publishing.

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

Katerina A. Finnis. 2014. Variation within a Greek-Cypriot community of practice in London: Code-switching, gender, and identity. *Language in Society*, 43(3):287–310.

Soledad Galli. 2023. SMOTE in Python: A guide to balanced datasets. Train in Data's Blog. Accessed: 2025-04-24.

Penelope Gardner-Chloros and Malcolm Edwards. 2004. Assumptions Behind Grammatical Approaches To Code-Switching: When The Blueprint Is A Red Herring. *Transactions of the Philological Society*, 102(1):103–129.

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.

Silvana Hartmann, Monojit Choudhury, and Kalika Bali. 2018. An Integrated Representation of Linguistic and Social Functions of Code-Switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Aravind K. Joshi. 1982. Processing of Sentences with Intra-Sentential Code-Switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Daniel Jurafsky and James H. Martin. 2007. Hidden Markov and Maximum Entropy Models (Chapter 6, Draft). Draft chapter from Speech and Language Processing. Accessed: 2025-04-24. Draft chapter, likely related to the 2nd Edition of their book.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.

Atia Pia Khan and Imran Siddiqi. 2020. W-Trans: A Weighted Transition Matrix Learning Algorithm for the Sensor-Based Human Activity Recognition. In *2020 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 769–775. IEEE. Accessed: 2025-04-24.

Grit Liebscher and Jennifer Dailey-O'Cain. 2005. Learner Code-Switching in the Content-Based Foreign Language Classroom. *The Modern Language Journal*, 89:234 – 247.

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.

Jeff MacSwan. 2000. The architecture of the bilingual language faculty: evidence from intrasentential code switching. *Bilingualism: Language and Cognition*, 3(1):37–54.

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

Carol Myers-Scotton. 1993. *Duelling Languages*. Oxford: Clarendon Press.

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding Linguistic Accommodation in Code-Switched Human-Machine Dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 565–577, Online. Association for Computational Linguistics.

Shana Poplack. 1978. *Syntactic Structure and Social Function of Code-switching*. Centro working papers. Centro de Estudios Puertorriqueños, [City University of New York].

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.

Iliana Reyes. 2004. Functions of Code Switching in Schoolchildren's Conversations. *Bilingual Research Journal*, 28(1):77–98.

Thamar Solorio. 2008. *A Computational Study in the Detection of English–Spanish Code-Switching*. Ph.d. dissertation, The Graduate Center, City University of New York. Accessed: 2025-04-24.

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.

Igor Sterner. 2024. Multilingual Identification of English Code-Switching. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 163–173, Mexico City, Mexico. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Christopher Stroud. 1992. The problem of intention and meaning in code-switching. *Text - Interdisciplinary Journal for the Study of Discourse*, 12(1):127–155.

Chara Tsoukala, Stefan L. Frank, Antal van den Bosch, Jorge Valdés Kroff, and Mirjam Broersma. 2019. Simulating Spanish-English Code-Switching: El Modelo Está Generating Code-Switches. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 20–29, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Chenxi Whitehouse, Fenia Christopoulou, and Ignacio Iacobacci. 2022. EntityCS: Improving Zero-Shot Cross-lingual Transfer with Entity-Centric Code Switching. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6698–6714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

A.1 Named entities

How do we determine the language of a NE? With respect to names of people and brands or products, there may be some ambiguity regarding the language in which such NEs are produced. Pronunciation helps to distinguish the language in which names of people are uttered (consider the "r" consonant in "Maria" as pronounced by a Spanish speaker in a Spanish sentence compared to the way it would be pronounced by an American-English speaker in an English sentence). These NE annotations were done by listening to the corresponding speech recording to make a labeling decision in the event of ambiguities. For brand names, certain named entities can only be produced in a single language, regardless of pronunciation, e.g. references to the retailer "Target" (English) or the coffee product "Cafe Bustelo" (Spanish). This knowledge was applied during annotation. In the event of ambiguities, Annotator B used a combination of pronunciation information, transcript context, and cultural knowledge to attempt to determine the language of the NE. In cases where such determination was not possible, NEs were marked as language-ambiguous (Figure 2). While reasonable, we acknowledge that this method may not have been perfect, and may leave room for potential improvement in future work.

Entity Type	Mean token distance
P	2.22
N	2.30
T	2.00
O	3.00
U	1.20
B	1.00
R	3.50

Table 5: Mean token distance to the next code-switched segment of NEs within code-switched utterances.

Entity Type	Mean	Median	Std. Dev.
P	1.77	1.0	1.65
N	1.92	2.0	1.90
T	2.16	2.0	2.01
O	1.07	0.0	1.39
U	1.80	2.0	1.82
B	1.00	1.0	1.41
R	5.42	5.5	1.51

Table 6: Dependency distances across NE types.

A.2 Dialogue acts

DA	Label
Statement-non-opinion	sd
Acknowledge (Backchannel)	b
Statement-opinion	sv-fx
Agree/Accept	aa
Abandoned | Turn-Exit | Uninterpretable	%
Appreciation	ba-fe
Yes-No-Question	qy
Non-verbal	x
Yes answers	ny
Conventional-closing	fc
Uninterpretable	%
Wh-Question	qw
No answers	nn
Response Acknowledgement	bk
Hedge	h
Declarative Yes-No-Question	qy^d
Other	fo-o-fw-by-bc
Backchannel in question form	bh
Quotation	^q
Summarize/reformulate	bf
Affirmative non-yes answers	na
Action-directive	ad
Collaborative Completion	^2
Repeat-phrase	b^m
Open-Question	qo
Rhetorical-Questions	qh
Hold before answer/agreement	^h
Reject	ar
Negative non-no answers	ng
Signal-non-understanding	br
Other answers	no
Conventional-opening	fp
Or-Clause	qrr
Dispreferred answers	arp-nd
3rd-party-talk	t3
Offers, Options, Commits	oo-co-cc
Self-talk	t1
Downplayer	bd
Maybe/Accept-part	aap-am
Tag-Question	^g
Declarative Wh-Question	qw^d
Apology	fa
Thanking	ft

Table 7: Set of DAs taken from SwDA.

Sensitivity analysis for clustering with additional values of k. We replicate our clustering analysis using $k = 2$ (Table 8), $k = 4$ (Table 9), $k = 5$ (Table 10), and $k = 10$ (Table 11), and find that the same patterns reported for $k = 3$ (Table 3) emerge in each case. For $k = 4$, $k = 5$, and $k = 10$ in particular, additional patterns emerge that support our current conclusions: CSW clusters are increasingly dominated by action directives and statements of agreement/acceptance, which are key to task-oriented conversations, while monolingual clusters are increasingly dominated by no- and other-answers, which are essential to providing feedback in more exploratory dialogue. While

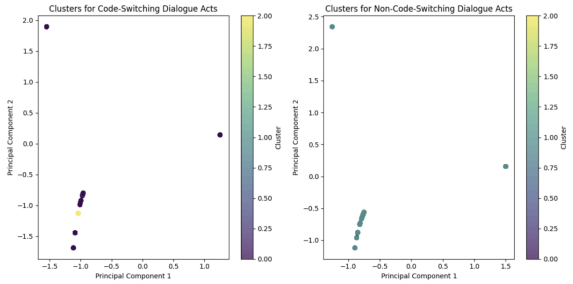

Figure 5: PCA scatterplots for code-switched (left) and monolingual (right) clusters for $k = 3$.

other dialogue acts are present in clusters associated with these k-values, these are in very small proportions in each case, and thus do not contribute much to the overall cluster composition or its interpretation. These results showcase that increasing k-values keeps cluster patterns generally consistent, with greater values allowing for additional patterns to emerge. We do note that fewer additional patterns emerge between $k = 5$ and $k = 10$ compared to those that emerge between $k = 3$, $k = 4$, and $k = 5$, suggesting that greater k-values are associated with diminishing returns in terms of further insights gained.

DA	Cluster #	
	0	1
sd	0	**0.57**
^q	**1**	0
qy	0	**0.03**
t1	0	**0.11**
ad	0	**0.06**

DA	Cluster #	
	0	1
sd	0	**0.44**
^q	0	**0.03**
sv-fx	0	**0.06**
t1	0	**0.10**
nn	**1**	0

Table 8: Cluster composition by proportion for $k = 2$: code-switched (left) and monolingual (right) utterances.

DA	Cluster #			
	0	1	2	3
sd	0	0	0	**0.60**
^q	**1**	0	0	0
qy	0	0	0	**0.03**
t1	0	0	0	**0.12**
aa	0	**1**	0	0
qw	0	0	**1**	0
ad	0	0	0	**0.07**

DA	Cluster #			
	0	1	2	3
sd	0	0	0	**0.44**
^q	0	0	**1**	0
sv-fx	0	0	0	**0.07**
t1	0	0	0	**0.10**
b	0	0	0	**0.06**
nn	**1**	0	0	0
no	0	**1**	0	0

Table 9: Cluster composition by proportion for $k = 4$: code-switched (left) and monolingual (right) utterances.

DA	Cluster #				
	0	1	2	3	4
sd	0	0	0	0	**0.61**
^q	0	**1**	0	0	0
qy	0	0	0	0	**0.03**
t1	0	0	0	0	**0.12**
aa	**1**	0	0	0	0
qw	0	0	0	**1**	0
ad	0	0	**1**	0	**0.07**

DA	Cluster #				
	0	1	2	3	4
sd	0	0	0	0	**0.49**
^q	0	0	0	**1**	**0.03**
sv-fx	0	0	0	0	**0.07**
t1	0	0	**1**	0	0
b	0	0	0	0	**0.06**
nn	**1**	0	0	0	0
no	0	**1**	0	0	0

Table 10: Cluster composition by proportion for $k = 5$: code-switched (left) and monolingual (right) utterances.

DA	Cluster #									
	0	1	2	3	4	5	6	7	8	9
sd	0	0	0	0	**1**	0	0	0	0	**0.48**
^q	0	**1**	0	0	0	0	0	0	0	**0.12**
qy	0	0	0	0	0	**1**	0	0	**0.21**	0
qh	0	0	0	0	0	0	0	**1**	**0.62**	0
t1	0	0	0	0	0	0	**1**	0	0	0
aa	**1**	0	0	0	0	0	0	0	0	**0.03**
qw	0	0	0	**1**	0	0	0	0	**0.08**	0
ad	0	0	**1**	0	0	0	0	0	0	**0.02**

DA	Cluster #									
	0	1	2	3	4	5	6	7	8	9
sd	0	0	0	0	**1**	0	0	**0.45**	0	0
^q	0	0	0	**1**	0	0	0	**0.11**	0	0
sv-fx	0	0	0	0	0	**1**	0	**0.05**	0	0
t1	0	0	**1**	0	0	0	0	0	0	**0.17**
b	0	0	0	0	0	0	**1**	0	0	**0.62**
nn	**1**	0	0	0	0	0	0	0	**0.43**	0
no	0	**1**	0	0	0	0	0	0	**0.02**	**0.06**
ar	0	0	0	0	0	0	0	0	**0.13**	0

Table 11: Cluster composition by proportion for $k = 10$: code-switched (left) and monolingual (right) utterances.

A.3 Predictive models

Why logistic regression? We select logistic regression with default hyperparameter values[13] as the base framework for one of our predictive transition models due to its robustness and interpretability in classification tasks, particularly in language identification and other CSW scenarios (Solorio, 2008). We use it at the utterance level, where the model treats each utterance as an independent data point and asks: given the linguistic evidence available now, how likely is the upcoming stretch of speech to be code-switched? In doing so, we can effectively test for micro-level, turn-by-turn predictions where only the current utterance's data is taken into account. In other words, the logistic regression transition model focuses on local cues embedded within a single utterance.

Why hidden Markov model? We choose a hidden Markov model with default hyperparameter values[14] as the base framework for our other predictive transition model because of its ability to preserve intra-dialogue temporal dependencies, which are crucial for understanding sequential CSW behavior. This model operates at the dialogue level, and is expressly trained to model temporal dependencies across successive utterances within an entire conversation. The hidden states (monolingual vs. code-switched) learn based on the transition probabilities in utterance turns, which allows the model to learn relevant patterns, e.g. "once a speaker code-switches, the next few utterances are also more likely to be code-switched." Thus, the HMM transition model captures grouped utterance dynamics.

HMM reweighting scheme. We collect transition counts from the labeled training sequences and normalize these to yield the unweighted transition matrix (T_{orig}) and its weighted counterpart (T_{w}). These matrices show that monolingual-to-CSW transitions are rare in the raw data and are up-weighted in the modified model to encourage exploration of minority paths during Viterbi decoding. Emission parameters are adjusted analogously: observations attributed to the code-switched state are up-weighted to prevent their Gaussian statistics from being dominated by abundant monolingual data (Jurafsky and Martin, 2007).

T_{orig}

0.953	0.047
0.876	0.124

T_{w}

0.802	0.198
0.876	0.124

[13] Except an increased number of max. iterations (100 → 1000) and a random state setting (42) for reproducibility.

[14] Note that the supervised two-state HMM has few tunable hyperparameters and we rely on domain-informed initial probabilities [0.5,0.5] together with our previously described weighting scheme. Given this, we do not employ an extra validation fold. Each state uses Gaussian emissions with diagonal covariance.