Automatic Identification of Code-Switching Functions in Speech Transcripts

Ritu Belani
The Harker School
ritubelani@gmail.com

Jeffrey Flanigan
University of California, Santa Cruz
jmflanig@ucsc.edu

Abstract

Code-switching, or switching between languages, occurs for many reasons and has important linguistic, sociological, and cultural implications. Multilingual speakers code-switch for a variety of communicative functions, such as expressing emotions, borrowing terms, making jokes, introducing a new topic, etc. The function of code-switching may be quite useful for the analysis of linguists, cognitive scientists, speech therapists, and others, but is not readily apparent. To remedy this situation, we annotate and release a new dataset of functions of code-switching in Spanish-English. We build the first system (to our knowledge) to automatically identify a wide range of functions for which speakers code-switch in everyday speech, achieving an accuracy of 75% across all functions.

1 Introduction

Code-switching, or switching between languages within the same utterance or sentence (Poplack, 1980), commonly emerges in conversations between multilinguals and in written communication such as social media. In today's intersecting multilingual world, it is essential to develop computational tools that can process and analyze code-switched speech and text.

In recent years, there has been much progress in processing code-switched language. Many code-switched datasets have been collected for a diverse set of natural language processing tasks such as sentiment analysis, NER, conversational systems, and many others (Sitaram et al., 2019). Workshops held on computational approaches to code-switching have created shared tasks on language identification (Solorio et al., 2014) and Named Entity Recognition (NER) (Aguilar et al., 2019) in code-switched texts. Nuanced tasks like humor detection, sarcasm detection, and hate detection have been applied to Hindi-English code-switched data (Bansal et al., 2020).

Despite these achievements, there is relatively little work on identifying the functions of code-switching. Although there are annotation schemes (Zentella, 1998; Hartmann et al., 2018) and some annotated datasets (Dey and Fung, 2014; Begum et al., 2016; Lee and Wang, 2015; Rudra et al., 2019), to our knowledge, there is no work automatically identifying the communicative function of a code-switch across a full range of qualities (Zentella, 1998).

There are many potential applications for the task proposed in this paper, including improved cognitive models of bilingual processing, diagnosis of language disorders, and improved understanding of social factors of group membership and microaggressions. Code-switching analysis contributes to the development of cognitive models for bilingual language processing and production (Macnamara and Kushnir, 1971; Kecskes, 2006; Phillips and Pylkkänen, 2021; Kheder and Kaan, 2021). Understanding the functions of code-switching is critical for speech-language pathologists interacting with bilingual children, so as not to mistakenly diagnose them with a language disorder when in reality, children are taking advantage of a wide range of communicative strategies by code-switching (Miccio et al., 2009; De la Rosa, 2022). Studying code-switching in people with dementia and Alzheimer's disease can provide insights into language impairments experienced as their condition becomes more severe (Santi et al., 1990; Friedland, 1998; Svennevig et al., 2019). Code-switching is also important for pragmatics research of understanding social identities and group membership that speakers are trying to assert (Auer, 2005; Cashman, 2005). Because of political undertones of using one language over another (Heller, 1992), code-switching is useful for understanding linguistic microaggressions (Anchimbe, 2015; Takeuchi, 2022).

Our contributions are the following:

- An annotation scheme identifying the func-

tion of code-switching with 11 different labels, encompassing emotional, situational, and semantic functions of code-switching

- A new dataset applying this annotation scheme to code-switched utterances in the Spanish-English Bangor Miami Corpus (Deuchar, 2010)

- Trained models and experiments with XLM-RoBERTa (Conneau et al., 2019) and a baseline Naive Bayes model, demonstrating the feasibility of the proposed task

2 Related Work

2.1 Code-Switched Data Annotation

Several studies have annotated code-switched data according to their own frameworks (Lee and Wang, 2015; Begum et al., 2016; Hartmann et al., 2018; Rudra et al., 2019). Rudra et al. (2016) developed classifiers to determine whether Hindi-English code-switching on Twitter was opinionated or not and found that audiences preferred to use Hindi to express a negative sentiment and English to express a positive sentiment. Lee and Wang (2015) developed a system to identify the emotions in code-switched Chinese-English posts. Additionally, one corpus of Hindi-English code-switched conversations has broadly grouped the functions of code-switching in order to study the rules that govern code-switching (Dey and Fung, 2014). The framework we apply in this paper draws upon elements from Zentella (1998)'s framework, and it closely mirrors the approach of Begum et al. (2016). However, while their annotation scheme is based on Tweets, ours is specific to conversational code-switching. Linguists have also developed theoretical frameworks for code-switching without applying them to the systematic annotation of corpora (Poplack, 1980; Gumperz, 1982; Myers-Scotton, 1997; Zentella, 1998; Halim and Maros, 2014).

2.2 Code-Switching and Multilingual Language Models

Previous research has proven the success of fine-tuning the pre-trained models Multilingual BERT and XLM-RoBERTa for tasks such as offensive language identification (Jayanthi and Gupta, 2021) and named entity recognition and part-of-speech tagging (Winata et al., 2021) in code-switched texts.

Because of these models' state-of-the-art performance, we decided to fine-tune Multilingual BERT and XLM-RoBERTa on our tasks.

3 Annotation

We describe the data annotated, present our annotation scheme, and give a comparison of our annotation to previous annotation schemes.

3.1 Data

We annotate data from the Bangor Miami corpus (Deuchar, 2010), a publicly available anonymized code-switched Spanish-English conversational dataset consisting of audio recordings and human-created transcripts between two or more speakers. This dataset was selected for annotation because of its diverse examples of natural code-switching in spontaneous conversations, as opposed to datasets with synthetically manufactured examples of code-switching.[1] We filter the data from the transcripts for sentences with instances of code-switching and annotate the first 26 transcripts of the 56 total transcripts. The statistics of our filtered dataset are: number of utterances = 1,379; number of sentences = 7,547; words in Spanish = 15,796; words in English = 20,357; ambiguous words (both Spanish and English) = 3,393.

3.2 Annotation Scheme

We identify eleven labels in our annotation scheme as a mix of emotional, situational, and semantic functions of code-switching. Like Begum et al. (2016), we identify that a single code-switch could serve multiple functions because each code-switch can be seen as a sum of its semantic, structural, and sentiment-related dimensions. Thus, the labels are not mutually exclusive, and one code-switch can have multiple labels.

Change topic: code-switch to introduce another viewpoint, change the tone, or clarify something. Ex: I'm not ready at all, *¿y qué tal tú?* (I'm not ready at all, and what about you?)

Borrowing: a short word or phrase substitution in the other language, then returning to the original language. Ex: *Mi amiga de* high school *va a casarse en dos semanas.* (My friend from high school is going to get married in two weeks.)

Joke: code-switch for comedic effect or a sarcastic quip. Ex: You're making such a big deal

---

[1]Our dataset and code can be found at https://github.com/ritumb0/Automatic-Identification-Code-Switching.

about it, *como si murieran las personas en la calle.* (You're making such a big deal about it, as if people were dying in the street.)

Quote: code-switch to be true to how a statement was spoken by someone else. Ex: So my Spanish teacher said, "*Oye, necesitas estudiar más.*" (So my Spanish teacher said, "Hey, you need to study more.")

Translate: code-switch to repeat a statement or phrase, perhaps for the sake of emphasis or clarity. Ex: *A veces*, sometimes, I like to be by myself. (Sometimes, sometimes, I like to be myself.)

Command: code-switch for imperative or mandate intended to get the addressee to do something. Ex: *Él no sabe lo que está diciendo*, just don't listen to him. (He doesn't know what he's saying, just don't listen to him.)

Filler: a filler, brief interjection, or short noise intended to communicate meaning from the other language. Ex: *Y yo me callé,* you know, *porque no quería ofender a nadie.* (And I stopped talking, you know, because I didn't want to offend anybody.)

Exasperation: code-switch to complain or emphasize anger or frustration. Ex: *Ay, cómo me sigues molestando,* I should just get up and leave! (Oh, how you keep annoying me, I should just get up and leave!)

Happiness: code-switch to make a compliment or positive interjection. Ex: I just saw her dress, *¡qué lindo!* (I just saw her dress, how pretty!)

Proper noun: code-switch to talk about people or places whose names are in the other language or pronounced according to the other language. Ex: *Escogimos* United Airlines *porque ellos ofrecen las mejores meriendas.* (We chose United Airlines because they offer the best snacks.)

Surprise: code-switch to interject or relay that something was unexpected. Ex: *¿Qué hizo ella?* Oh my god. (What did she do? Oh my god.)

61.6% of the utterances in the dataset contain more than one type of code-switching. It is possible for an utterance to contain code-switching that does not fall under our scheme and therefore gets no label, but this does not occur in our dataset.

3.3 Comparison to Previous Annotation Schemes

Because of the broad range of domains to which our task and dataset can be applied, we choose to include a diverse set of tags to account for all the functions of code-switching we observe. Our categories quote, command, and translate are similar to categories in Begum et al. (2016) and Zentella (1998). However, we use courser-grained categories to expedite annotation and improve agreement. Our changing topic category is closely modeled after Zentella (1998)'s designation of Realignment, which includes a topic shift, rhetorical question, break from a narrative, aside comment, and checking with the listener. Begum et al. (2016) includes sarcasm and negative sentiment categories, which are subsets of our more expansive joke and exasperation categories. Fine-grained categories that Begum et al. (2016) include which we do not are the more fine-grained breakdowns of Narrative-Evaluative, Reinforcement, Cause-Effect, and Reported Speech. Table 4 in the appendix includes this comparison between annotation schemes in table form.

We include emotion categories for code switching, which are not included in Begum et al. (2016) and Zentella (1998), as we find this to be an important reason for code switching in dialogues. Lee and Wang (2015)'s annotation scheme for emotions in Chinese-English code-switching includes happiness, sadness, anger, fear, and surprise, three of which we share in our categories of happiness, exasperation, and surprise. We have included categories such as using a filler and expressing happiness, frustration, or surprise which we find occurs during a conversation in which someone is reacting to the statements made by the other person.

In a related annotation scheme, Dey and Fung (2014) establish a set of functions of code-switching among the speakers in their Hindi-English code-switching conversation corpus, which consists of Ease of Use, Comment, Referential Function, Topic Shift, Dispreference, Personalisation, Emphasis, No Substitute Word, Name Entity, and Clarification. However, they do not go in depth into their reasoning behind choosing these functions and offer little elaboration upon what each one entails.

A few of the functions that we identify have typically not been regarded as instances of code-switching, such as borrowing and proper nouns (Scotton and Ury, 1977). However, these features may still be of interest for downstream applications, so we include them here.

Label	Naive Bayes	mBERT	mBERT with adapter	XLM-R	XLM-R with adapter
Change topic	63.2	**86.3** ±1	85.7 ±1.7	**86.3** ±0.9	**86.3** ±0.4
Borrowing	57.3	**78.5** ±6.7	77.4 ±3.1	75 ±2.3	70.9 ±2.1
Joke	59.6	**79.8** ±13.6	37.0 ±28.0	68.5 ±15.6	68.7 ±9.8
Quote	40.9	**75.6** ±2.4	74.3 ±5.2	69.3 ±4.9	70.3 ±4.6
Translate	46.4	72.2 ±10.7	73.9 ±9.6	**74.6** ±17.6	74 ±10.5
Command	70.5	59.6 ±31	**74.3** ±8.2	66.4 ±20.6	66.2 ±7.1
Filler	57.8	70.5 ±3.2	72.2 ±5.3	73.4 ±2.5	**74.4** ±2.5
Exasperation	62.3	53.2 ±16.8	51.4 ±14.2	70.5 ±14.4	**77.1** ±8.7
Happiness	64.1	**83.6** ±6.1	80.2 ±8.7	78.4 ±4.3	70.5 ±6.3
Proper noun	61.0	84.5 ±3.3	85.4 ±1.6	**85.5** ±1.9	83.6 ±1.9
Surprise	68.2	75.0 ±4.9	66.4 ±3.9	**79.4** ±3.6	73.3 ±7.4
Average	59.2	74.4 ±2.8	70.7 ±5.2	**75.4** ±3.6	74.1 ±3.1

Table 1: Accuracy (in %) of label detection in code-switching dialogue. We report the standard deviation from training with 5 different random seeds.

3.4 Statistics and Inter-Annotator Agreement

In the annotated data, the frequency of some functions of code-switching over others validates theories about code-switching. For example, code-switching to change topics is regarded as the most frequent type of code-switching (Zentella, 1998), a trend which is present in Table 2. There are three filtered entries which contain markers that a code-switch is near, but are all spoken in one language, so they receive no label.

To compute inter-annotator agreement, a subset of 100 code-switched utterances was labeled by another annotator. The trained annotator was fluent in English and Spanish. After engaging in a presentation which included the same information as Section 3.2 and discussing five examples with the principal annotator, the trained annotator labeled 100 code-switched utterances independently. Because our dataset is multi-label, Cohen-Kappa is computed for each label as a binary classification task. The agreement scores are shown in Table 2 for each category.

4 Automatic Detection of the Code-Switching Functions

To demonstrate the feasibility of the proposed task, we fine-tune classifiers on our annotated corpus to predict labels for code-switching in our data. For the train/dev/test split, four conversations (16% of the annotated data, 220 code-switched utterances) are randomly set aside as test data, and the rest of the data is organized into a 75/25 train/dev split.

Label	Frequency	Cohen-Kappa Score
Change topic	65.0%	0.60
Borrowing	26.0%	0.45
Joke	3.7%	0.14
Quote	6.5%	0.52
Translate	5.9%	0.30
Command	8.3%	0.50
Filler	31.0 %	0.30
Exasperation	7.7%	0.23
Happiness	4.1%	0.54
Proper noun	25.9%	0.40
Surprise	11.6%	0.07

Table 2: Distribution of labels in the dataset (Frequency) and agreement between annotators (Cohen-Kappa Score).

Results show the most effective approach is by building unique classifiers for each label. Because over half of the labels appear in less than 10% of the data, we find that the classifiers always predict 0 for these labels if provided with all of the training data. Thus, we create balanced training datasets for each label so that half of the examples are an instance of the label, and the other half are not.

In addition to a baseline Naive Bayes classifier, we fine-tune bert-base-multilingual-cased (mBERT) and xlm-roberta-base (XLM-RoBERTa) classifiers using Huggingface.[2] Because of the rel-

---

[2]https://huggingface.co/models. mBERT base has 110M parameters, and XLM-RoBERTa base has 125M. We use the

Transcript	Gold	System
Original: MAR: That my children were being welcomed into the — *olvídate si tiene como tres* trainers! *Tiene un cocinero!*	Borrowing	Borrowing
Translation: MAR: That my children were being welcomed into the — forget it if he has like three trainers! <u>He has a chef!</u>		
Original: JES: *Invita a a alguna de las celebraciones.* NIC: I don't know. I have JES: *Tú sabes se caen bien.* NIC: Yeah I'll tell her. *Bueno* not her I gotta tell sister.	Filler	No Filler
Translation: JES: <u>Invite [her] to one of the celebrations.</u> NIC: I don't know. I have JES: <u>You know they like each other.</u> NIC: Yeah I'll tell her. <u>Well</u> not her I gotta tell sister.		
Original: IRI: *Ajá.* JAM: If if I happens to see like running blood or something like that I feel disgust and I feel weak and I IRI: My dad was just the same. *Sí, sí, sí, sí,* kryptonite. JAM: Kryptonite, yeah. IRI: *No mi pa mi papá era igual.*	Translate	Translate
Translation: IRI: <u>Uh huh.</u> JAM: If if I happens to see like running blood or something like that I feel disgust and I feel weak and I IRI: My dad was just the same. <u>Yes, yes, yes, yes,</u> kryptonite. JAM: Kryptonite, yeah. IRI: No my da- my dad was the same.		
Original: PAI: *En qué lo puedo ayudar?* SAR: He's going to the airport. PAI: What up. SAR: *Discúlpame.* PAI: It sounds like you're saying *escúpame.*	No Command	Command
Translation: PAI: <u>How can I help you?</u> SAR: He's going to the airport. PAI: What up. SAR: <u>Excuse me.</u> PAI: It sounds like you're saying spit me.		

Table 3: Sample system outputs on Spanish-English code-switched data with speaker IDs. We show gold and system outputs for only one label type. However, these examples may have additional labels.

atively small training set, to combat overfitting, we experiment with adapter layers for the two Transformers, but find that they do not perform as well. Training details and hyperparameters are in the appendix. We find the best model to be the XLM-RoBERTa model.

4.1 Results

The accuracy for each label with each model is shown in Table 1. Since the dataset is small, in order to quantify the statistical significance, we compute the mean accuracy of each model on each task and report the standard deviation across 5 training runs.

4.2 Qualitative Analysis of Results

In a qualitative analysis of the models' predictions, we observe that models are more likely to notice a borrowed word when it is surrounded by a longer string in the other language. In addition, when

there are multiple code-switching points, it is more difficult for models to identify the full range of functions. Example outputs are shown in Table 3.

5 Conclusion

This paper presents a corpus of Spanish and English code-switching with labels for the different functions for code-switching. We collect the data from the Bangor Miami corpus, create an annotation scheme for functions of code-switching, and annotate the data. We propose a classifier-based approach to detect the functions of code-switching in the annotated code-switching corpus. Results show that the XLM-RoBERTa model is the most effective at predicting functions of code-switching. We believe that analysis of functions of code-switching is an innovative approach towards bilingual speech diagnosis as well as contributing to a linguistic model of code-switching.

Google Colab Pro+ Tesla V100-SXM2-16GB GPU to train the models, and each model trains in less than 15 minutes.

6 Limitations

Our system has been trained on everyday conversations from Spanish-English bilinguals and may not be applicable to other domains. Additionally, the accuracy of the classifier varies depending on the label type. We use human-created transcripts, so results may not apply for automatic transcripts. There is a risk that incorrect conclusions can be drawn if the system does not meet the performance requirements.

Acknowledgements

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2019. Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. *arXiv preprint arXiv:1906.04138*.

Eric A Anchimbe. 2015. Code-switching: Between identity and exclusion. *Code-switching between structural and sociolinguistic perspectives*, page 139.

Peter Auer. 2005. A postscript: code-switching and social identity. *Journal of Pragmatics*, 37(3):403–410. Conversational Code-Switching.

Srijan Bansal, Vishal Garimella, Ayush Suhane, Jasabanta Patro, and Animesh Mukherjee. 2020. Code-switching patterns can be an effective route to improve performance of downstream nlp applications: A case study of humour, sarcasm and hate speech detection. *arXiv preprint arXiv:2005.02295*.

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.

Holly R. Cashman. 2005. Identities at play: language preference and group membership in bilingual talk in interaction. *Journal of Pragmatics*, 37(3):301–315. Conversational Code-Switching.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Ivonne Marie Maldonado De la Rosa. 2022. *Speech Language Pathologists' Approach to Code-Switching with Culturally Linguistically Diverse Clients*. Ph.D. thesis, Inter-American University of Puerto Rico (Puerto Rico).

Margaret Deuchar. 2010. Bilingbank spanish-english miami corpus.

Anik Dey and Pascale Fung. 2014. A hindi-english code-switching corpus. In *LREC*, pages 2410–2413.

Deborah Cecily Friedland. 1998. *Language loss in bilingual speakers with Alzheimer's disease*. Ph.D. thesis, Newcastle University.

John J Gumperz. 1982. *Discourse strategies*. 1. Cambridge University Press.

Nur Syazwani Halim and Marlyna Maros. 2014. The functions of code-switching in facebook interactions. *Procedia-Social and Behavioral Sciences*, 118:126–133.

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)*.

Monica Heller. 1992. The politics of codeswitching and language choice. *Journal of Multilingual & Multicultural Development*, 13(1-2):123–142.

Sai Muralidhar Jayanthi and Akshat Gupta. 2021. Sj_aj@ dravidianlangtech-eacl2021: Task-adaptive pre-training of multilingual bert models for offensive language identification. *arXiv preprint arXiv:2102.01051*.

Istvan Kecskes. 2006. The dual language model to explain code-switching: A cognitive-pragmatic approach.

Souad Kheder and Edith Kaan. 2021. Cognitive control in bilinguals: Proficiency and code-switching both matter. *Cognition*, 209:104575.

Sophia Lee and Zhongqing Wang. 2015. Emotion in code-switching texts: Corpus construction and analysis. In *Proceedings of the Eighth SIGHAN workshop on chinese language processing*, pages 91–99.

John Macnamara and Seymour L Kushnir. 1971. Linguistic independence of bilinguals: The input switch. *Journal of Verbal Learning and Verbal Behavior*, 10(5):480–487.

Adele W Miccio, Carol Scheffner Hammer, and Bárbara Rodríguez. 2009. *Code-switching and language disorders in bilingual children*. Cambridge University Press.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Sarah F Phillips and Liina Pylkkänen. 2021. Composition within and between languages in the bilingual mind: Meg evidence from korean/english bilinguals. *Eneuro*, 8(6).

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter? In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1131–1141.

Koustav Rudra, Ashish Sharma, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2019. Identifying and analyzing different aspects of english-hindi code-switching in twitter. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):1–28.

Susan De Santi, Loraine K Obler, Helene Sabo-Abramson, and Joan Goldberger. 1990. Discourse abilities and deficits in multilingual dementia. In *Discourse ability and brain damage*, pages 224–235. Springer.

Carol Myers Scotton and William Ury. 1977. Bilingual strategies: The social functions of code-switching.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Jan Svennevig, Pernille Hansen, Hanne Gram Simonsen, and Anne Marie Dalby Landmark. 2019. Code-switching in multilinguals with dementia: patterns across speech contexts. *Clinical linguistics & phonetics*, 33(10-11):1009–1030.

Jae DiBello Takeuchi. 2022. Code-switching as linguistic microaggression: L2-japanese and speaker legitimacy. *Multilingua*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *arXiv preprint arXiv:2103.13309*.

Ana Celia Zentella. 1998. *Growing up bilingual: Puerto rican children in New York*. Blackwell.

A Appendix

Hyperparameters For the mBERT and XLM-RoBERTa models as well as their respective adapter models, our hyperparameters were 20 epochs, a weight decay of 0.01, and we tuned the batch size from the set $4, 16$ and the learning rate from the set $2e^{-5}, 0.0001$ with grid search. In order to account for the variance between different initial seeds, we first found the best performing hyperparameter combination for each model on each task with the default seed of 42, then we ran the model four additional times with the same hyperparameters but with a different seed, from 30 to 20 to 10 to 5.

Comparison to Previous Annotation Schemes We give a mapping between labels in our annotation scheme and labels in other code-switching annotation schemes in Table 4.

Zentella (1998)	Begum et al. (2016)	Our paper
Topic shift, Declarative/question shift	Narrative-Evaluative, Cause-Effect	Change topic (discourse)
Narrative frame break	Sarcasm	Joke (sociological)
Direct Quotations	Quotations	Quote
Indirect Quotations	Reported Speech	(discourse)
Aggravating requests	Imperative	Command
Mitigating requests		(sociological)
Attention attraction		
Translations	Translation	Translate (clarify)
-	Reinforcement	-
Crutching	-	Borrowing (lexical)
Filling in	-	Filler (discourse)
-	-	Proper noun (lexical)
-	-	Happiness (express emotion)
-	Abuse/Neg. Sentiment	Exasperation (express emotion)
-	-	Surprise (express emotion)

Table 4: Comparison of our annotation scheme with other frameworks for code-switching

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3.1, 4

☑ B1. Did you cite the creators of artifacts you used?
3.1, 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3.1, 4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
5

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3.1

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4, Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4.1

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
3.4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3.4

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We recruited the annotator in an informal context; they were a friend who agreed to spare a few hours of their time to help us out.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
We used a pre-existing, publicly available and anonymized dataset collected by Bangor Miami researchers.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We used a pre-existing, publicly available and anonymized dataset collected by Bangor Miami researchers.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3.1