# Speech Analysis for Code-Switched Settings


Debasmita Bhattacharya

Advanced Spoken Language Processing

November 25th, 2025

# What is code-switching?

Dice que le dieron ese ...

puesto de interno .

internship .

(= He said they gave him this internship.)

# Why study code-switching? And how?

- Most of the world speaks more than one language!
  - ➢ It is important to build language technologies that are robust to diverse linguistic settings
- Draw on a variety of methods
  - ➢ Computational paralinguistics, discourse-functional analysis, acoustic-prosodic techniques, …

- ***How*** do speakers code-switch? → proficiency, entrainment, prosody
- ***Why*** do speakers code-switch? → empathy, discourse function/content

# From Context to Code-switching: Examining the Interplay of Language Proficiency and Multilingualism in Speech

**Debasmita Bhattacharya**, Aanya Tolat, Julia Hirschberg.

Published at INTERSPEECH 2025

# Motivation: language proficiency x code-switching

- **Language proficiency**
  - A speaker's competence and capability to use oral and/or written language accurately and appropriately in a variety of settings.

  - Individual linguistic characteristics → years of experience
  - Language exposure aspects → medium of schooling

# Research Question

- How is **language proficiency**, as encoded by demographic and linguistic factors, related to the **quantity, language distribution, and syntactic complexity of code-switching** in a conversational domain?

# Background: quantifying code-switching

- CSW ratio: # switches / length
- **M**ultilingual-index

$$\text{M-Index} \equiv \frac{1 - \sum p_j^2}{(k-1) \cdot \sum p_j^2}.$$

- **I**ntegration-index

$$\text{I-Index} \equiv \frac{1}{n-1} \sum_{1 \leq i = j-1 \leq n-1} S(l_i, l_j).$$

# Background: strategies of code-switching

| Strategy | | Example Sentence |
|---|---|---|
| Monolingual | *EN* | Do you have any friend who studies linguistics? |
| | *SP* | *¿Tienes algún amigo que estudie lingüística?* |
| Insertional | $SP \xrightarrow{ins} EN$ | Do you have any *amigo* who studies *lingüística*? |
| | $EN \xrightarrow{ins} SP$ | *¿Tienes algún* friend *que estudie* linguistics? |
| Alternational | $EN \xrightarrow{alt} SP$ | Do you have any friend *que estudie lingüística*? |
| | $SP \xrightarrow{alt} EN$ | *Tienes algún amigo* that studies linguistics? |
| Neither | – | *pero* she is the case manager for those patients |

# Data: Bangor Miami corpus

- **Conversational data**
  - CSW quantity and frequency: CSW ratio, M-index, I-index
  - CSW language distribution: relative token counts
  - CSW syntactic complexity: CSW strategy
- **Questionnaire data**
  - Linguistic: age, years of experience, self-reported ability
  - Demographic: parents' primary language, medium of schooling

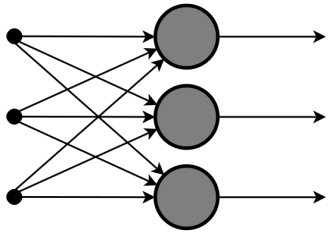
# Method

## Statistical analysis

- Chi squared tests
- ANOVA tests
- Pearson correlation
- Logistic regression

## Predictive modeling

- Logistic regression and SVM
  - 3-fold CV
  - 80-20 train-test split



# Results: language proficiency is significantly related to CSW quantity

Table 1: *F-statistics from one-way ANOVA tests for quantity of CSW. All values are statistically significant with* $p < 0.001$.

| Factor | CSW ratio | M-index | I-index |
|---|---|---|---|
| Mother's language | 304.17 | 35.70 | 281.96 |
| Father's language | 205.05 | 20.37 | 191.51 |
| Primary school language | 162.05 | 45.37 | 156.38 |
| Secondary school language | 201.17 | 65.24 | 194.54 |

# Results: language proficiency is significantly related to predominant language in CSW

| Factor | Odds Ratio |
|---|---|
| Primary school language | 4.97 |
| Secondary school language | 7.32 |

# Results: language proficiency is significantly related to predominant language in CSW

Table 2: *Summary of logistic regression for Spanish-dominant (coded 1) versus English-dominant (coded 0) CSW.*

| Factor | coef | std err | t |
|---|---|---|---|
| Intercept | 1.1875 | 0.993 | 1.196 |
| Years of experience (es:en) | -1.5283 | 0.944 | -1.619 |
| Reported ability (es:en) | 0.7417 | 0.248 | 2.993 |
| Age | 0.0032 | 0.004 | 0.834 |

# Results: language proficiency is not directly related to CSW strategy

Table 3: *Summary of logistic regression for alternational (coded 1) versus insertional (coded 0) CSW.*

| Factor | coef | std err | t |
|---|---|---|---|
| Intercept | 2.7343 | 1.020 | 2.681 |
| Years of experience (es:en) | -1.7861 | 0.970 | -1.841 |
| Reported ability (es:en) | -0.3678 | 0.255 | -1.444 |
| Age | 0.0037 | 0.004 | 0.919 |

# Results: language proficiency can predict CSW behaviour

Table 4: *Evaluating model performance on predicting CSW behavior from language proficiency factors.*

| Prediction task | Mean Accuracy (SD) Logistic regression | SVM |
|---|---|---|
| High CSW ratio | 0.85 (0.04) | 0.88 (0.00) |
| High M-index | 0.71 (0.03) | 0.73 (0.02) |
| High I-index | 0.90 (0.02) | 0.90 (0.02) |
| Dominant CSW language | 0.75 (0.12) | 0.69 (0.02) |
| Dominant CSW strategy | 0.63 (0.06) | 0.62 (0.01) |

# Conclusions

- How is **language proficiency**, as encoded by demographic and linguistic factors, related to the **quantity, language distribution, and syntactic complexity of code-switching** in a conversational domain?

# Conclusions

1. Parents' primary language → code-switching quantity and frequency

# Conclusions

1. Parents' primary language → code-switching quantity and frequency
2. Medium of instruction + self-reported ability → dominant code-switching language

# Conclusions

1. Parents' primary language → code-switching quantity and frequency
2. Medium of instruction + self-reported ability → dominant code-switching language
3. Language proficiency →? code-switching strategy

# Conclusions

- Language proficiency has notable relationships with several aspects of code-switching.
  - These relationships can even be learned and applied by simple predictive models of code-switching behaviour.

Paper

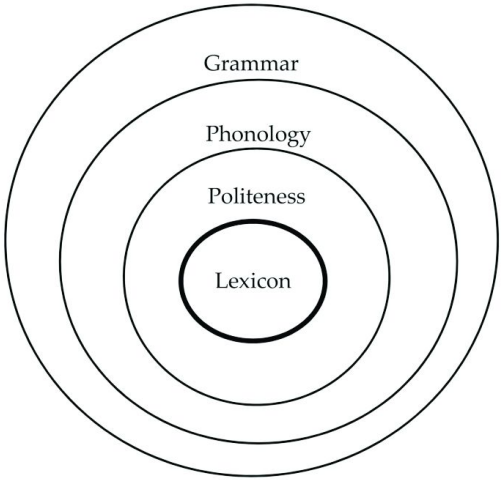
# Measuring Entrainment in Spontaneous Code-switched Speech

**Debasmita Bhattacharya**, Siying Ding, Alayna Nguyen, Julia Hirschberg.

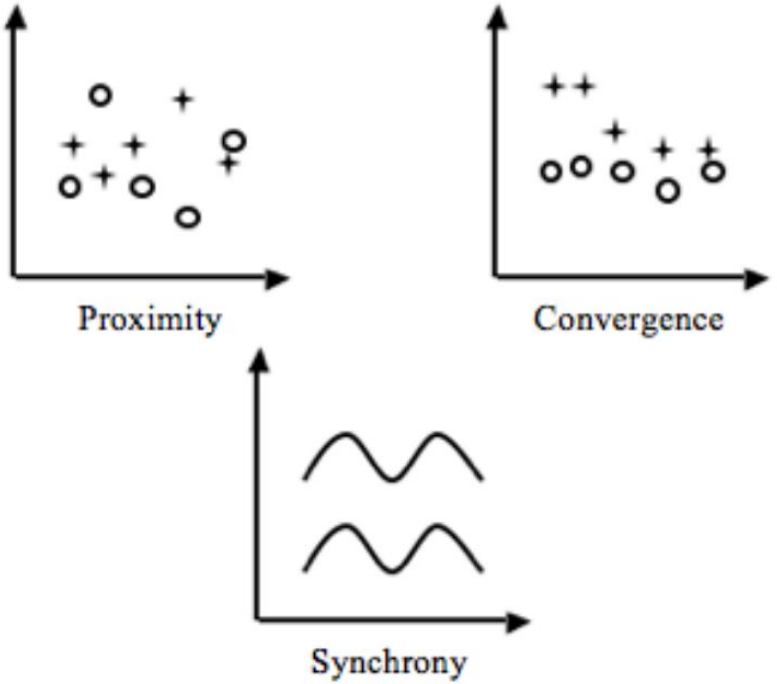
Published at NAACL 2024

# Background: what is entrainment?

- Speakers become more like their interlocutors across various dimensions
  - (= alignment = accommodation = coordination)



# Background: types of entrainment



# Motivation: entrainment in new multilingual domain

- Prior work: entrainment occurs in **monolingual** writing and speech
  - Nenkova et al., 2008.
  - Levitan and Hirschberg, 2011.

- Prior work: entrainment occurs in code-switched **human-machine** communication
  - Ahn et al., 2020.
  - Parekh et al., 2020.

- **What about entrainment in multilingual human speech?**

# Research Questions

1. Do patterns of entrainment in monolingual settings generalize to code-switched settings?

2. Do patterns of entrainment on code-switching in text — some of which is produced by virtual dialogue agents — generalize to spontaneous code-switching in speech?

# Data: augmenting the Bangor Miami corpus

- Bangor Talk: Miami corpus
  - ➕ Code-switching strategies: I, A, O
  - ➖ Background noise

- Features
  - Lexical
  - Acoustic-prosodic
  - Code-switching

# Method: statistical analyses

- Lexical features
  - Count-based and probabilistic measures

$$entr(S_A, S_B) = -\sum_{w \in W} |\frac{count_{S_A}(w)}{ALL_{S_A}} - \frac{count_{S_B}(w)}{ALL_{S_B}}|$$

- Acoustic-prosodic // Code-switching features
  - Turn- and conversation-level
  - Proximity, convergence, synchrony

# Research Questions

1. **Do patterns of entrainment in monolingual settings generalize to code-switched settings?**

2. Do patterns of entrainment on code-switching in text — some of which is produced by virtual dialogue agents — generalize to spontaneous code-switching in speech?

# Results: interlocutors entrain on lexical features of code-switched conversations

- **Overall language use**
  - Including OOVs: 97% of conversations
  - Excluding OOVs: 74% of conversations

- **Word sets**
  - Top-100 words in corpus: 100% of conversations
  - Affirmative cue words: 100% of conversations
  - Top-25 words in corpus: 97% of conversations
  - Filler words: 97% conversations

# Results: interlocutors entrain on lexical features of code-switched conversations

top-25$_{corpus}$ CSW:A

$S_{A1}$: pero no la puedes hacer BECAUSE YOU CAN'T START CHECKING IT.

CSW ratio = 0.545

top-25$_{corpus}$ CSW:A

$S_{B1}$: pero es que HE IS SO TECHNOLOGICALLY ADVANCED.

CSW ratio = 0.625

| Speaker | Mean pitch (Hz) | Speaking rate (syllables/sec) |
|---|---|---|
| A | 212 | 5.71 |
| B | 179 | 4.27 |

# Results: interlocutors entrain on most acoustic-prosodic features of code-switched conversations

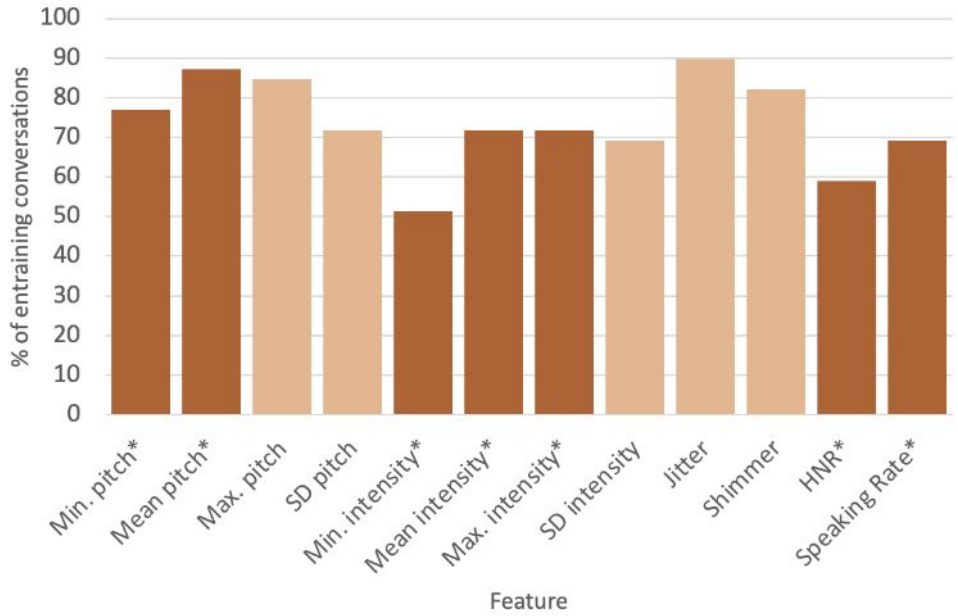



Figure 1: Proximity at the turn-level.

# Results: interlocutors entrain on most acoustic-prosodic features of code-switched conversations

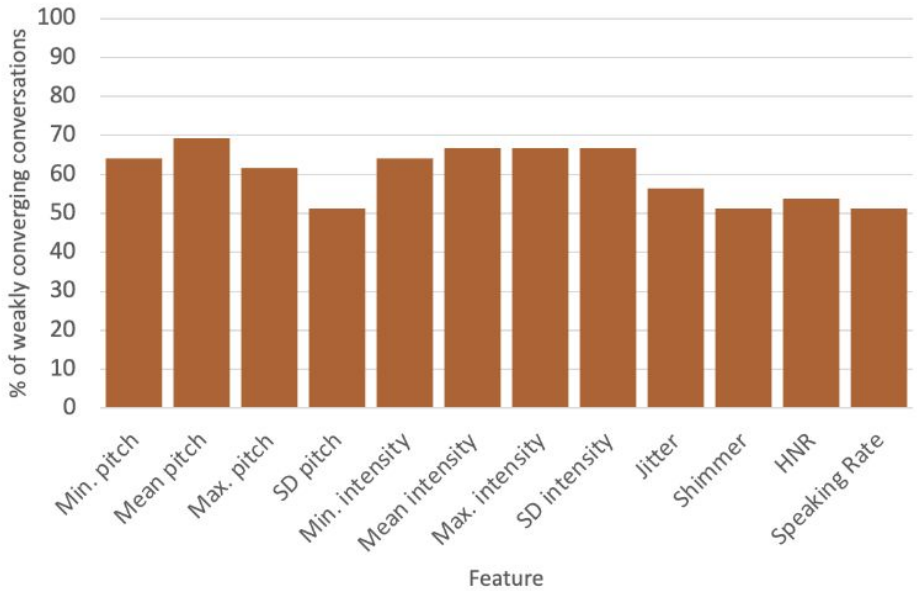



Figure 2: Convergence at the turn-level.

# Results: interlocutors entrain on most acoustic-prosodic features of code-switched conversations

CSW:I

$S_{A2}$: pero mi | PRINTER DOESN'T WORK.

CSW:O

CSW ratio = 0.6

$S_{B2}$: eh | WHAT PRINTER?

CSW ratio = 0.667

| Speaker | Mean pitch (Hz) | Speaking rate (syllables/sec) |
|---|---|---|
| A | 159 | 4.15 |
| B | 164 | 4.31 |

# Research Questions

1. Do patterns of entrainment in monolingual settings generalize to code-switched settings?

2. **Do patterns of entrainment on code-switching in text — some of which is produced by virtual dialogue agents — generalize to spontaneous code-switching in speech?**

# Results: interlocutors entrain on most code-switching features of code-switched conversations

- **Presence of code-switching** 🌎
  - Turn-level proximity: 74% of conversations
  - Conversation-level proximity: 82% of conversations
  - Turn-level synchrony: 54% of conversations

- **Quantity of code-switching** 🌎
  - Turn-level proximity: 85% of conversations
  - Conversation-level proximity: 87% of conversations

# Results: interlocutors entrain on most code-switching features of code-switched conversations

CSW:I

$S_{A2}$: pero mi PRINTER DOESN'T WORK.

CSW:O

CSW ratio = 0.6

$S_{B2}$: eh WHAT PRINTER?

CSW ratio = 0.667

| Speaker | Mean pitch (Hz) | Speaking rate (syllables/sec) |
|---|---|---|
| A | 159 | 4.15 |
| B | 164 | 4.31 |

# Results: interlocutors entrain on most code-switching features of code-switched conversations

- **Strategy of code-switching** 🌎
  - Turn-level synchrony: 56% of conversations - **insertional**
  - Conversation-level proximity: 74% of conversations - **alternational**

# Results: interlocutors entrain on most code-switching features of code-switched conversations

top-25$_{corpus}$ CSW:A

$S_{A1}$: pero no la puedes hacer BECAUSE YOU CAN'T START CHECKING IT.

CSW ratio = 0.545

top-25$_{corpus}$ CSW:A

$S_{B1}$: pero es que HE IS SO TECHNOLOGICALLY ADVANCED.

CSW ratio = 0.625

| Speaker | Mean pitch (Hz) | Speaking rate (syllables/sec) |
|---|---|---|
| A | 212 | 5.71 |
| B | 179 | 4.27 |

# Conclusions

# Conclusions

1. Do patterns of entrainment in monolingual settings generalize to code-switched settings? ✅
   a. 
   b. 

# Conclusions

1. Do patterns of entrainment in monolingual settings generalize to code-switched settings? ✅
   a. 
   b. 

2. Do patterns of entrainment on code-switching in text — some of which is produced by virtual dialogue agents — generalize to spontaneous code-switching in speech? ✅ ➕
   a. 

# Conclusions

1. Do patterns of entrainment in monolingual settings generalize to code-switched settings? ✅
   a. 🖋️
   b. 🗣️

Paper

2. Do patterns of entrainment on code-switching in text — some of which is produced by virtual dialogue agents — generalize to spontaneous code-switching in speech? ✅ ➕
   a. 🌎

# The Sound of Code-Switching: Prosodic Signatures of Spanish-English Speech

**Debasmita Bhattacharya***, Michela Marchini*Julia Hirschberg.

Under review at Speech Prosody 2026

# Motivation: prosodic production x code-switching

- Prior work on code-switching has effectively captured its morpho-syntactic, sociolinguistic, etc. characteristics
  - ➢ But this has primarily been tied to *writing*!
- **Prosody** captures a uniquely *spoken* aspect of speech production
  - ➢ What can we learn about spoken code-switching that we can't discover from studying its transcripts alone?

# Research Questions

1. Is there utterance-level **variation** across a suite of language-independent **pitch**, **energy**, and **duration** features between code-switched and monolingual spontaneous speech?

2. How is this variation **influenced** by (a) speaker **proficiency** and (b) **linguistic characteristics** of multilingual speech?

# Research Questions

1. Is there utterance-level **variation** across a suite of language-independent **pitch**, **energy**, and **duration** features between code-switched and monolingual spontaneous speech?

2. How is this variation **influenced** by (a) speaker **proficiency** and (b) **linguistic characteristics** of multilingual speech?

# Data: Bangor Miami corpus (again)

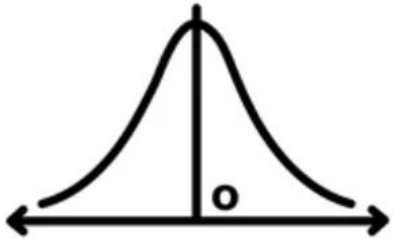
- **Conversational data**
  - 103 prosodic features spanning pitch, energy, duration
    - Initial and final voiced and unvoiced segments
    - 6 functionals: mean, SD, max., min., skew, kurtosis
  - CSW quantity and frequency: M-index, I-index
  - CSW syntactic complexity: CSW strategy

- **Questionnaire data**

# Method

## ❖ Statistical analysis

- ➢ Independent *t*-tests
- ➢ Cohen's *d* effect sizes
- ➢ *z*-tests of proportions



## ❖ Modeling

- ➢ Unsupervised analysis
  - ■ *K*-means clustering
- ➢ Supervised analysis
  - ■ Binary LID: Whisper-base

# Research Questions

1. Is there utterance-level **variation** across a suite of language-independent **pitch**, **energy**, and **duration** features between code-switched and monolingual spontaneous speech?

2. How is this variation **influenced** by (a) speaker **proficiency** and (b) **linguistic characteristics** of multilingual speech?

# Results: Code-switching differs prosodically from monolingual speech

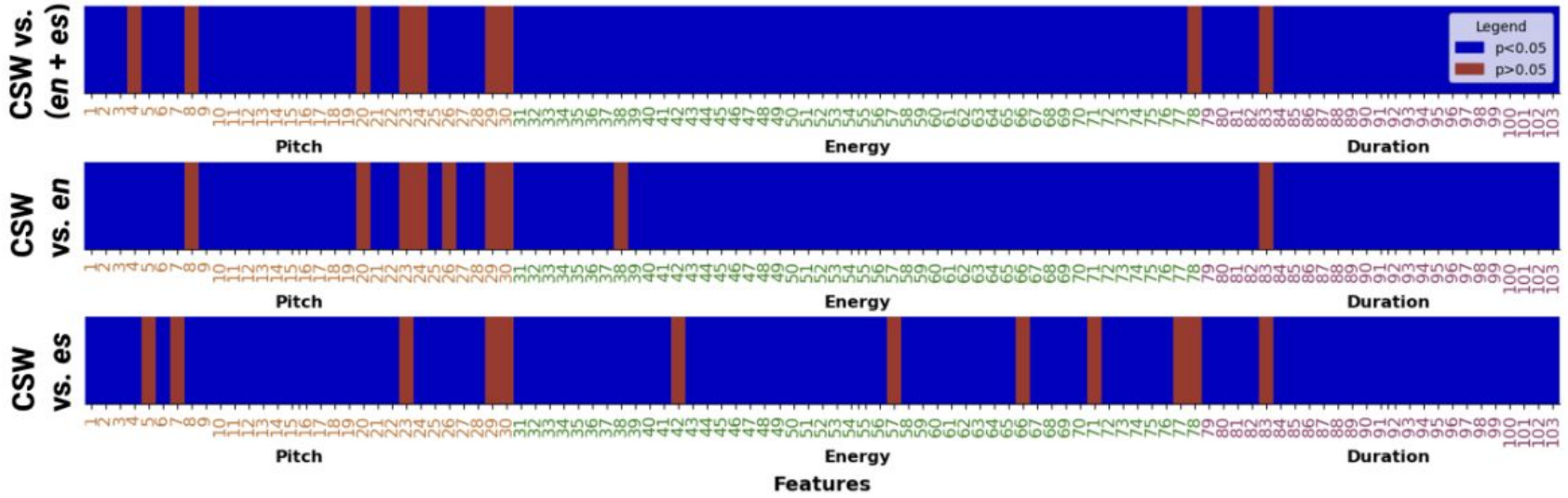



Figure 1: *p-values for prosodic feature distribution comparisons between code-switched and combined monolingual English & Spanish (top), monolingual English (middle), and monolingual Spanish (bottom) utterances. Blue indicates statistical significance; red indicates insignificance. Associated Cohen's d values generally fall within* $0.2 < d < 0.8$ *(i.e. small to medium effects) for significant features.*

# Results: Code-switching differs prosodically from monolingual speech

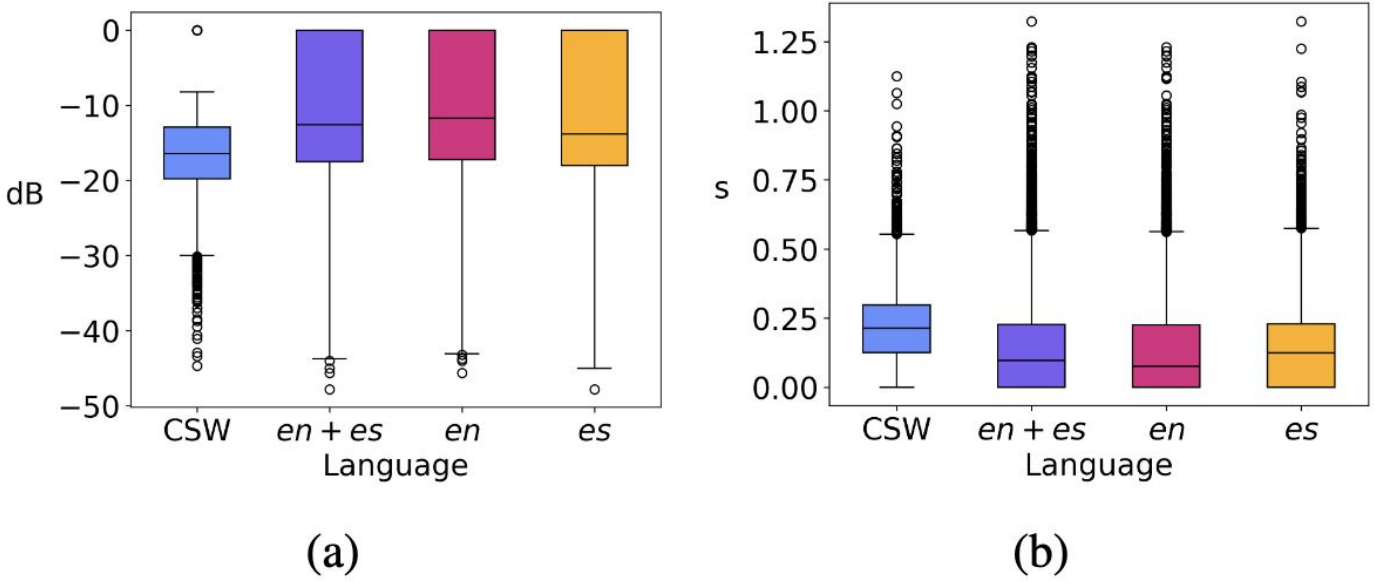



Figure 2: *Visualizing trends in (a) mean energy and (b) mean duration of intial voiced segments across utterance types.*

# Results: Code-switching differs prosodically from monolingual speech

Table 1: *k-means clustering performance across comparison settings. The first comparison* (en *vs.* es) *serves as a baseline.*

| Cluster comparison | Accuracy |
|---|---|
| Monolingual *en* vs. Monolingual *es* | 0.636 |
| All monolingual vs. code-switched | 0.843 |
| Monolingual *en* vs. code-switched | 0.837 |
| Monolingual *es* vs. code-switched | 0.868 |

# Results: Code-switching differs prosodically from monolingual speech

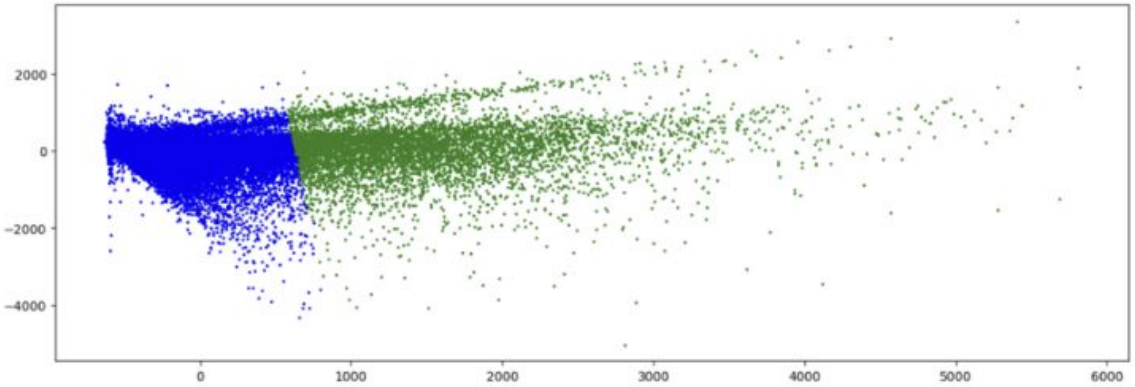



Figure 3: *Clustering all monolingual (blue) vs. code-switched (green) utterances. Axes represent the relevance of selected prosodic feature groups post-PCA; the highest loadings for PC1 (x-axis) and PC2 (y-axis) correspond to greater mean , max., and SD in energy and pitch features. Visualizations for clustering English vs. CSW and Spanish vs. CSW are almost identical.*

# Research Questions

1. Is there utterance-level **variation** across a suite of language-independent **pitch**, **energy**, and **duration** features between code-switched and monolingual spontaneous speech?

2. How is this variation **influenced** by (a) speaker **proficiency** and (b) **linguistic characteristics** of multilingual speech?

# Conclusions

1. Is there utterance-level **variation** across a suite of language-independent **pitch**, **energy**, and **duration** features between code-switched and monolingual spontaneous speech? ✅

2. How is this variation **influenced** by (a) speaker **proficiency** and (b) **linguistic characteristics** of multilingual speech?

🔝 »

# Switching Tongues, Sharing Hearts: Identifying the Relationship Between Empathy and Code-Switching in Speech

**Debasmita Bhattacharya***, Eleanor Lin*, Run Chen, Julia Hirschberg.

Published at INTERSPEECH 2024

# Background: why do speakers code-switch?

account for speaker competence

relate to audience identity

express (in)formality

express solidarity

express group identity

adapt to linguistic context

perform affective function

reflect shared experiences

adapt to conversation topic

# Background: why do speakers code-switch?

account for speaker competence

relate to audience identity

express (in)formality

**express solidarity**

**express group identity**

adapt to linguistic context

**perform affective function**

**reflect shared experiences**

adapt to conversation topic

# Motivation: empathy x code-switching

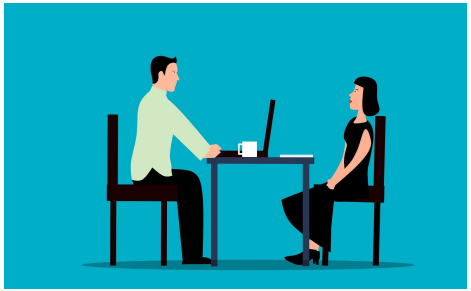
- **Empathy** → the ability to…
  - understand other people's feelings as if one were having them oneself, and
  - respond accordingly

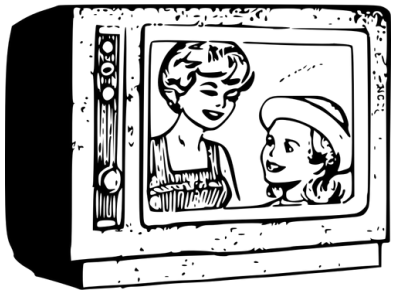
# Research Questions

1. Is there a relationship between ***code-switching*** prevalence in speech and the lexical and/or acoustic-prosodic correlates of ***empathy***?
2. Does the answer to **1.** ***generalize across language pairs*** involving different language families?

# Code-switched corpora

- **Spanish-English**: conversational
  - Bangor Miami
- **Mandarin-English**: conversational and interview
  - SEAME
- **Hindi-English**: soap opera
  - MaSaC





# Method

- **Approximate ground truth empathy labels**
  - RoBERTa base model (text only) & multimodal model (text +speech)
  - Fine tune on English empathetic utterances
  - Translate code-switched utterances to monolingual English
- **Compute code-switching metrics**
  - M-index & I-index
  - Code-switched vs. monolingual
- **Statistical analysis**
  - Chi squared test
  - Odds ratio
  - Pearson correlation

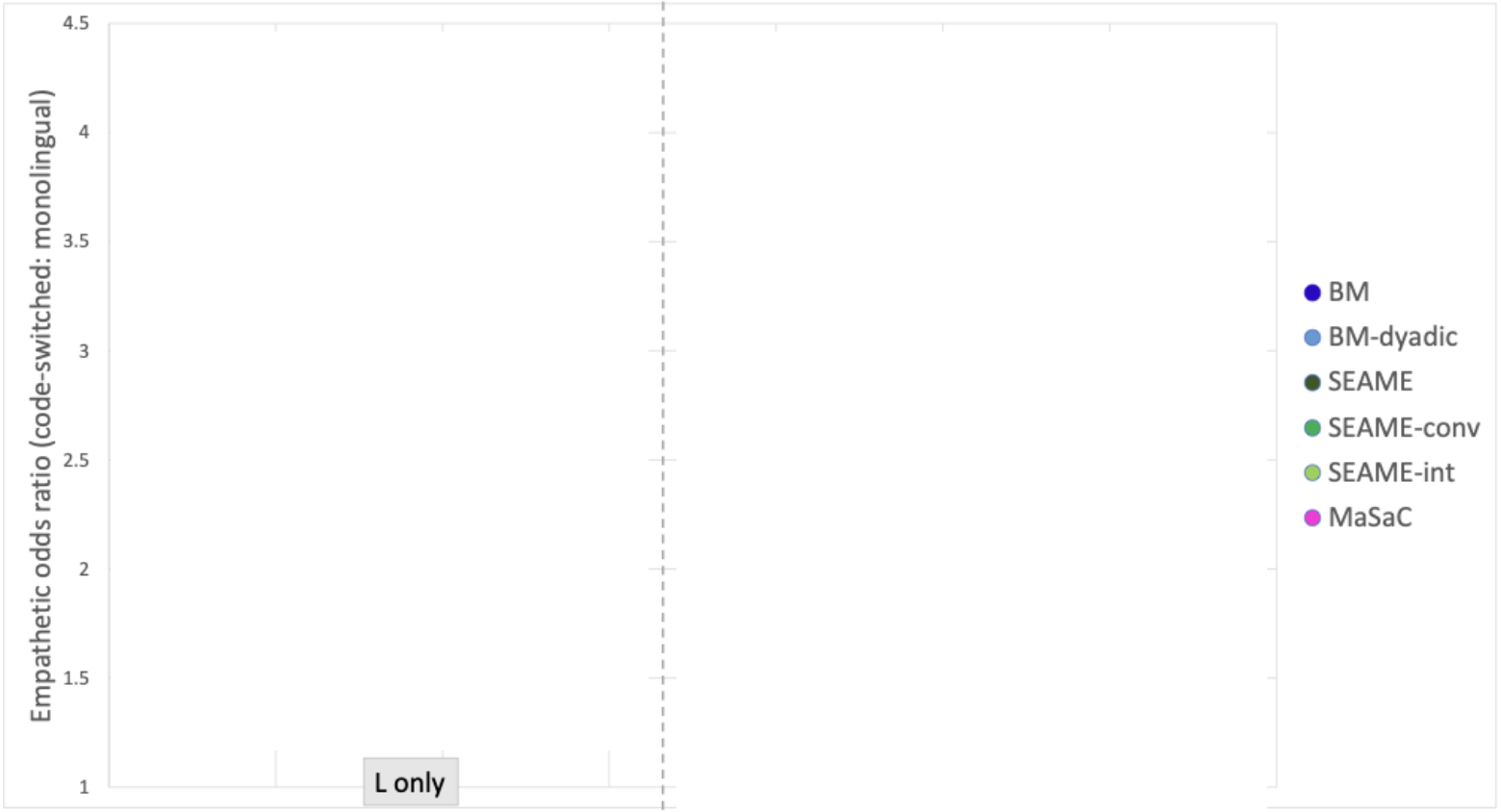
# Method

- **Approximate ground truth empathy labels**
  - RoBERTa base model (text only) & multimodal model (text +speech)
  - Fine tune on English empathetic utterances
  - Translate code-switched utterances to monolingual English
- **Compute code-switching metrics**
  - M-index & I-index
  - Code-switched vs. monolingual
- **Statistical analysis**
  - Chi squared test
  - Odds ratio
  - Pearson correlation

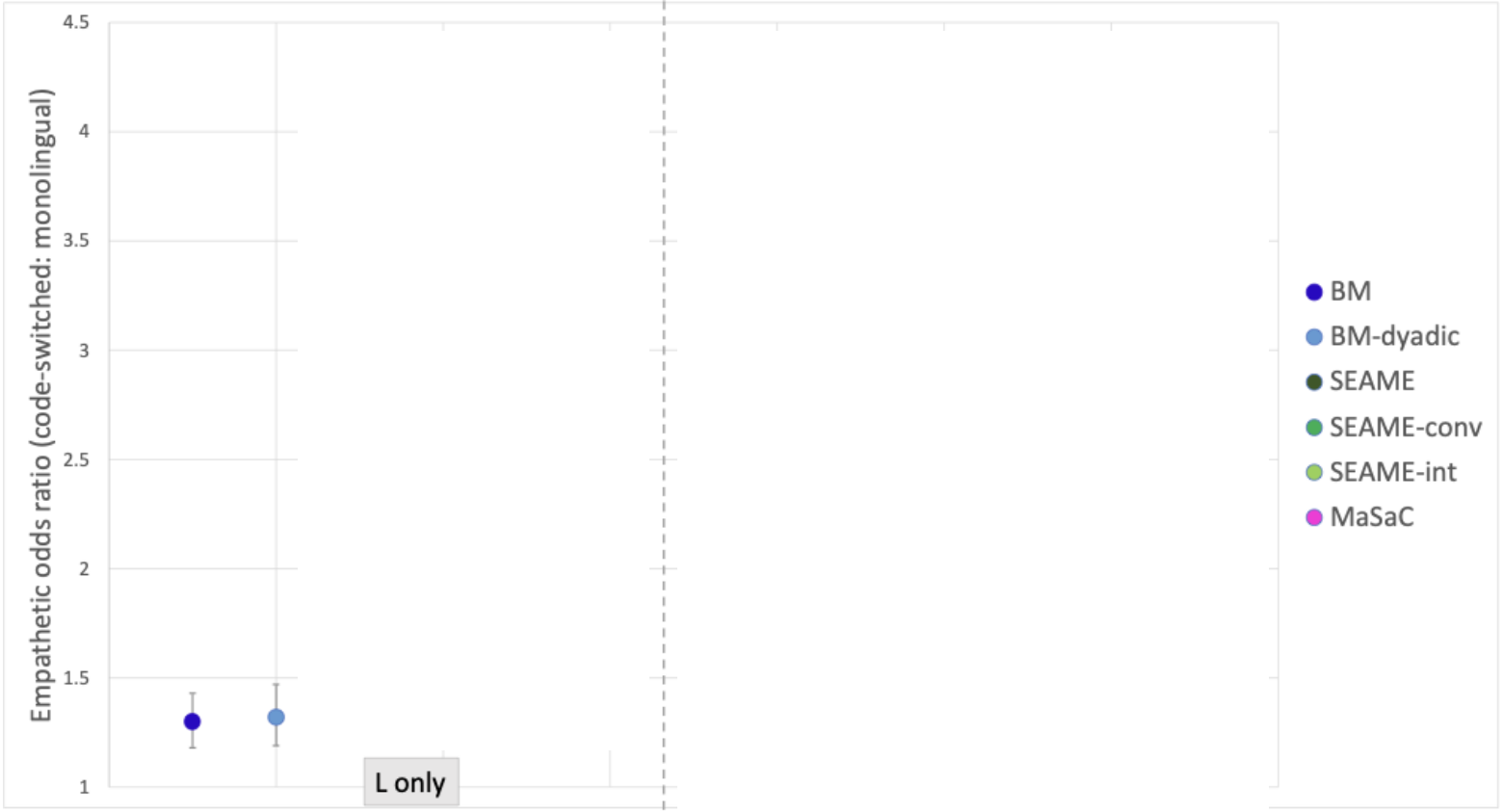
# Method

- **Approximate ground truth empathy labels**
  - RoBERTa base model (text only) & multimodal model (text +speech)
  - Fine tune on English empathetic utterances
  - Translate code-switched utterances to monolingual English
- **Compute code-switching metrics**
  - M-index & I-index
  - Code-switched vs. monolingual
- **Statistical analysis**
  - Chi squared test
  - Odds ratio
  - Pearson correlation

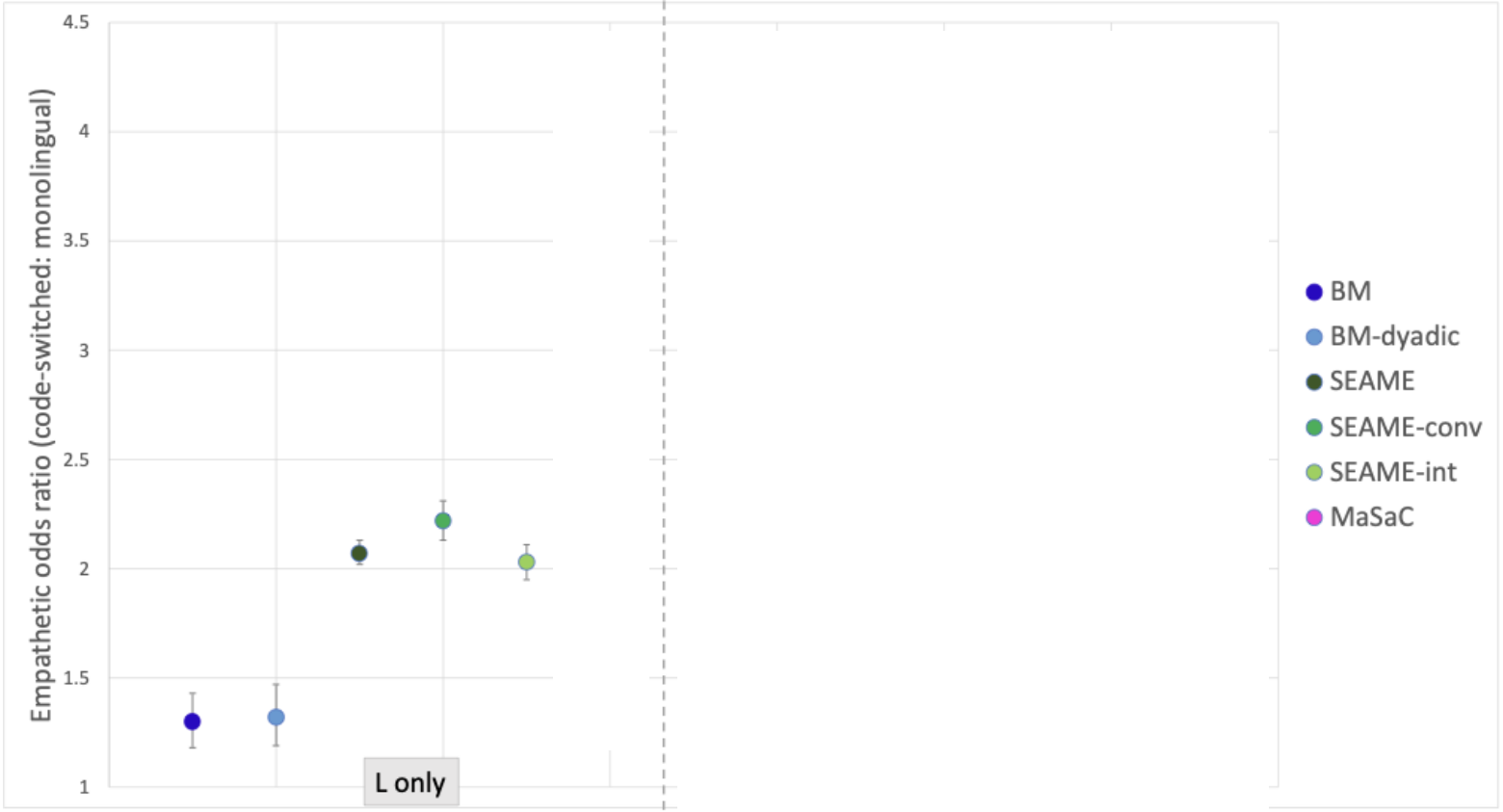
# Results: spoken CSW aligns with lexical correlates of empathy



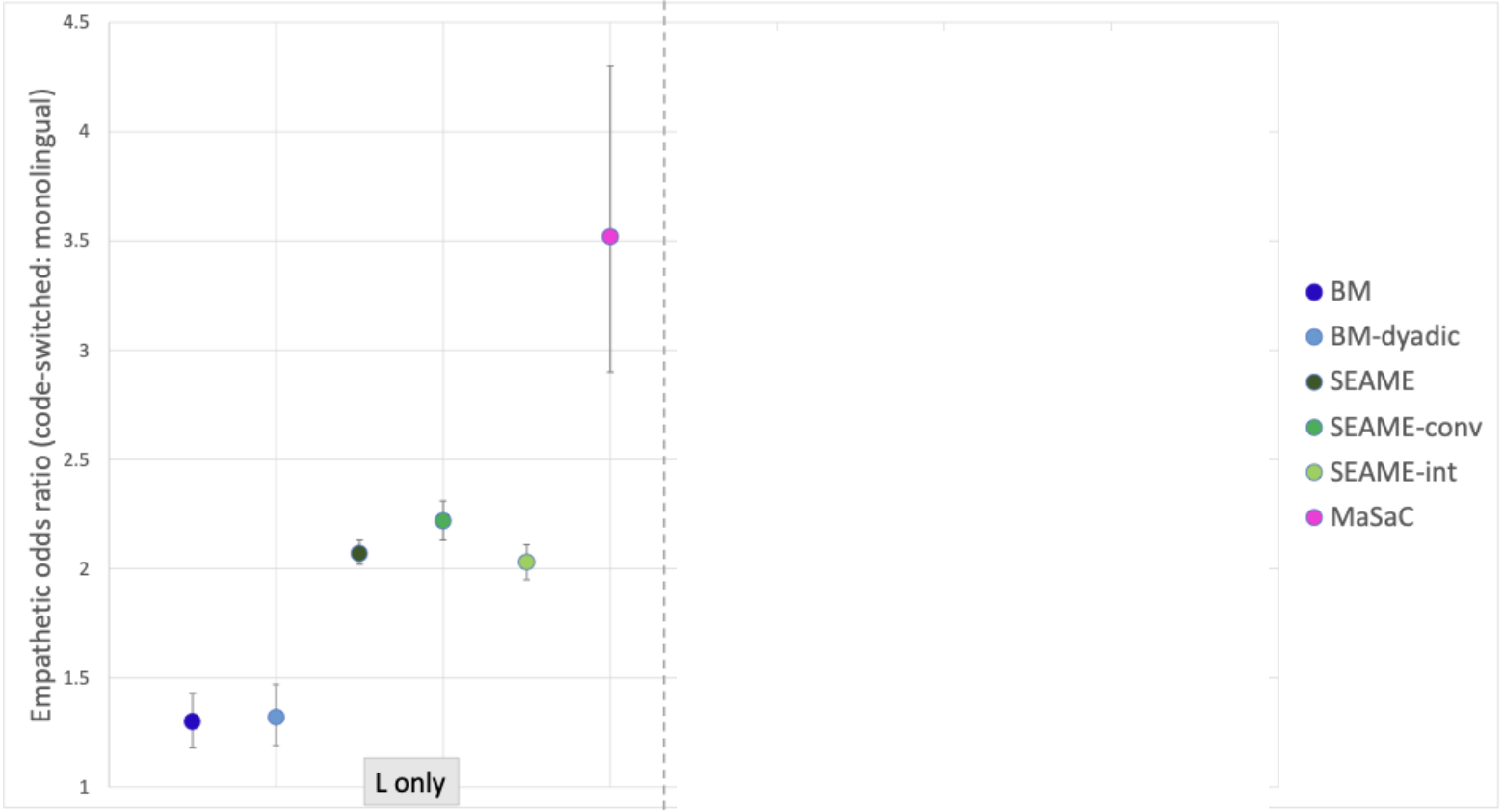
# Results: spoken CSW aligns with lexical correlates of empathy



# Results: spoken CSW aligns with lexical correlates of empathy



# Results: spoken CSW aligns with lexical correlates of empathy



# Results: spoken CSW aligns with lexical correlates of empathy

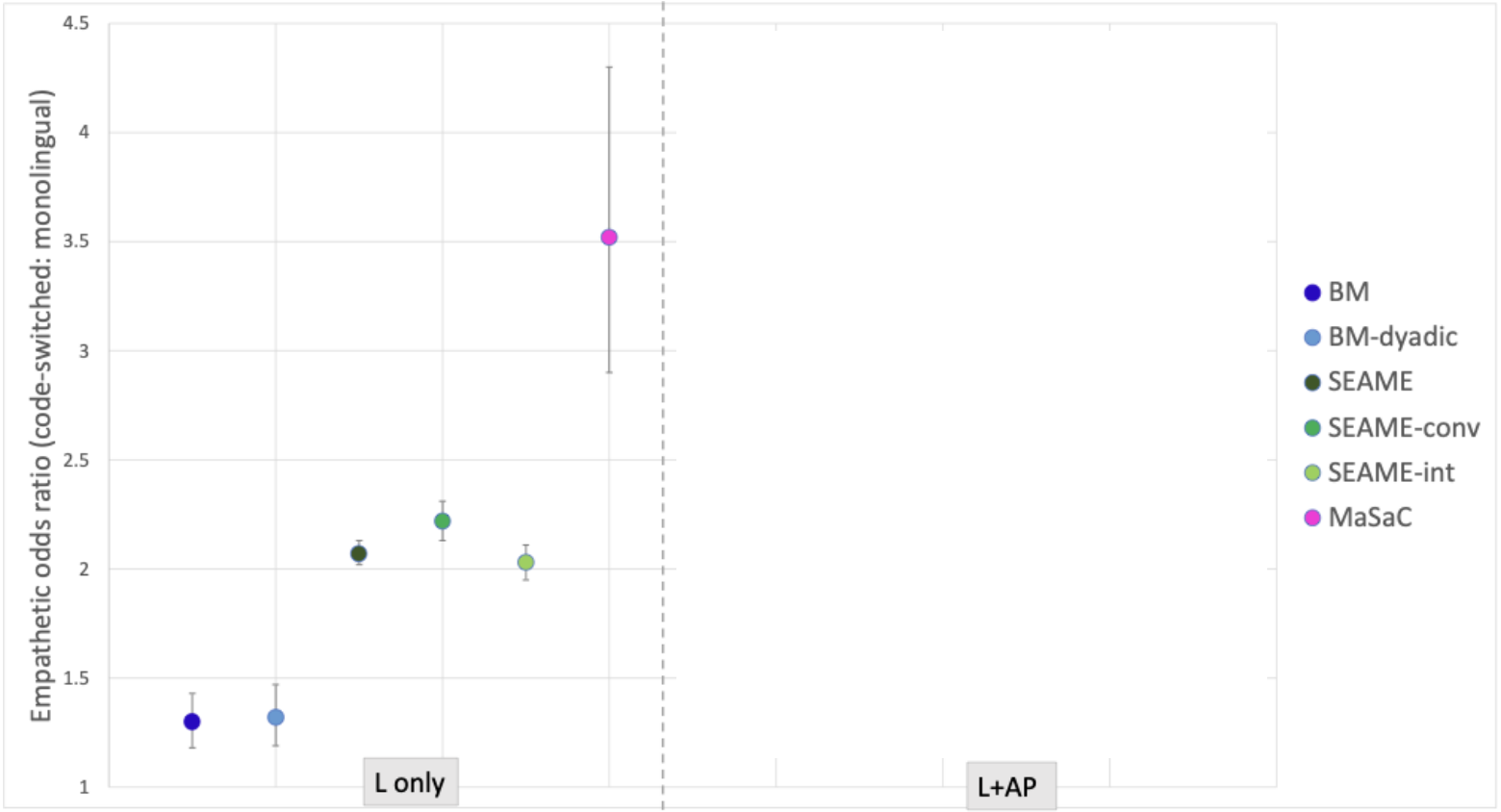
| | **Odds CSW:ML** | **CSW∝empathy** |
|---|---|---|
| Bangor Miami | ~1.3 | weak - |
| SEAME | ~2.1 | weak + |
| MaSaC | ~3.5 | weak + |

# Results: spoken CSW aligns with lexical correlates of empathy

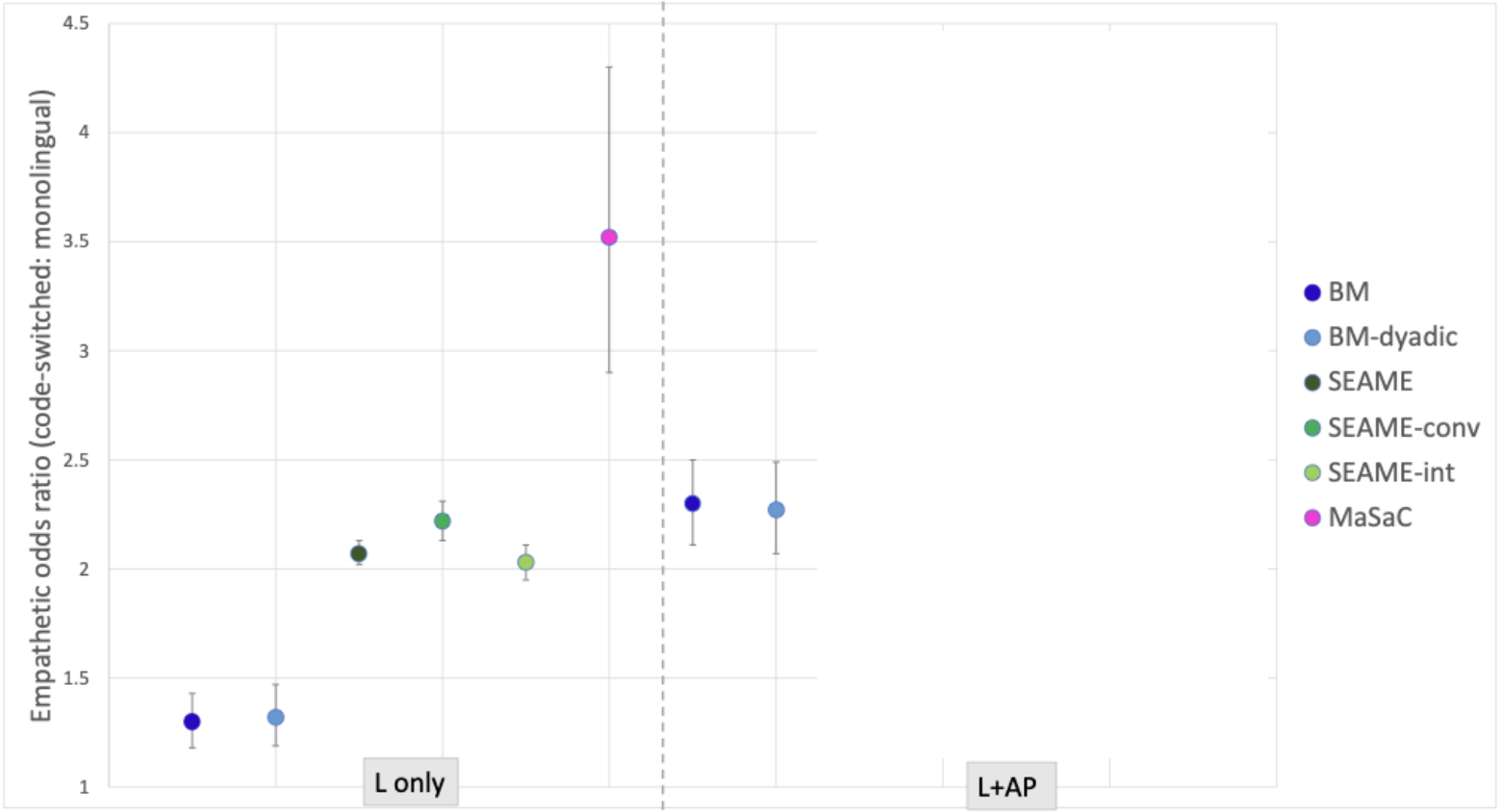
*Pero ahora mismo tú me estabas diciendo que te gustaba y te parecía* nice

= But right now you were telling me that you like it and you think it's nice

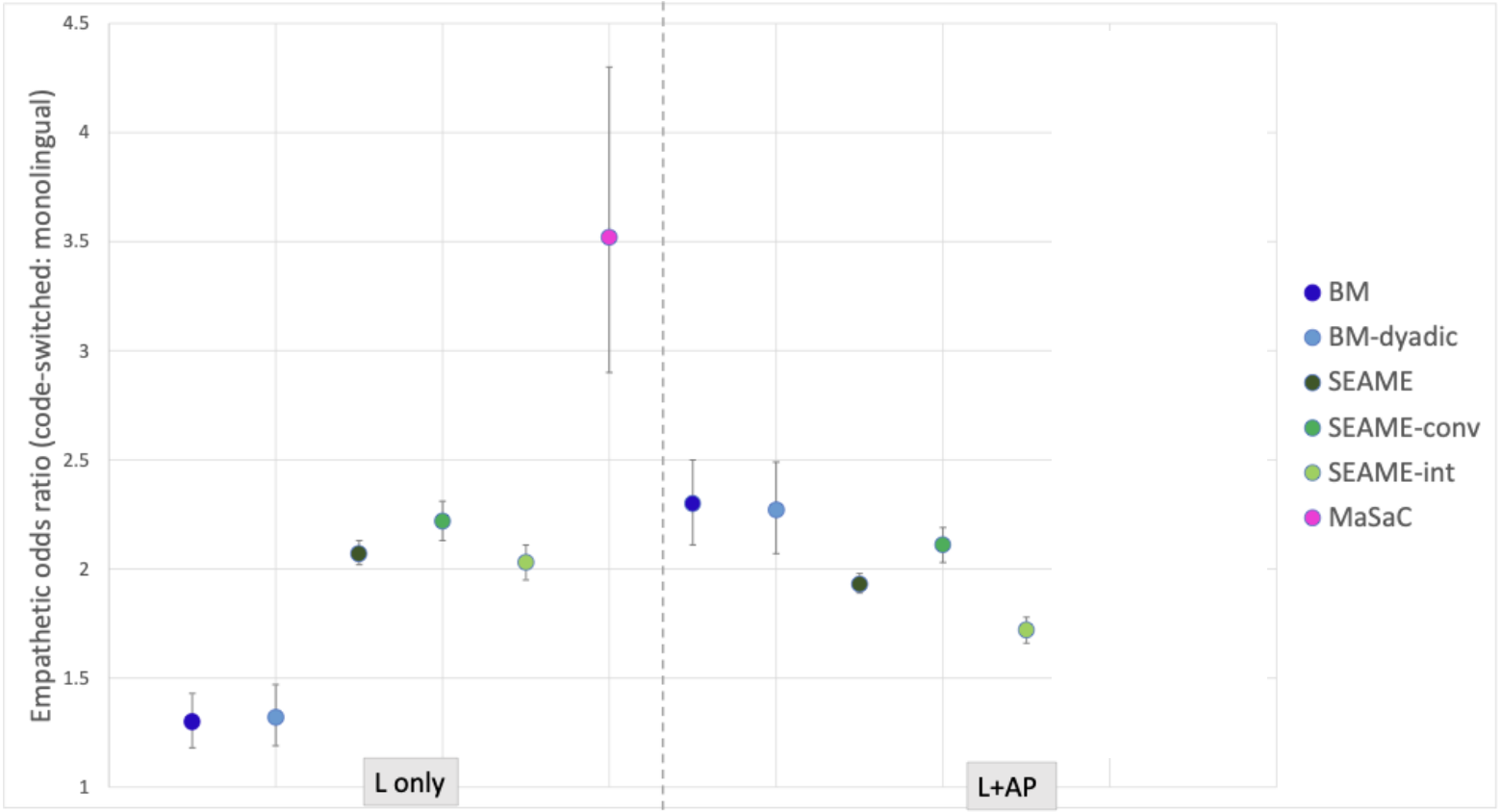
# Results: spoken CSW aligns (somewhat) with acoustic-prosodic correlates of empathy



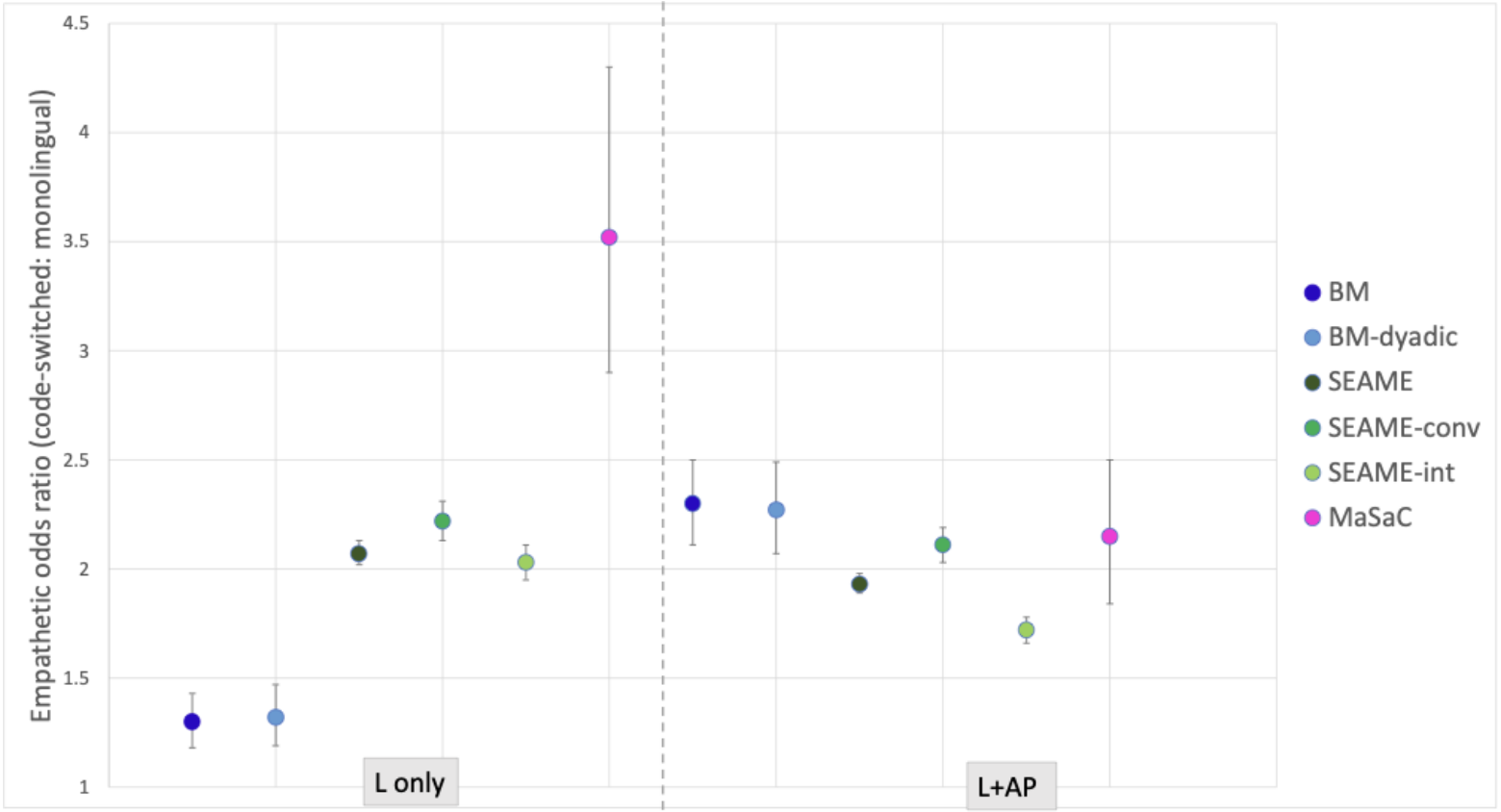
# Results: spoken CSW aligns (somewhat) with acoustic-prosodic correlates of empathy



# Results: spoken CSW aligns (somewhat) with acoustic-prosodic correlates of empathy



# Results: spoken CSW aligns (somewhat) with acoustic-prosodic correlates of empathy



# Results: spoken CSW aligns (somewhat) with acoustic-prosodic correlates of empathy

| | Odds CSW:ML | CSW∝empathy | Odds CSW:ML | CSW∝empathy |
|---|---|---|---|---|
| Bangor Miami | ~1.3 | weak - | ~2.3 | weak + |
| SEAME | ~2.1 | weak + | ~1.9 | weak + |
| MaSaC | ~3.5 | weak + | ~2.1 | none |

# Results: spoken CSW aligns (somewhat) with acoustic-prosodic correlates of empathy

Of course *mujhe pata hai, lekin ek baar tum kaho na* you know 24th February *ke baare mein, jab tum kehti ho na, to aur accha lagta hai*

= Of course I know, but once you say that you know about the 24th of February, when you say it, it feels better

# Conclusions

# Conclusions

1. Is there a relationship between ***code-switching*** prevalence in speech and the lexical and/or acoustic-prosodic correlates of ***empathy***? ✅

# Conclusions

1. Is there a relationship between ***code-switching*** prevalence in speech and the lexical and/or acoustic-prosodic correlates of ***empathy***? ✅
2. Does the answer to **1.** ***generalize across language pairs*** involving different language families? 🤔

# Conclusions

- Current metrics of empathy in speech generally align with the incidence of code-switching.

# Conclusions

- Current metrics of empathy in speech generally align with the incidence of code-switching.

# Conclusions

- The relationship between acoustic-prosodic empathetic features and code-switching may be more subtle than expected.

# Conclusions

- The relationship between acoustic-prosodic empathetic features and code-switching may be more subtle than expected.

- Next steps
  - Further exploration of acoustic-prosodic features
  - Multilingual (≠ English) empathy models
  - Determining causal relationship between empathy and code-switching

Paper

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

**Debasmita Bhattacharya**, Juan Junco, Divya Tadimeti, Julia Hirschberg.

Published at EMNLP 2025

# Background: what is discourse?


- How written or spoken language is used to produce meaning, through interactions of context and form
- Discourse content → named entities
- Discourse function → dialogue acts

"We enjoyed going to the Heat game!"

**DA: Statement of opinion** **NE type: Organization**

# Motivation: discourse x code-switching?

"*Entonces, ella trabaja* this Friday?"

**Utt. type: ML** **DA: Statement of opinion** **NE type: Organization** **NE Lang.: English**

"*Creo que* it's about 350 square feet."

**Utt. type: CSW - alternational** **DA: Statement of non-opinion**

"*Mi mamá dijo que* she didn't know about it."

**Utt. type: CSW - alternational** **DA: Quotation**

"*A ver*, let's see… "

**Utt. type: CSW - insertional**

**DA: Self talk**

# Research Questions

1. How are patterns of **CSW** influenced by the **content** of the utterance?
2. How is **CSW** influenced by the **function** of the utterance?
    a. Does this **interact** with patterns stemming from the content of code-switched speech?

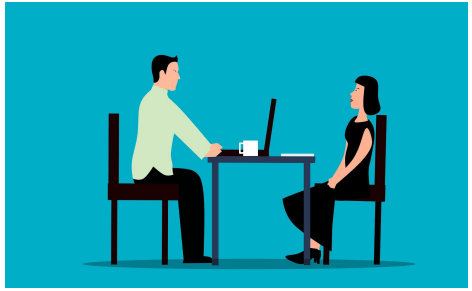
# Research Questions

1. How are patterns of **CSW** influenced by the **content** of the utterance?
2. How is **CSW** influenced by the **function** of the utterance?
    a. Does this **interact** with patterns stemming from the content of code-switched speech?

# Data: Augmenting the Bangor Miami corpus (again)

- **Spanish-English**
  - Mix of monolingual and code-switched speech
- Spontaneous and informal

- **Annotations**
  - **NEs**: “N”, “O”, “P”, “B”, “T”, “R”, “U"
  - **DAs**: Switchboard Dialogue Act Set
  - **CSW metrics**: CSW ratio, M-index, I-index, CSW strategies

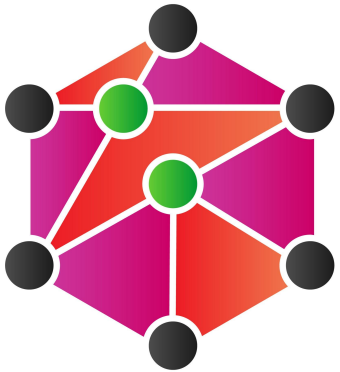


# Method

- **Statistical methods → NEs and DAs**
  - Chi-squared tests
  - Odds ratio calculations
  - One-way ANOVA test
  - *z*-tests of proportions
- **Unsupervised methods → DAs**
  - *k*-means clustering
- **Supervised methods → NEs and DAs**
  - Dependency parsing (NEs only)
  - Transition modeling:
    - Logistic regression
    - Hidden Markov Model

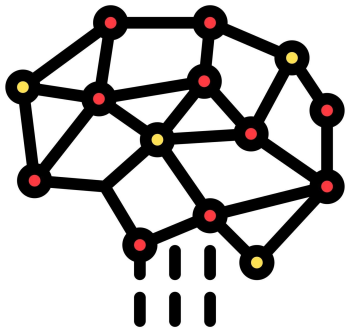
# Method

- **Statistical methods → NEs and DAs**
  - Chi-squared tests
  - Odds ratio calculations
  - One-way ANOVA test
  - *z*-tests of proportions
- **Unsupervised methods → DAs**
  - *k*-means clustering
- **Supervised methods → NEs and DAs**
  - Dependency parsing (NEs only)
  - Transition modeling:
    - Logistic regression
    - Hidden Markov Model



# Method

- **Statistical methods → NEs and DAs**
  - Chi-squared tests
  - Odds ratio calculations
  - One-way ANOVA test
  - *z*-tests of proportions
- **Unsupervised methods → DAs**
  - *k*-means clustering
- **Supervised methods → NEs and DAs**
  - Dependency parsing (NEs only)
  - Transition modeling:
    - Logistic regression
    - Hidden Markov Model



# Research Questions

1. How are patterns of **CSW** influenced by the **content** of the utterance?
2. How is **CSW** influenced by the **function** of the utterance?
    a. Does this **interact** with patterns stemming from the content of code-switched speech?

# Results: Dialogue acts differ between code-switched and monolingual contexts

- DAs appear similarly distributed across CSW and ML contexts

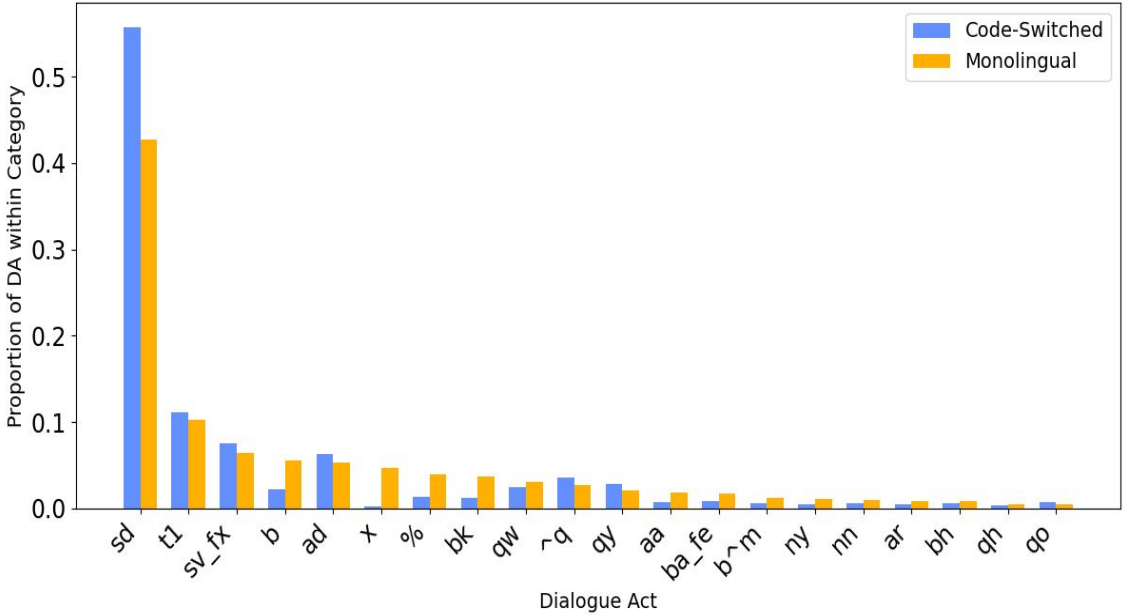



Fig. 4: Distribution of DA types across CSW and ML utterances.

# Results: Dialogue acts differ between code-switched and monolingual contexts

- Most commons DAs: statements of opinion (sv-fx; e.g. “oh no, *qué estúpida*.”) and non-opinion (sd; e.g. “*creo que* it’s about 350 square feet.”) and self-talk (t1; e.g. “*a ver*, let’s see…”)

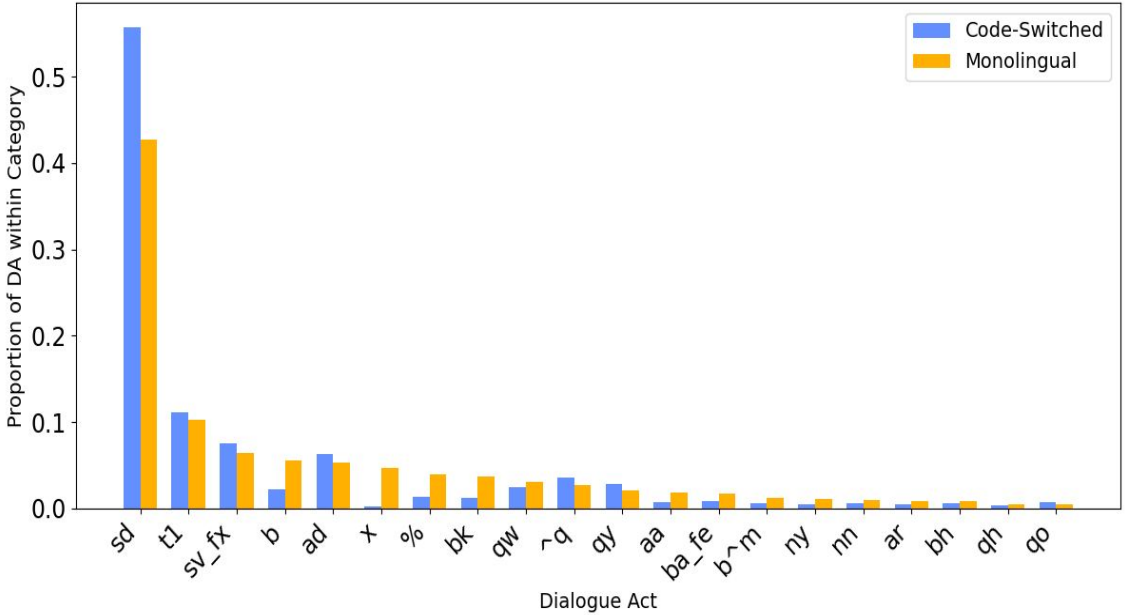



Fig. 4: Distribution of DA types across CSW and ML utterances.

# Results: Dialogue acts differ between code-switched and monolingual contexts

- DAs conveying factual information, providing explanations, confirming understanding → CSW
- DAs advancing a conversation, supporting thematic exploration → ML

Tab. 1: χ2 tests and odds ratios comparing DA distribution across CSW and ML utterances.

| DA type | χ2 | *p*-val | OR |
|---|---|---|---|
| **sd** | **119.2** | **<0.01** | **1.59** |
| t1 | 1.13 | – | 1.08 |
| sv-fx | 2.08 | – | 1.13 |
| **b** | **39.1** | **<0.01** | **0.41** |
| ad | 1.86 | – | 1.13 |
| **x** | **28.3** | **<0.01** | **0.11** |
| **%** | **38.7** | **<0.01** | **0.34** |
| **bk** | **35.8** | **<0.01** | **0.33** |
| **qw** | **4.48** | **<0.05** | **0.74** |
| **^q** | **4.81** | **<0.05** | **1.30** |

# Results: Dialogue acts differ between code-switched and monolingual contexts

- Clustering reinforces previous DA results
  - CSW in BM leans toward task-oriented, structured exchanges, with a focus on clarity and goal completion
  - ML speech is more varied, incorporating feedback and exploratory acts.

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

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

Tab. 3: Cluster composition by proportion: CSW (left) and ML (right) utterances.

# Results: Dialogue acts differ between code-switched and monolingual contexts

- Overall: conversational functions have *preferred modes of linguistic expression*

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

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

Tab. 3: Cluster composition by proportion: CSW (left) and ML (right) utterances.

# Results: Dialogue acts vary within types of CSW

- One-way ANOVA test results on CSW richness (CSW ratio, M-index) align with our previous DA groupings
  - DAs that are more likely to be expressed in CSW speech require smaller quantities of CSW (e.g. ^q; "*Mi mamá dijo* que she didn't know.")
  - DAs that are more likely to be expressed in ML speech require greater quantities of CSW (e.g. ad; "Okay, *ponlo ahí*.")
    - It seems there is a compensatory multilingual mechanism at play.

# Results: Dialogue acts vary within types of CSW

- Introspective and information-conveying DAs are more effectively expressed via shorter, simpler, insertional code-switches
  - "*Pues,* what was I saying…"
  - *"Para mí, es igual que cuando uno manda los* checks *al* IRS."

- DAs expressed with distinct, connected clauses are more effectively expressed via longer, more complex, alternational CSW
  - "*Cuándo vas a ver el apartamento* and how many bedrooms does it have?"

# Results: Dialogue acts vary within types of CSW

- Overall: different discourse functions support CSW of *varying quantity, syntactic structure, and complexity*

# Research Questions

1. How are patterns of **CSW** influenced by the **content** of the utterance?
2. How is **CSW** influenced by the **function** of the utterance?
    a. Does this **interact** with patterns stemming from the content of code-switched speech?

# Conclusions

1. Patterns of **CSW** are somewhat influenced by the discourse **content** of the utterance expressed via NEs
2. Discourse **function** encoded by DAs has a relatively greater influence on patterns of **CSW**
   - ➢ The relationships we discover are salient enough to be learned by transition models that predict CSW, which point to the two discourse aspects **interacting** in modeling scenarios

Paper

# Big Picture: How & why do speakers code-switch?

- **How?**
  - According to individual proficiency
  - In dynamic interaction with an interlocutor
  - In distinctive prosodic styles
- **Why?**
  - To convey empathy
  - To achieve specific conversational goals