How and Why Speakers Code-Switch: Insights for Naturalistic Multilingual Speech Generation

Debasmita Bhattacharya
Columbia University

Spring 2026

What is code-switching?

- **The alternation between one language variety and another**
  - In writing or in speech

Dice que le dieron ese ...

puesto de interno .

internship .

(= He said they gave him this internship.)

Why study code-switching?

- Most of the world speaks more than one language!
  - It is important to build language technologies that are robust to diverse linguistic settings

Why study code-switching?

- As a building block for developing generative speech technologies capable of producing naturalistic multilingual speech
  - People trust and prefer systems that resemble them
  - Particular importance in high-stakes domain areas
  - Serving previously underserved communities of speakers

Why study code-switching?

- As a building block for developing generative speech technologies capable of producing naturalistic multilingual speech
  - Applying insights gleaned in
    - training data curation recommendations, e.g. for SFT
    - priors for generative multilingual models
    - steering prompts for model controllability

How can we study code-switching?

Given naturally occurring, spontaneous speech:

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

- Draw on a variety of methods
  - ➢ Computational paralinguistics, acoustic-prosodic modeling, information theory…

How & why do speakers code-switch?

- **How?**
  - In distinctive prosodic styles

- **Why?**
  - To convey empathy
  - *Not only* to ease production

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

Debasmita Bhattacharya*, Michela Marchini*, Julia Hirschberg.

Under review at INTERSPEECH 2026



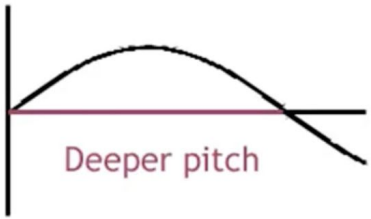
Motivation: prosodic production x code-switching

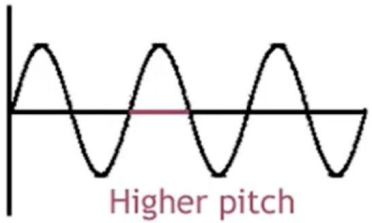
- Prior work on code-switching has effectively captured its morpho-syntactic, sociolinguistic, etc. characteristics
  - ➢ But this has primarily been tied to *writing*!

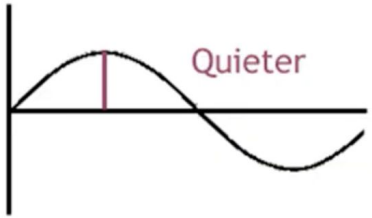
- **Prosody:** variations in pitch, loudness, tempo, stress that convey meaning (e.g. emotion, emphasis) in spoken language
  - ➢ It captures a uniquely *spoken* aspect of speech production
    - What can we learn about spoken code-switching that we can't discover from studying its transcripts alone?

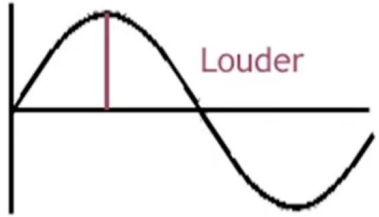
Research Questions

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?

Background: language proficiency

- **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 → e.g., years of experience
  - Language exposure aspects → e.g., medium of schooling

Background: quantifying code-switching

Code-switching richness metrics

If interested, ask me for formulae during Q&A at the end!

- M-index
  - How ***much*** code-switching is present
  - min = 0 → no code-switching
  - max = 1 → even distribution between languages
- I-index
  - How ***frequently*** do language varieties alternate
  - min = 0 → no code-switching
  - max = 1 → every switch point has a code-switch

Background: quantifying code-switching

- "***Andale pues* and do come again**"
  - M-index = 0.8
  - I-index = 0.2

es tokens: 2
en tokens: 4

- "***Bueno*, in other words, *el* flight *que sale de Chicago* around three o'clock"**
  - M-index = 0.95
  - I-index = 0.25

es tokens: 5
en tokens: 8

Background: strategies of code-switching

Code-switching syntactic complexity

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

Figure credit: Ahn et al., SCiL 2020

Background: strategies of code-switching

Code-switching syntactic complexity

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

Figure credit: Ahn et al., SCiL 2020

Background: strategies of code-switching

Code-switching syntactic complexity

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

Figure credit: Ahn et al., SCiL 2020

Data: Bangor Miami Spanish-English corpus

- **Conversational data:** recorded speech + transcripts
  - 103 prosodic features: 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 complexity: insertional = low; alternational = high

mostly dyadic dialogue, i.e., #speakers = 2

- **Questionnaire metadata**
  - Linguistic: age, years of experience, self-reported ability
  - Demographic: parents' primary language, medium of schooling

Method

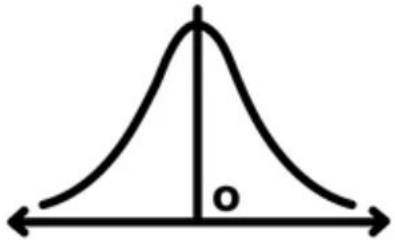
❖ Statistical analysis

- ➢ Compare prosodic feature distributions of
  - ■ Monolingual speech (*en* and *es*)
  - ■ Code-switched speech (*es-en*)

each speaker produces CSW *AND* monolingual speech, i.e., they appear on both sides of each comparison

❖ Modeling

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

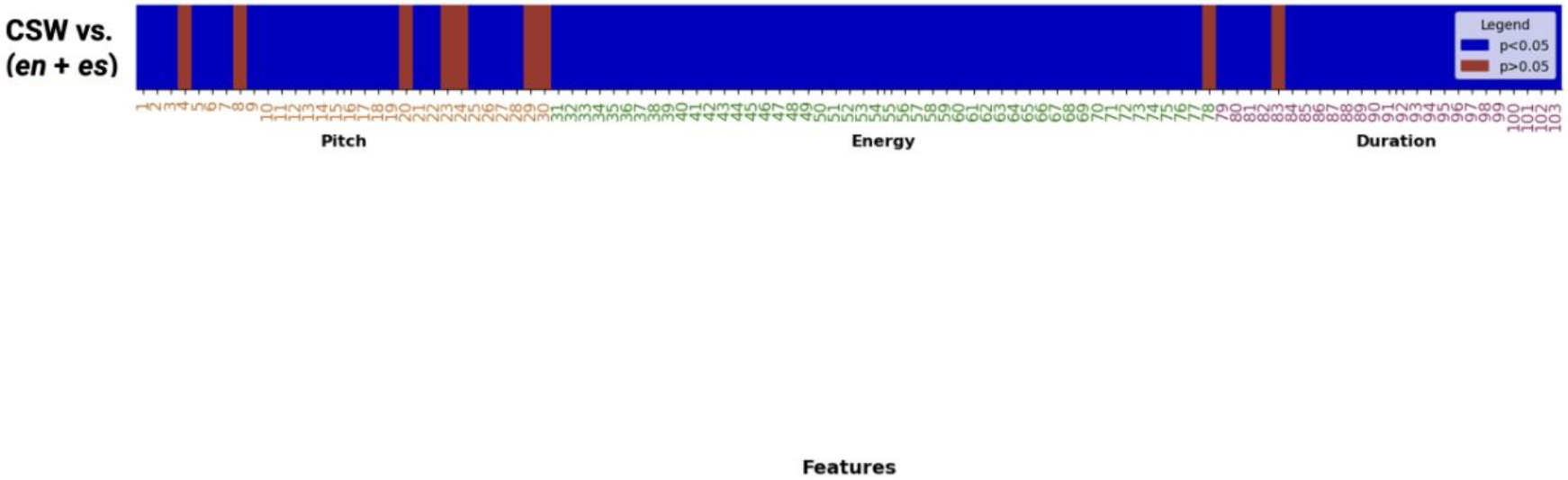


Research Questions

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

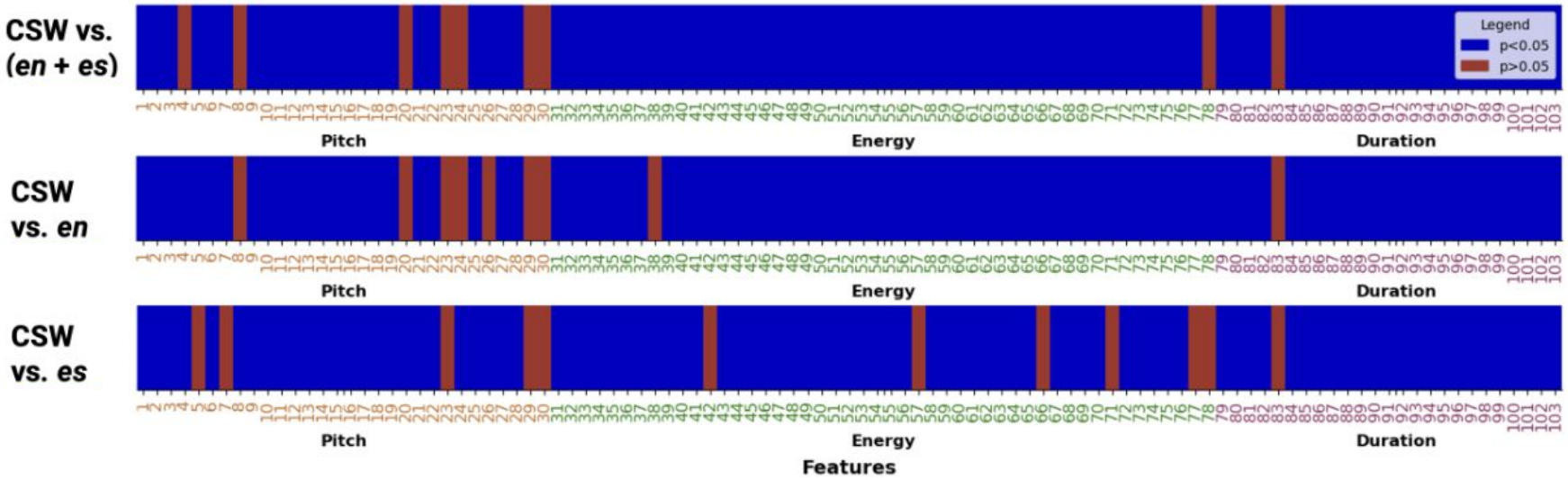
2. How is this variation **influenced** by (a) speaker **proficiency** and (b) **linguistic characteristics** of multilingual speech?

Results: Code-switching differs prosodically from monolingual speech



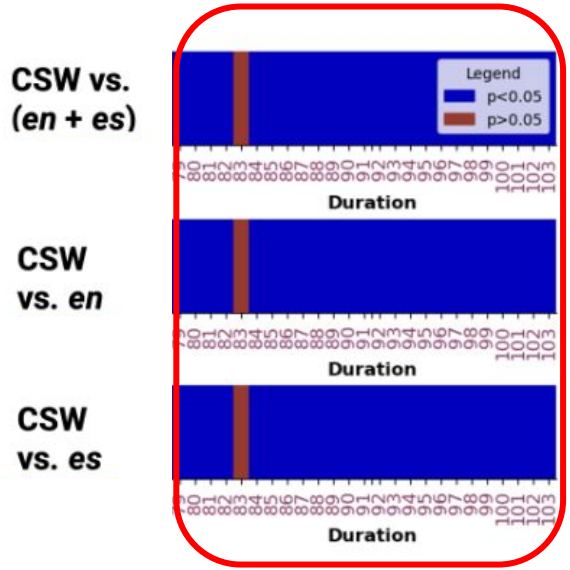
- Most features differ between code-switching and monolingual speech

Results: Code-switching differs prosodically from monolingual speech



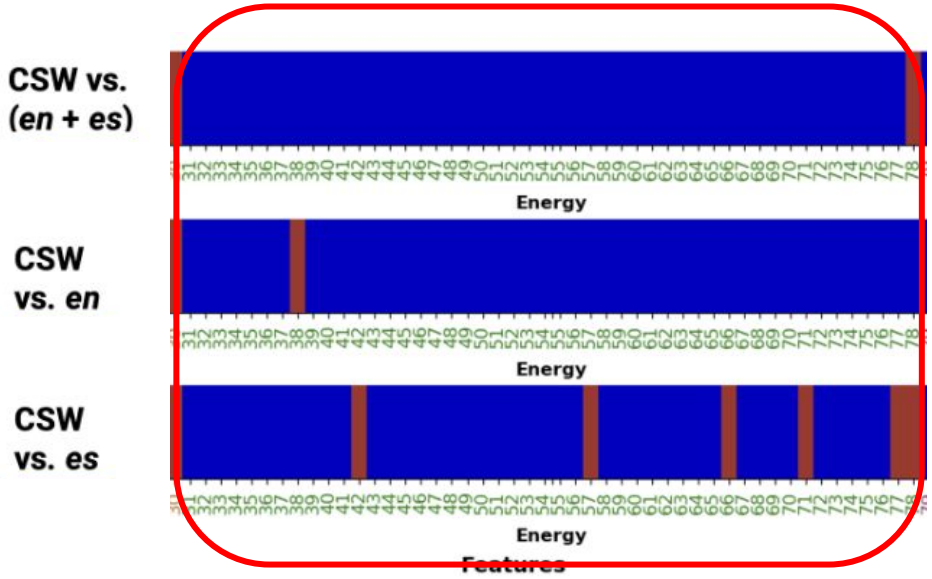
- Most features differ between code-switching and monolingual speech

Results: Code-switching differs prosodically from monolingual speech



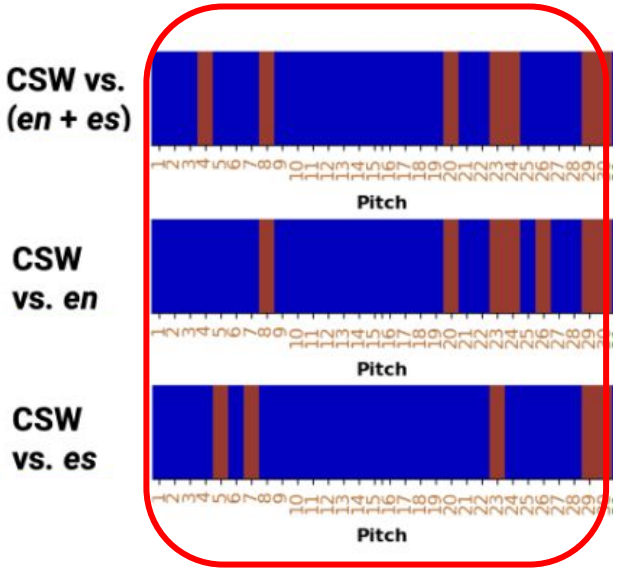
- > 96% of features differ!

Results: Code-switching differs prosodically from monolingual speech



- > 87.5% of features differ!

Results: Code-switching differs prosodically from monolingual speech



- >70% of features differ

Results: Code-switching differs prosodically from monolingual speech

	Pitch	Loudness	Rhythm
Code-switching	***Higher***	***Lower***	***Disjointed***
Monolingual	Lower	Higher	Fluid

Results: Code-switching differs prosodically from monolingual speech

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

- Prosodic differences are salient in clustering models

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: Prosodic variation in CSW is shaped by speaker-level language proficiency characteristics

Proficiency factor	**% similar: CSW vs. *es*** utterances	**% similar: CSW vs. *en*** utterances	**Δ(CSW vs. *es* – CSW vs. *en*)**
Secondary school: *es*	**26**	11	15%
Primary school: *en*	5	**14**	9%
Secondary school: *en*	7	**17**	10%

❖ Do speakers produce code-switched prosody that is ***more similar*** to the monolingual prosody of their ***higher proficiency*** language?

Results: Prosodic variation in CSW is shaped by speaker-level language proficiency characteristics

Proficiency factor	**% similar: CSW vs. *es*** utterances	**% similar: CSW vs. *en*** utterances	**Δ(CSW vs. *es* – CSW vs. *en*)**
Primary school: *es*	**28**	12	16%
Secondary school: *es*	**26**	11	15%
Primary school: *en*	5	**14**	9%
Secondary school: *en*	7	**17**	10%

Results: Prosodic variation in CSW is shaped by speaker-level language proficiency characteristics

Proficiency factor	**% similar: CSW vs. *es*** utterances	**% similar: CSW vs. *en*** utterances	**Δ(CSW vs. *es* – CSW vs. *en*)**
Primary school: *es*	**28**	12	16%
Secondary school: *es*	**26**	11	15%
Primary school: *en*	5	**14**	9%
Secondary school: *en*	7	**17**	10%

Results: Prosodic variation in CSW is shaped by speaker-level language proficiency characteristics

Proficiency factor	**% similar: CSW vs. *es*** utterances	**% similar: CSW vs. *en*** utterances	**Δ(CSW vs. *es* – CSW vs. *en*)**
Primary school: *es*	**28**	12	16%
Secondary school: *es*	**26**	11	15%
Primary school: *en*	5	**14**	9%
Secondary school: *en*	7	**17**	10%

❖ Language proficiency drives prosodic differences

Results: Prosodic variation in CSW is shaped by speaker-level language proficiency characteristics

- Do speakers produce code-switched prosody that is ***more similar*** to the monolingual prosody of their ***higher proficiency*** language?

- Binary LID task → predict language of code-switched utterance
  - **Input**: prosodic features of code-switched utterances
  - **Output**: [*en, es*]
  - **Hypotheses**
    - Input: CSW by *es*-proficient speaker → Output: *es*
    - Input: CSW by *en*-proficient speaker → Output: *en*

Results: Prosodic variation in CSW is shaped by speaker-level language proficiency characteristics

Model	Accuracy	F1-Score	
		en	*es*
`Whisper-base` (PT)	0.64	0.74	0.44
`Whisper-base` (FT)	0.83	0.86	0.77
`Whisper-base` (FT + CH)	0.92	0.93	0.89

Results: Prosodic variation in CSW is shaped by speaker-level language proficiency characteristics

Model	Accuracy	F1-Score	
		en	*es*
`Whisper-base` (PT)	0.64	0.74	0.44
`Whisper-base` (FT)	0.83	0.86	0.77
`Whisper-base` (FT + CH)	0.92	0.93	0.89

Results: Prosodic variation in CSW is shaped by speaker-level language proficiency characteristics

Model	Accuracy	F1-Score	
		en	*es*
`Whisper-base` (PT)	0.64	0.74	0.44
`Whisper-base` (FT)	0.83	0.86	0.77
`Whisper-base` (FT + CH)	0.92	0.93	0.89

Results: Prosodic variation in CSW is shaped by speaker-level language proficiency characteristics

Model	Accuracy	F1-Score *en*	F1-Score *es*
`Whisper-base` (PT)	0.64	0.74	0.44
`Whisper-base` (FT)	0.83	0.86	0.77
`Whisper-base` (FT + CH)	0.92	0.93	0.89

- Language proficiency drives prosodic differences

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?

🔝 »

How & why do speakers code-switch?

- **How?**
  - In distinctive prosodic styles

- **Why?**
  - To convey empathy
  - *Not only* to ease production

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

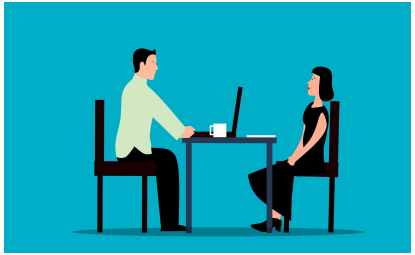
- **Empathy** → the ability to…
  - understand other people's feelings as if one were having them oneself, and
  - respond accordingly
- Characterized by, e.g. (among others):
  - slow, soft, low-pitched, even-toned speech
  - focus on thoughts, feelings, the interlocutor in general
    - "I feel that way too sometimes" ≠ "I attend class on Tuesdays"
      empathetic neutral

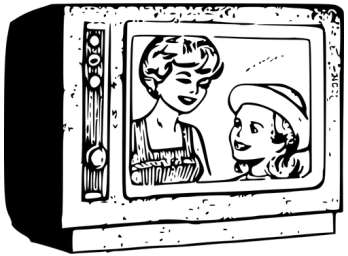
Research Questions

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 informal speech corpora

- **Spanish-English**: conversational
  - Bangor Miami
  - Mostly =2 speakers, some ≥3 speakers
- **Mandarin-English**: conversational
  - SEAME
  - All =2 speakers
- **Hindi-English**: soap opera
  - MaSaC
  - All =2 speakers





Method

- **Approximate utterance-level 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 utterance-level code-switching metrics**
  - M-index & I-index
  - Code-switched vs. monolingual
- **Statistical analysis**
  - Relationship between code-switching and empathy
    - Strength of relationship
    - Linearity of relationship

Method

- **Approximate utterance-level 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 utterance-level code-switching metrics**
  - M-index & I-index
  - Code-switched vs. monolingual
- **Statistical analysis**
  - Relationship between code-switching and empathy
    - Strength of relationship
    - Linearity of relationship

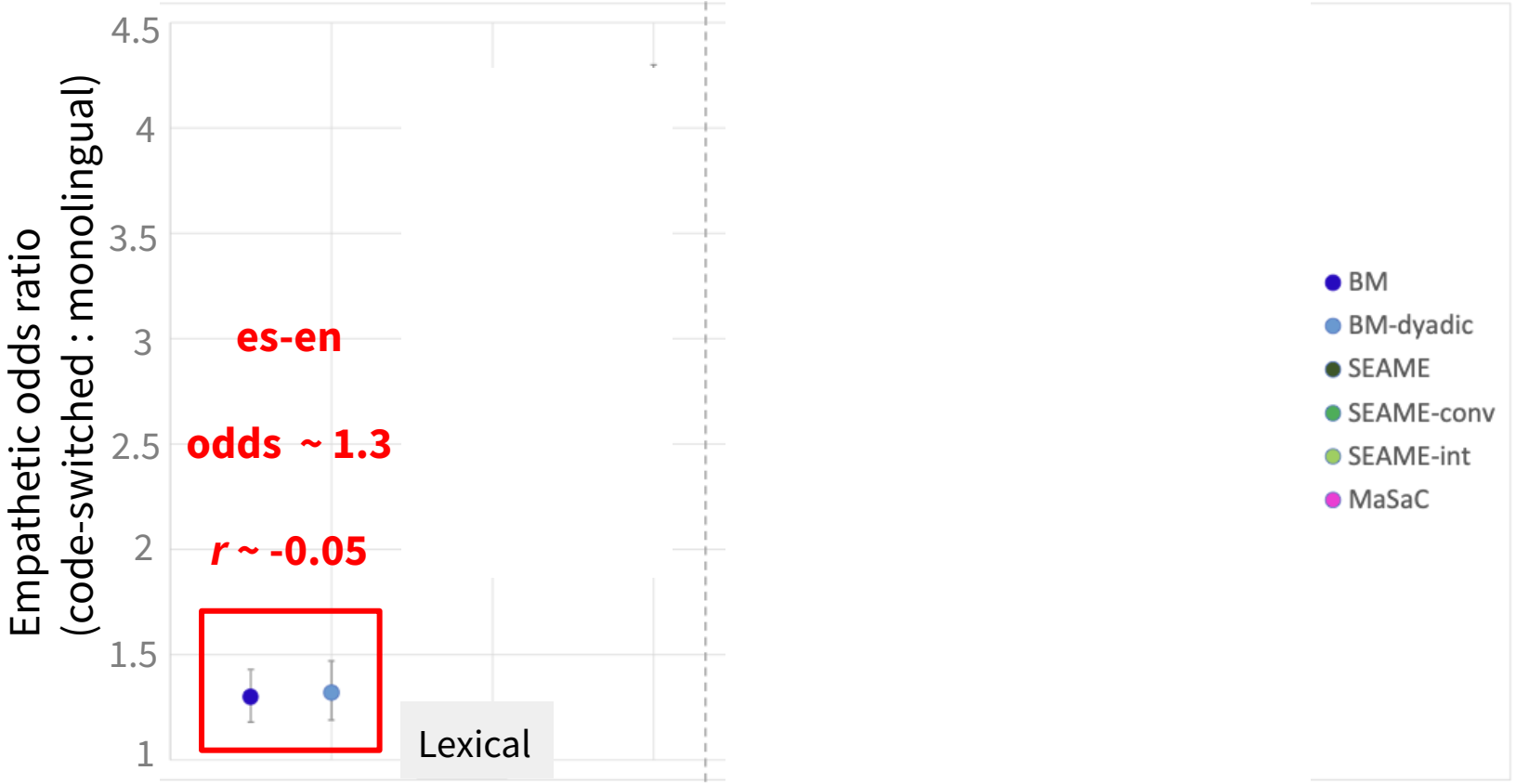
Method

- **Approximate utterance-level 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 utterance-level code-switching metrics**
  - M-index & I-index
  - Code-switched vs. monolingual
- **Statistical analysis**
  - Relationship between code-switching and empathy
    - Strength of relationship
    - Linearity of relationship

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?

Results: spoken CSW aligns with lexical correlates of empathy

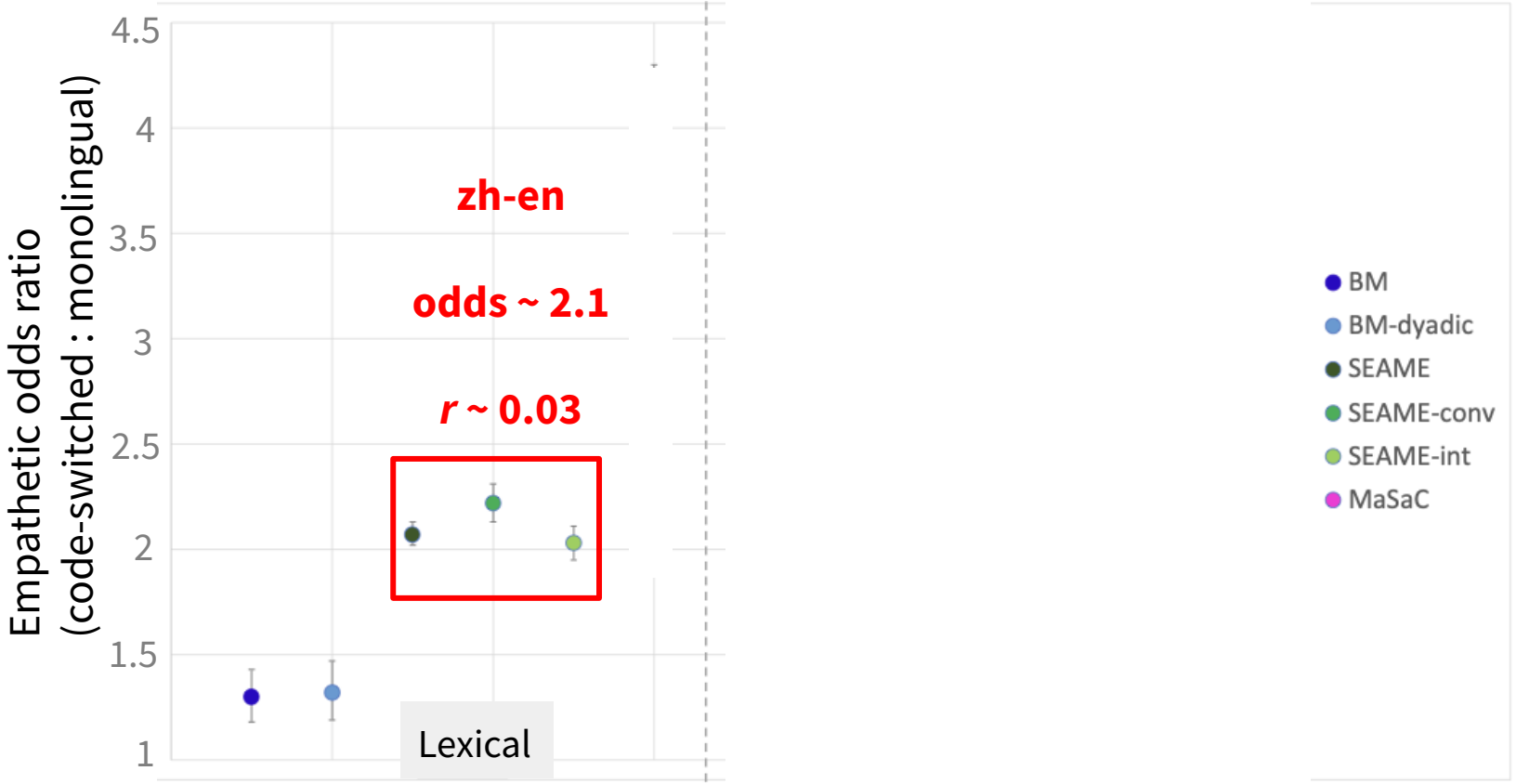


Results: spoken CSW aligns with lexical correlates of empathy

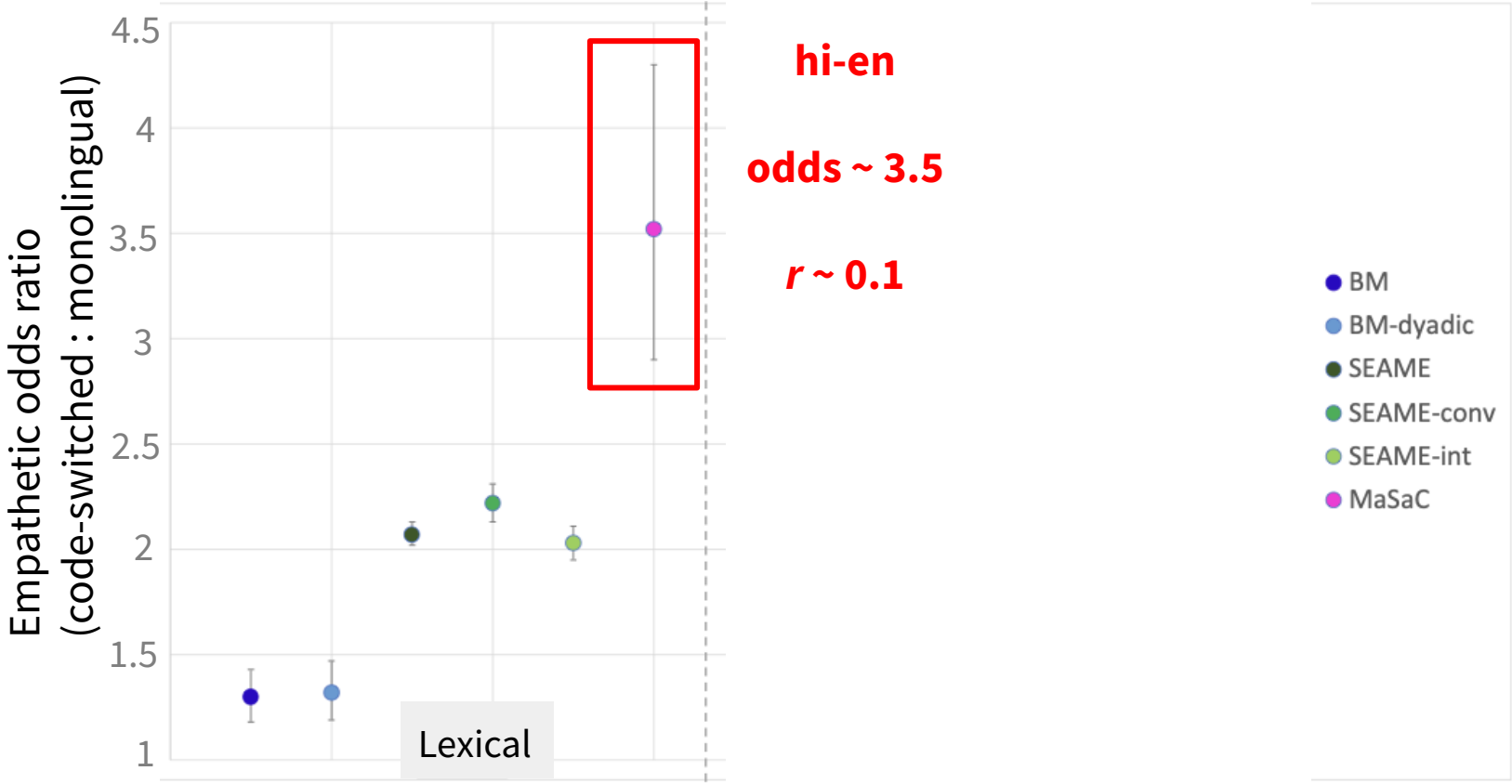
Pero ahora mismo tú me estabas diciendo que te gustaba y te parecía nice

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

Results: spoken CSW aligns 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	≈ 0
SEAME	~2.1	≈ 0
MaSaC	~3.5	≈ 0

} Positive, non-linear association

Results: spoken CSW aligns with lexical correlates of empathy

	Odds CSW:ML	CSW∝empathy
Bangor Miami	~1.3	≈ 0
SEAME	~2.1	≈ 0
MaSaC	~3.5	≈ 0

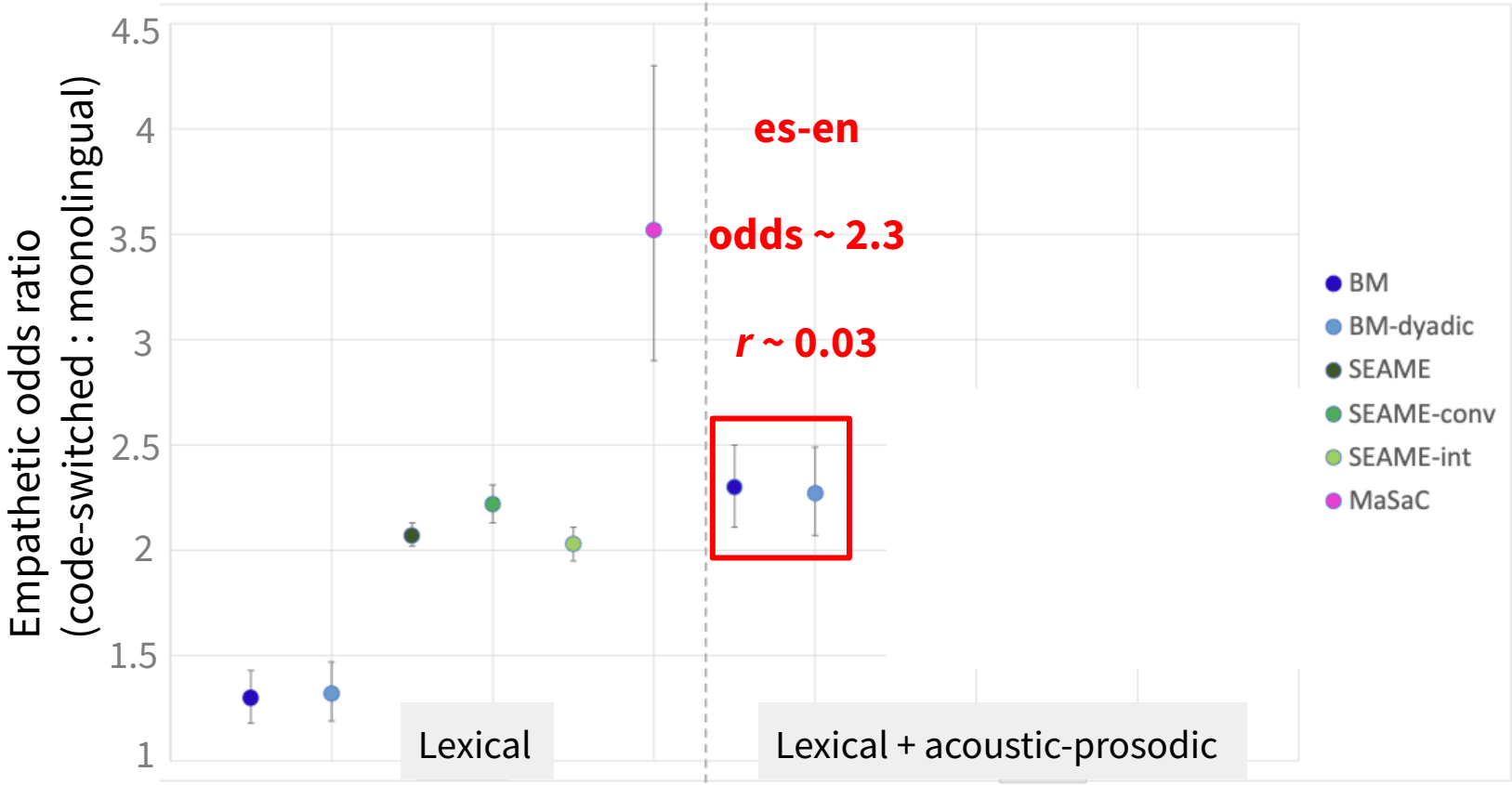
} Positive, non-linear association

RQ2: Does the answer to **1. *generalize across language pairs*** involving different language families? ✅

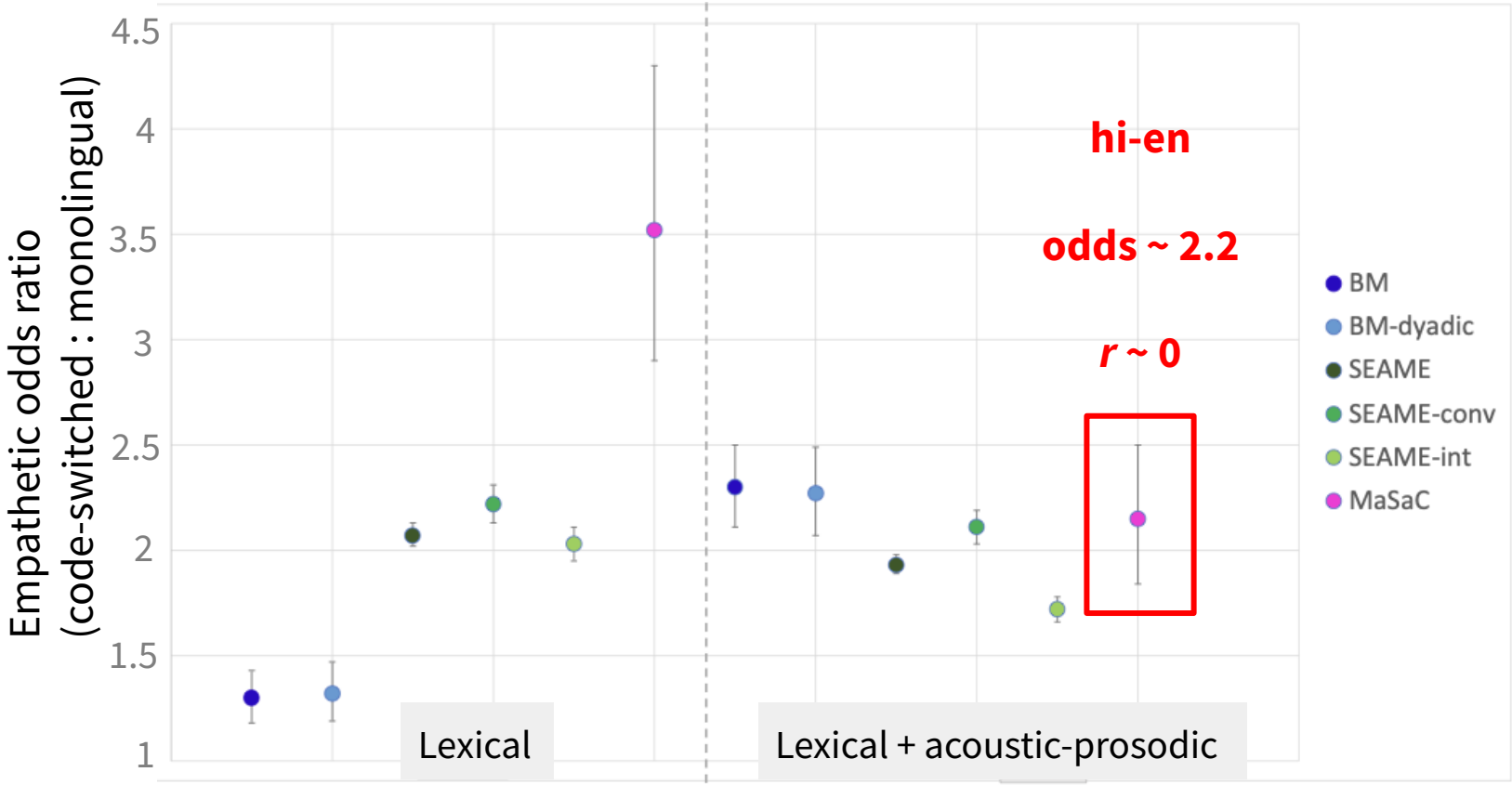
Research Questions

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?

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

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*

Jitter: 2.1%

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

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	≈ 0	~2.3	≈ 0
SEAME	~2.1	≈ 0	~1.9	≈ 0
MaSaC	~3.5	≈ 0	~2.1	= 0

Positive, non-linear association

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	≈ 0	~2.3	≈ 0
SEAME	~2.1	≈ 0	~1.9	≈ 0
MaSaC	~3.5	≈ 0	~2.1	= 0

RQ2: Does the answer to **1. *generalize across language pairs*** involving different language families? 🤷🏽‍♀️

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? 🤔

❖ 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.

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

Paper

How & why do speakers code-switch?

- **How?**
  - In distinctive prosodic styles

- **Why?**
  - To convey empathy
  - *Not only* to ease production

Code-switching in text and speech challenges information-theoretic speaker design

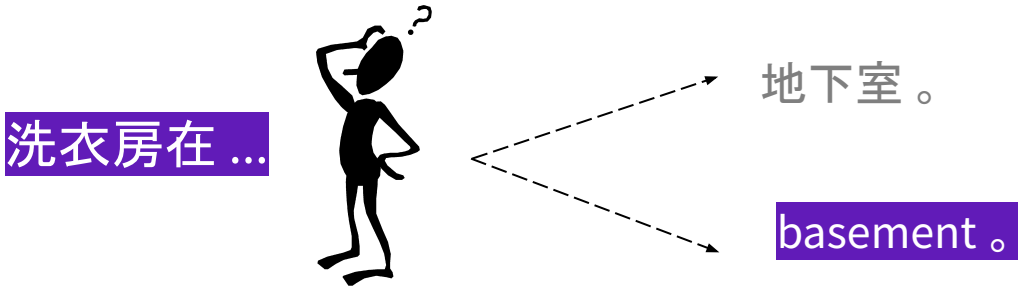
Debasmita Bhattacharya and Marten van Schijndel.

Accepted for publication at Language and Cognition

Motivation: information load in CSW

- Bilinguals are known to code-switch in writing and speech
- Is this ***purely*** to make language production easier?
  - ➢ Motivation: **Speaker-**centric; Mechanism: **less** info. load
- Or is there a potential additional signaling benefit for the listener?
  - ➢ Motivation: **Audience** design(?); Mechanism: **more**(?) info. load
- Can information load analysis help to answer these questions?
  - ➢ Key idea: ⬆️ communicative function ⚖️ ⬇️ listener overload

Background: predictability x code-switching

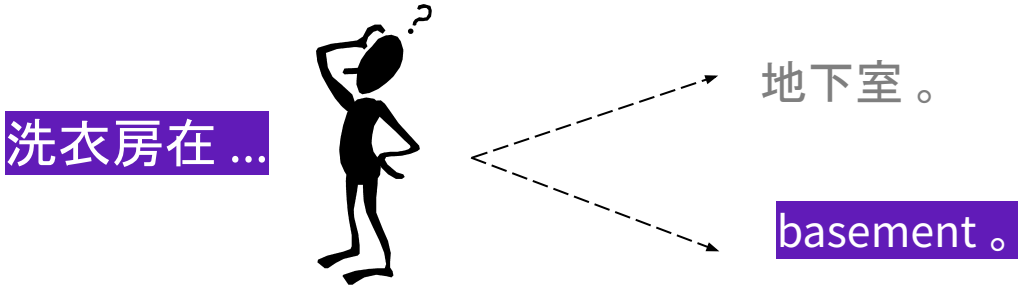


(= The laundry room is in the basement.)

CSW (ZH→EN):

At code-switch site:
pred(ZH) <<< 1

Background: predictability x code-switching



(= The laundry room is in the basement.)

CSW (ZH→EN):

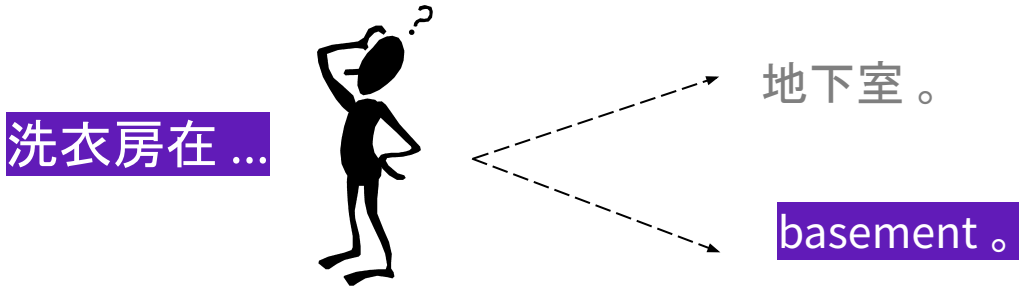
At code-switch site:
pred(ZH) <<< 1

ZH (trans):

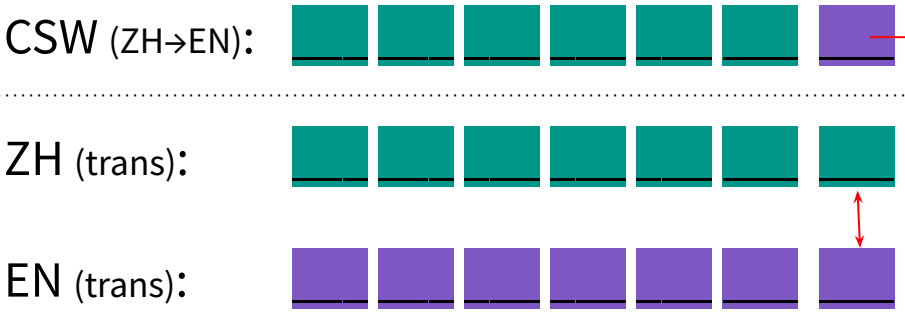
EN (trans):

Speaker Design:
pred(ZH) < pred(EN)
⟹ code-switch from ZH to EN

Background: predictability x code-switching



(= The laundry room is in the basement.)



At code-switch site:
pred(ZH) <<< 1

Audience Design:
pred(ZH) ≥? pred(EN)
⟹ code-switch from ZH to EN

Research Questions

1. Does the influence of predictability on code-switching reflect ***only* speaker-centric pressures**?

2. Does the influence of predictability on code-switching **generalize** across both **writing and speech**?

$$surp(w_i) \quad = \quad -logP(w_i|w_{i-1}, ..., w_{i-t})$$

Code-switched corpora

- **SEAME**

- **Calvillo et al. (2020) corpus**
  - CSW (ZH→EN)
  - ZH (CSW trans)
  - *EN (CSW **trans**)*
  - ZH (Non-CSW **control**)
  - *EN (Non-CSW **trans**)*

Sentence type	Sentence sub-type	Example	Notes
(1) CS [orig.]	--	你 是 full-time 学生 吗 ?	CS1
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(2) CS [trans.]	--	你 是 全日制 学生 吗 ?	
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(3) Non-CS	--	是 里程 票 吗 ?	CS1 Non-CS
	Gloss	Is mileage ticket *part.* ?	
	Translation	*Is it a mileage ticket?*	

Code-switched corpora

- SEAME

- **Calvillo et al. (2020) corpus**
  - CSW (ZH→EN)
  - ZH (CSW trans)
  - *EN (CSW **trans**)*
  - ZH (Non-CSW **control**)
  - *EN (Non-CSW **trans**)*

Sentence type	Sentence sub-type	Example	Notes
(1) CS [orig.]	--	你 是 full-time 学生 吗 ？	CS1
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(2) CS [trans.]	--	你 是 全日制 学生 吗 ？	
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(3) Non-CS	--	是 里程 票 吗 ？	CS1 Non-CS
	Gloss	Is mileage ticket *part.* ?	
	Translation	*Is it a mileage ticket?*	

Code-switched corpora

- SEAME

- **Calvillo et al. (2020) corpus**
  - CSW (ZH→EN)
  - ZH (CSW trans)
  - *EN (CSW **trans**)*
  - ZH (Non-CSW **control**)
  - *EN (Non-CSW **trans**)*

Sentence type	Sentence sub-type	Example	Notes
(1) CS [orig.]	--	你 是 full-time 学生 吗 ？	CS1
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(2) CS [trans.]	--	你 是 全日制 学生 吗 ？	
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(3) Non-CS	--	是 里程 票 吗 ？	CS1 Non-CS
	Gloss	Is mileage ticket *part.* ?	
	Translation	*Is it a mileage ticket?*	

Code-switched corpora

- SEAME

- **Calvillo et al. (2020) corpus**
  - CSW (ZH→EN)
  - ZH (CSW trans)
  - *EN (CSW **trans**)*
  - ZH (Non-CSW **control**)
  - *EN (Non-CSW **trans**)*

Sentence type	Sentence sub-type	Example	Notes
(1) CS [orig.]	--	你 是 full-time 学生 吗 ？	CS1
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(2) CS [trans.]	--	你 是 全日制 学生 吗 ？	
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(3) Non-CS	--	是 里程 票 吗 ？	CS1 Non-CS
	Gloss	Is mileage ticket *part.* ?	
	Translation	*Is it a mileage ticket?*	

Code-switched corpora

- SEAME

- **Calvillo et al. (2020) corpus**
  - CSW (ZH→EN)
  - ZH (CSW trans)
  - *EN (CSW **trans**)*
  - ZH (Non-CSW **control**)
  - *EN (Non-CSW **trans**)*

Sentence type	Sentence sub-type	Example	Notes
(1) CS [orig.]	--	你 是 full-time 学生 吗 ？	CS1
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(2) CS [trans.]	--	你 是 全日制 学生 吗 ？	
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(3) Non-CS	--	是 里程 票 吗 ？	CS1 Non-CS
	Gloss	Is mileage ticket *part.* ?	
	Translation	*Is it a mileage ticket?*	

Code-switched corpora

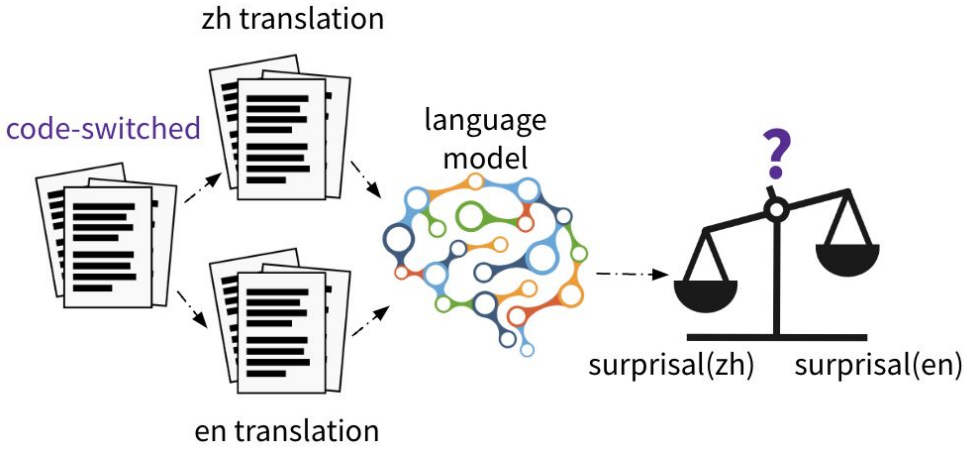
- **SEAME**

- **Calvillo et al. (2020) corpus**
  - 1.5k sents:
    - CSW (ZH→EN)
    - ZH (CSW trans)
    - *EN (CSW **trans**)*
  - 1.5k sents:
    - ZH (Non-CSW **control**)
    - *EN (Non-CSW **trans**)*

Sentence type	Sentence sub-type	Example	Notes
(1) CS [orig.]	--	你 是 full-time 学生 吗 ？	CS1
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(2) CS [trans.]	--	你 是 全日制 学生 吗 ？	
	Gloss	You are full-time student *part.* ?	
	Translation	*Are you a full-time student?*	
(3) Non-CS	--	是 里程 票 吗 ？	CS1 Non-CS
	Gloss	Is mileage ticket *part.* ?	
	Translation	*Is it a mileage ticket?*	

Method

- Train monolingual language models to infer surprisal statistics for (monolingually translated) Mandarin-to-English code-switched corpora (written and spoken).
  - 5-gram LM (*en* & *zh*)
  - GPT-2 (*en* & *zh*)



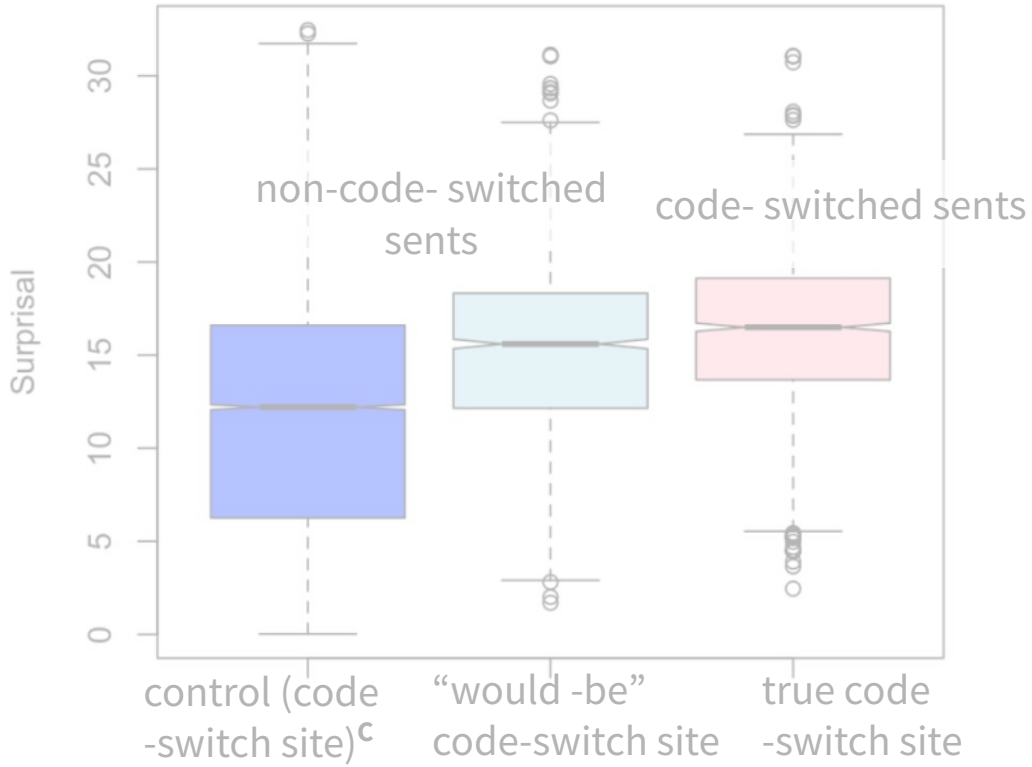
$$surp(w_i) = -logP(w_i|w_{i-1}, ..., w_{i-t})$$

Research Questions

1. Does the influence of predictability on code-switching reflect ***only* speaker-centric pressures**?

2. Does the influence of predictability on code-switching **generalize** across both **writing and speech**?

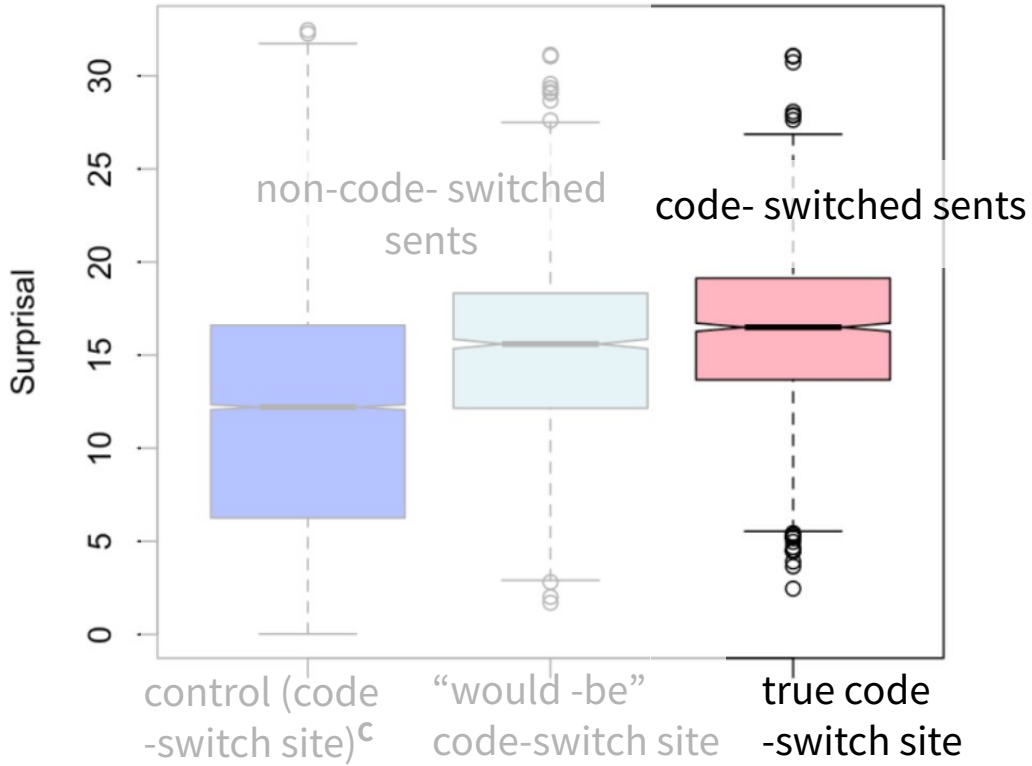
Results: Code-switching occurs in regions of low predictability



All comparison data are in ***Mandarin***.

We determine significance using Welch’s *t*-tests.

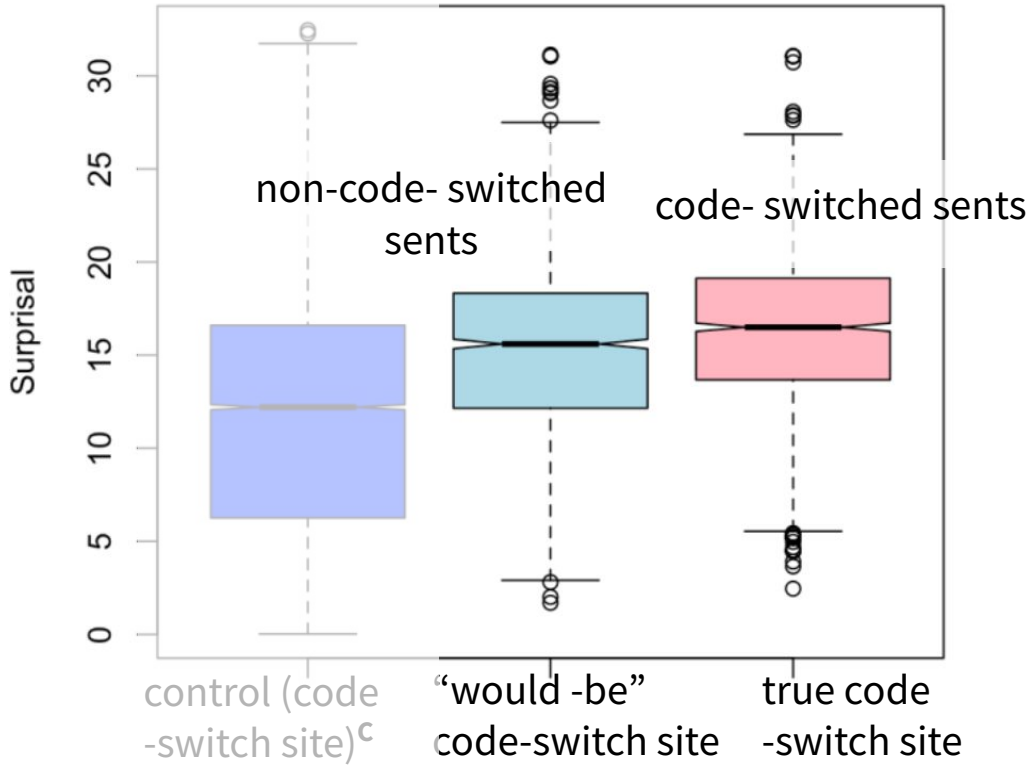
Results: Code-switching occurs in regions of low predictability



All comparison data are in ***Mandarin***.

We determine significance using Welch’s *t*-tests.

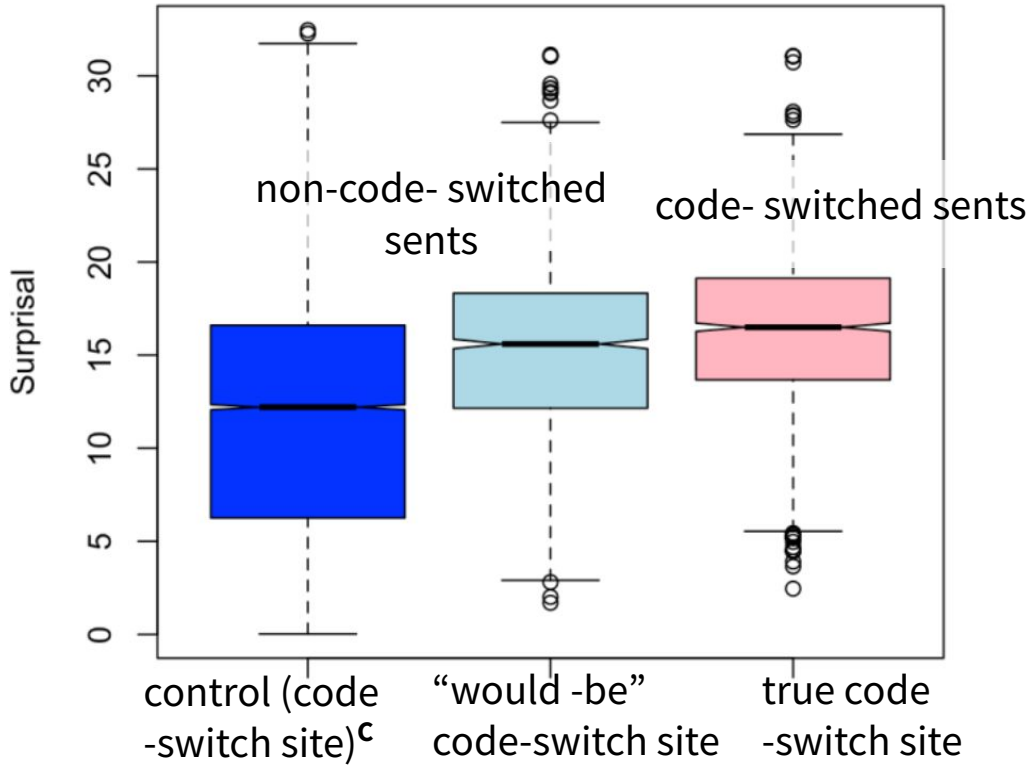
Results: Code-switching occurs in regions of low predictability



All comparison data are in ***Mandarin***.

We determine significance using Welch’s *t*-tests.

Results: Code-switching occurs in regions of low predictability



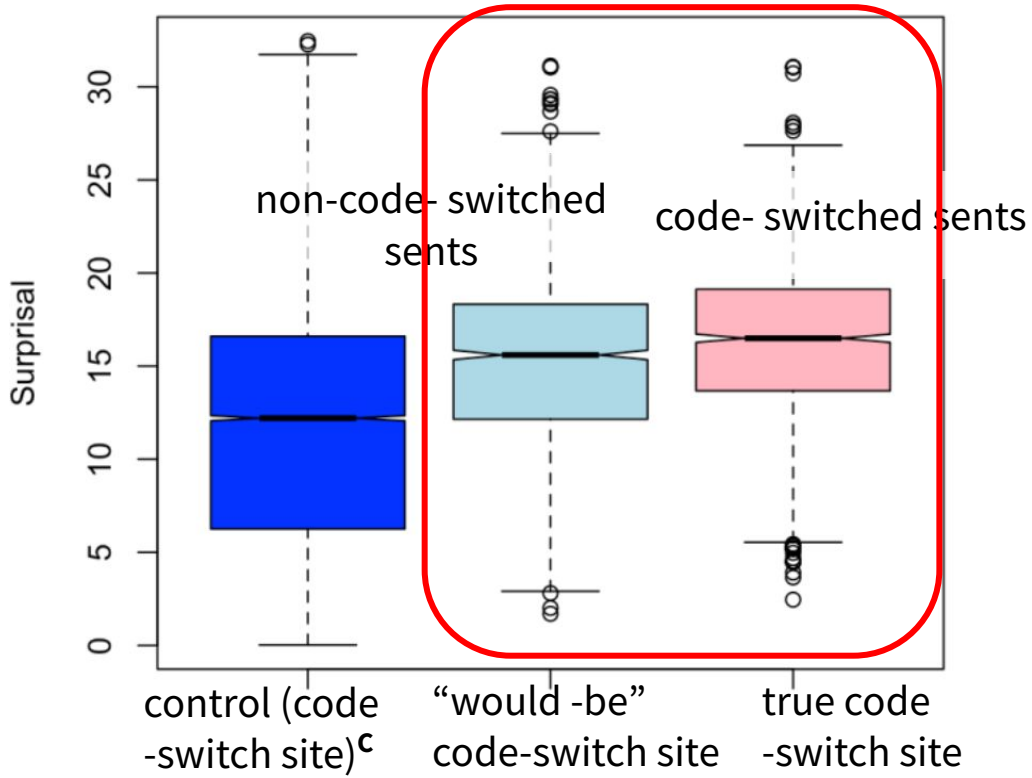
All comparison data are in ***Mandarin***.

We determine significance using Welch’s *t*-tests.

Results: Code-switching occurs in regions of low predictability

Sentence type	Sentence sub-type	Example
(1) CS [orig.]	--	你 是 full-time 学生 吗 ?
	Gloss	You are full-time student *part.* ?
	Translation	*Are you a full-time student?*
(2) CS [trans.]	--	你 是 全日制 学生 吗 ?
	Gloss	You are full-time student *part.* ?
	Translation	*Are you a full-time student?*
(3) Non-CS	--	是 里程 票 吗 ?
	Gloss	Is mileage ticket *part.* ?
	Translation	*Is it a mileage ticket?*

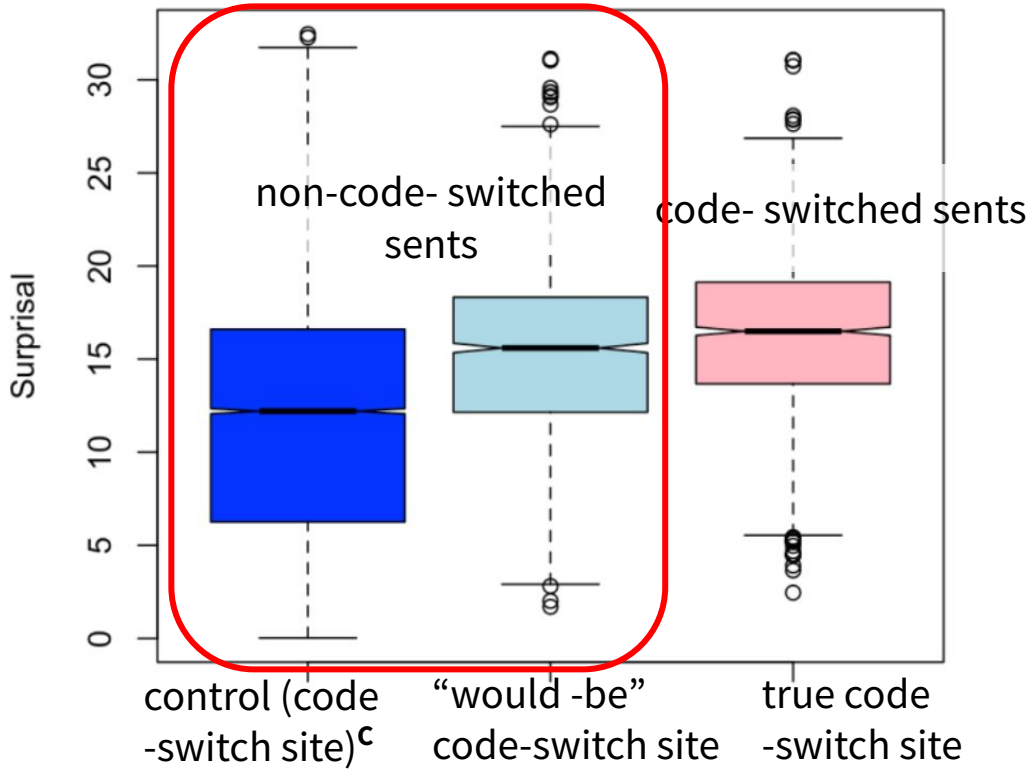
Results: Code-switching occurs in regions of low predictability



All comparison data are in ***Mandarin***.

We determine significance using Welch’s *t*-tests.

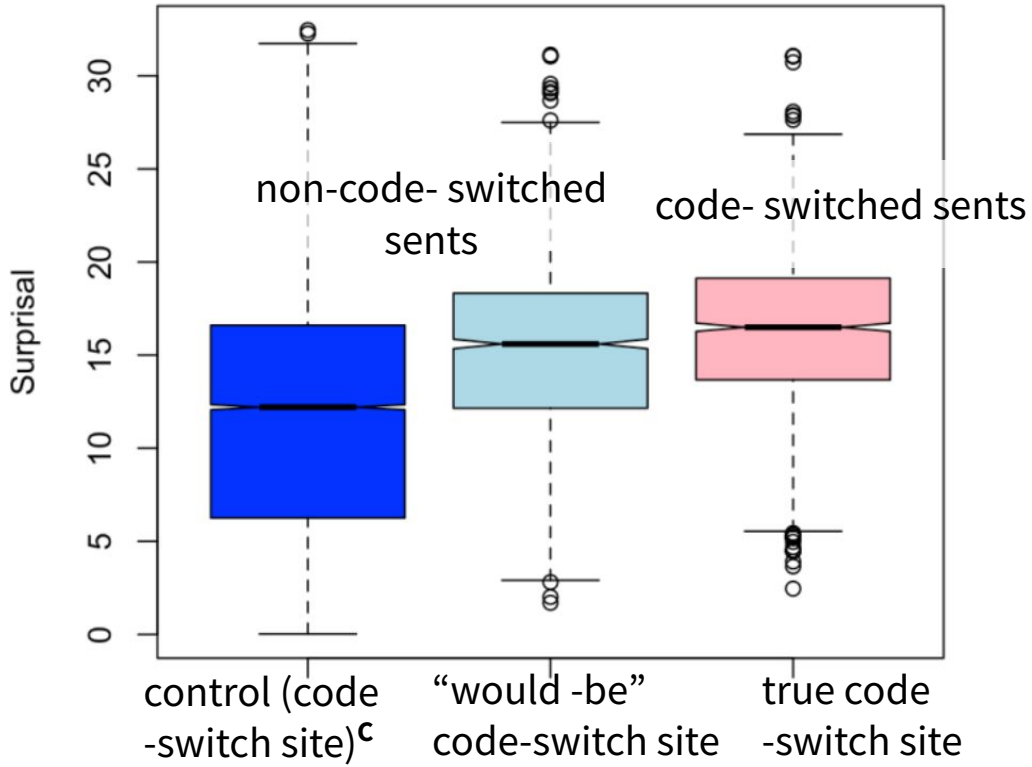
Results: Code-switching occurs in regions of low predictability



All comparison data are in ***Mandarin***.

We determine significance using Welch’s *t*-tests.

Results: Code-switching occurs in regions of low predictability



All comparison data are in ***Mandarin***.

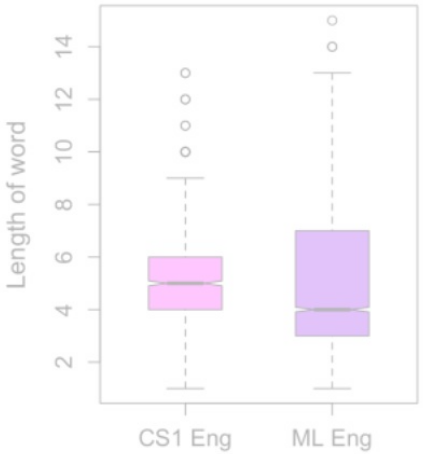
We determine significance using Welch’s *t*-tests.

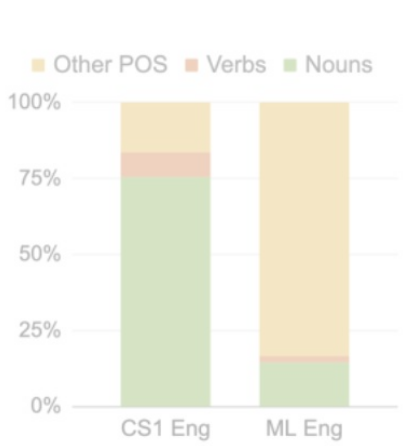
Research Questions

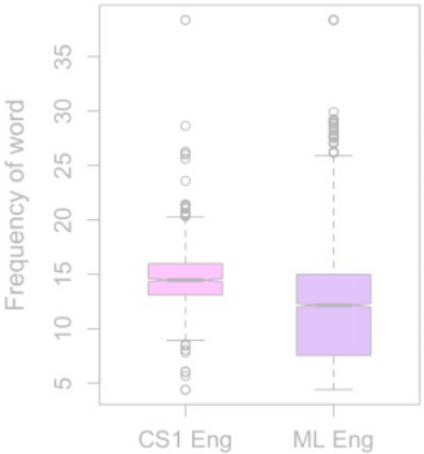
1. Does the influence of predictability on code-switching reflect ***only* speaker-centric pressures**?

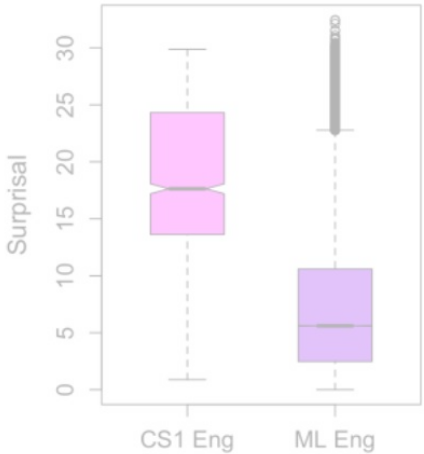
2. Does the influence of predictability on code-switching **generalize** across both **writing and speech**?

Results: Code-switching has information-theoretic components that are not solely speaker-driven





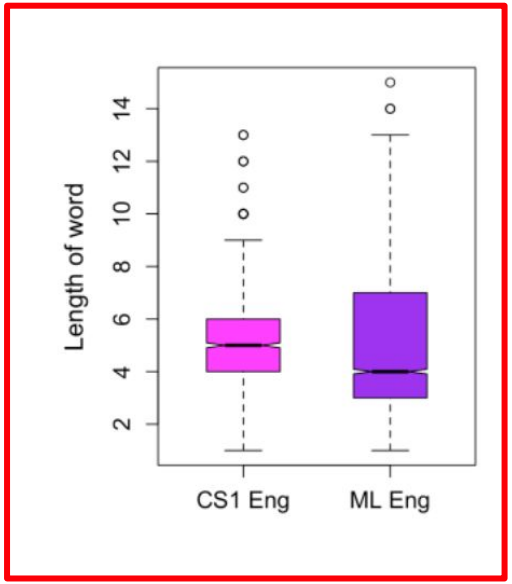


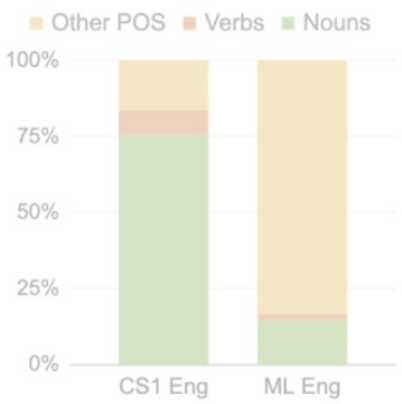


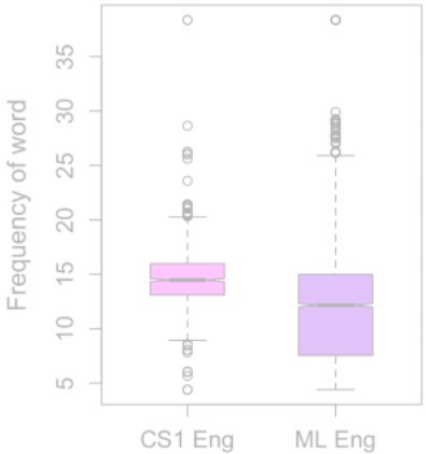
All comparison data are in ***English***.

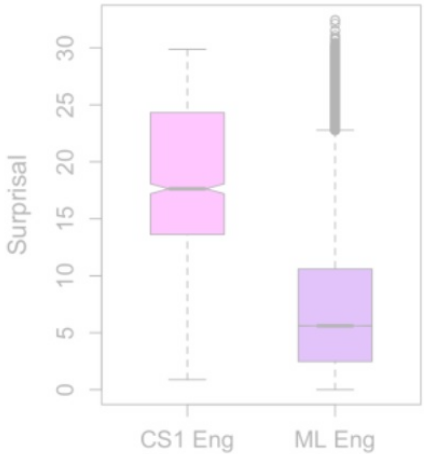
We determine significance using Welch's *t*-tests.

Results: Code-switching has information-theoretic components that are not solely speaker-driven





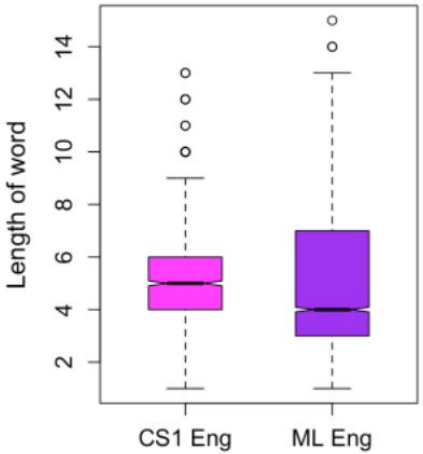


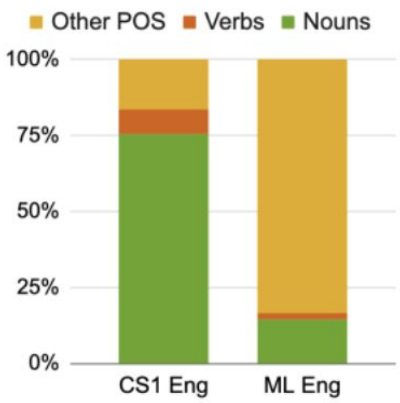


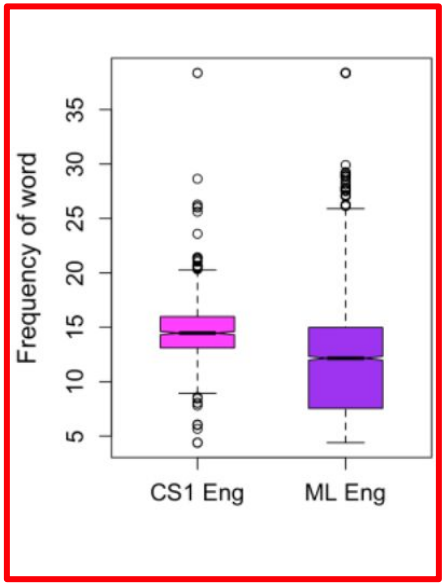
All comparison data are in ***English***.

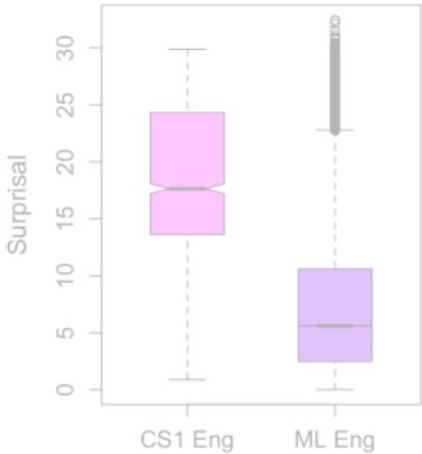
We determine significance using Welch's *t*-tests.

Results: Code-switching has information-theoretic components that are not solely speaker-driven





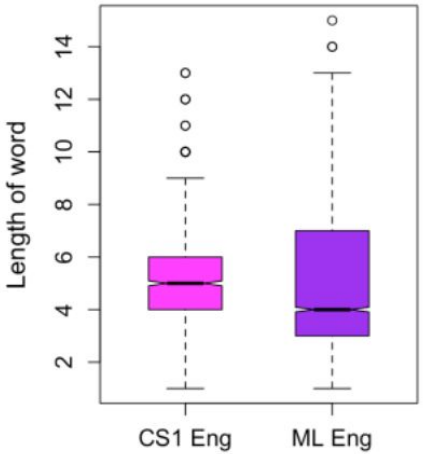


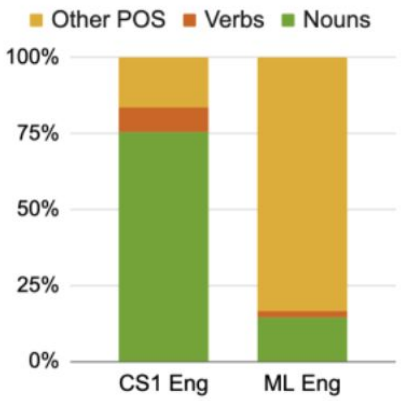


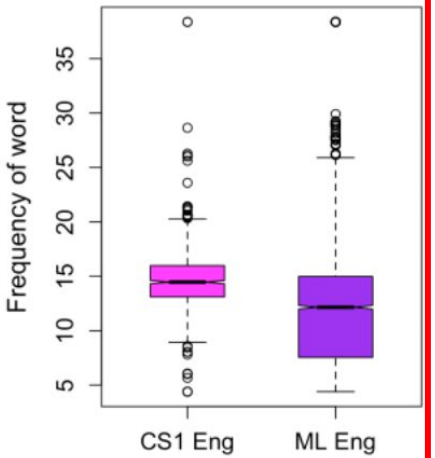
All comparison data are in ***English***.

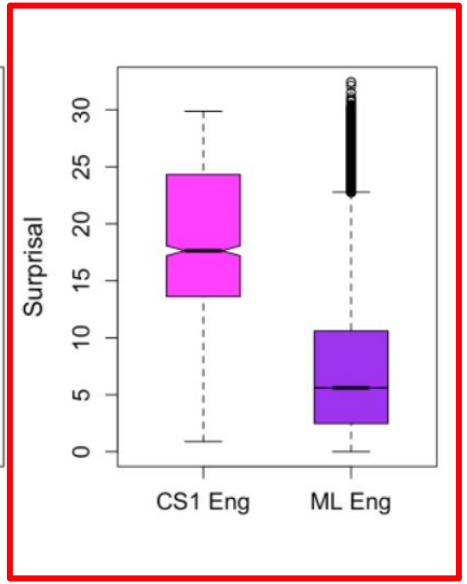
We graph $-\log_2(\text{frequency})$ ≡unigram surprisal
⟹ large frequency value ~ less common word

Results: Code-switching has information-theoretic components that are not solely speaker-driven





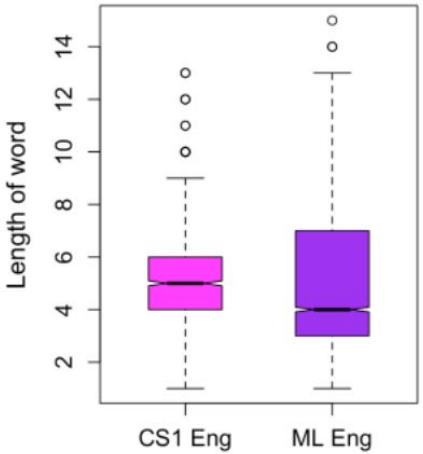


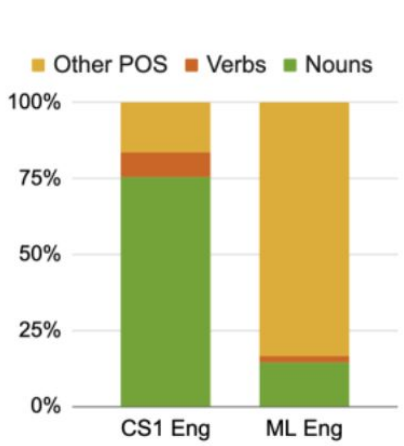


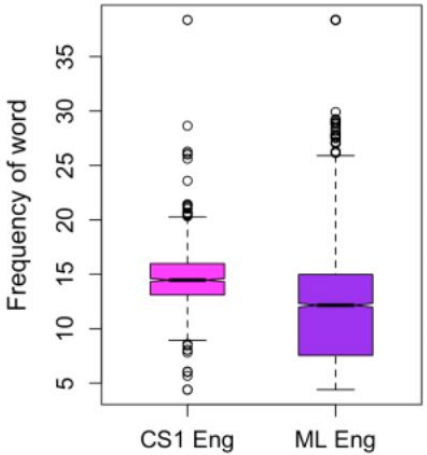
All comparison data are in ***English***.

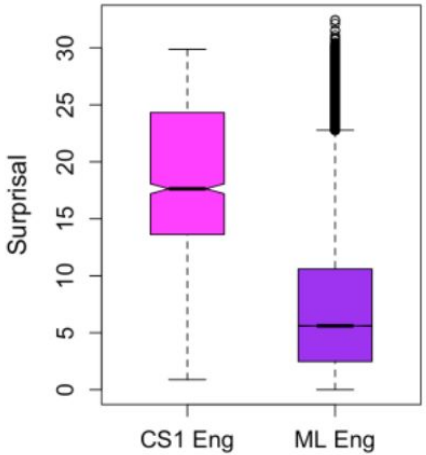
We determine significance using Welch's *t*-tests.

Results: Code-switching has information-theoretic components that are not solely speaker-driven



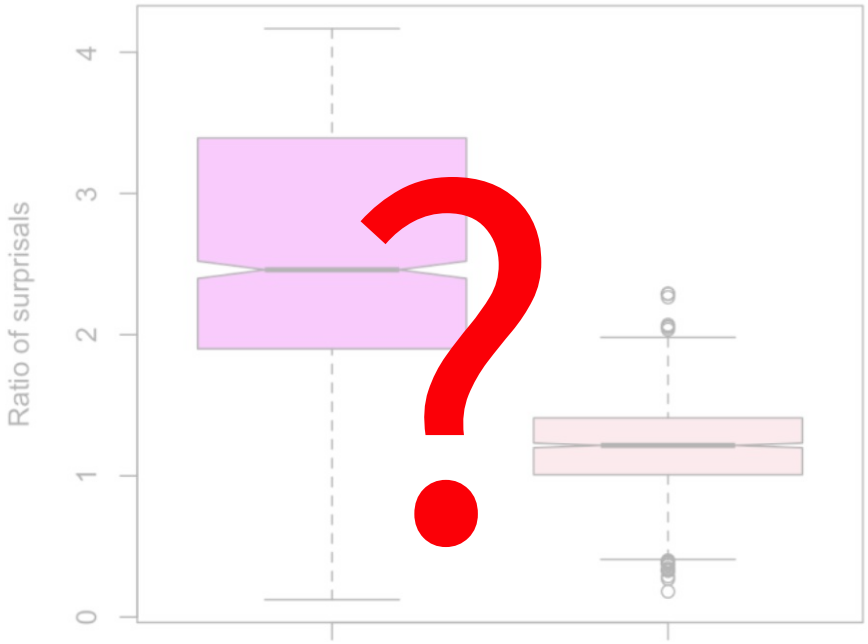




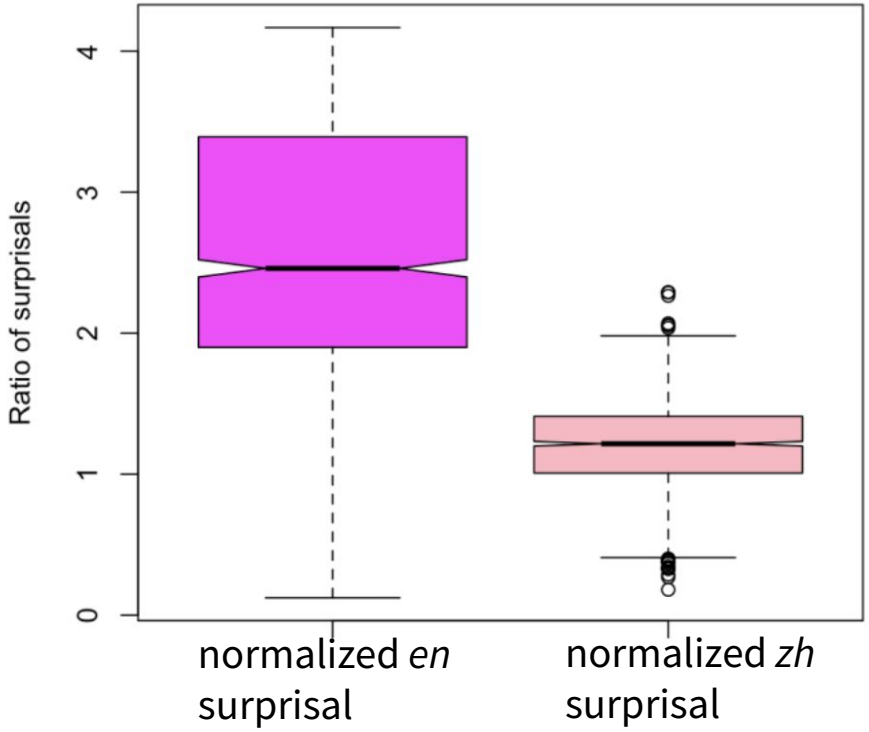


- **CS1 English is more complex than monolingual English**

Results: Code-switching has information-theoretic components that are not solely speaker-driven



Results: Code-switching has information-theoretic components that are not solely speaker-driven



- **CS1 English is relatively more complex than CS1 Chinese**

We determine significance using Welch's *t*-tests.

Conclusions

1. Does the influence of predictability on code-switching reflect ***only* speaker-centric pressures**? 👎

2. Does the influence of predictability on code-switching **generalize** across both **writing and speech**? 👍

- ❖ Language models (both *n*-gram *and* LLMs) can be used to identify the patterns we found!

Wrap Up

How & why do speakers code-switch?

- **How?**
  - In distinctive prosodic styles
  - According to individual proficiency [INTERSPEECH ‘25]
  - In dynamic interaction with an interlocutor [NAACL ‘24]
- **Why?**
  - To convey empathy
  - *Not only* to ease production
  - To express and achieve specific discursive goals [EMNLP ‘25; CODI-CRAC ‘25]

Up Next: Bridging speech analysis and generation

- **Ongoing work**
  - Do models attend to the same entraining features that humans produce in code-switched conversations across language pairs?
  - How does conversational code-switched writing generated by LLMs currently compare to that produced by humans?
  - Do speakers perceive and prefer to hear code-switching produced with “true” code-switched prosody?

Up Next: Generating naturalistic code-switching

- **Upcoming work**
  - Applying insights gleaned in
    - training data curation recommendations, e.g. for SFT
    - priors for generative models
    - steering prompts for model controllability

Thank you!

Contact:
debasmita.b@cs.columbia.edu