Switching Contexts

Pragmatic Effects of Code-switching in Spanish-English Dialogue

Fahima Ayub Khan Bill Noble Christine Howes

9th October 2025, SweCog2025

Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability
University of Gothenburg

Table of contents

Background

Code-switching (CS)

Code-switching is the **alternation of multiple languages** in the same discourse or conversation.

Code-switching (CS)

Code-switching is the **alternation of multiple languages** in the same discourse or conversation.

Language choice has been analysed as a function of the linguistic structures, speaker proficiency, and language identity of the speaker.

Code-switching (CS)

Code-switching is the **alternation of multiple languages** in the same discourse or conversation.

Language choice has been analysed as a function of the linguistic structures, speaker proficiency, and language identity of the speaker.

More recent studies argue that **code-switching is used as a communicative resource** in multilingual interaction to emphasise relevant information.

Code-switching (CS)

Consider the following example:

(1) IS. *yeah but tú sabes qué?*

yeah but you know what?

Code-switching (CS)

Consider the following example:

(2) IS. *yeah but tú sabes qué?*

yeah but you know what?

IS. *el problema de aquí que no me gustan son los taxes is too high*

the problem that I dislike over here is that taxes are too high

MA. *I see*

MA. *siempre por lo general el impuesto es del dos por ciento taxes is too high*

generally, tax is always at 2 per cent
(Zeledon5, Miami-Bangor, Deuchar, 2010)

Code-switching in dialogue

Code-switching often coincides with pragmatic functions like repair Hlavac (2014) and turn-taking in dialogue.

Code-switching in dialogue

Code-switching often coincides with pragmatic functions like repair Hlavac (2014) and turn-taking in dialogue.

However, existing interactional approaches to code-switching do not account for the processing difficulties of code-switched utterances and the information distribution of CS in interaction.

Code-switching in dialogue

Code-switching often coincides with pragmatic functions like repair Hlavac (2014) and turn-taking in dialogue.

However, existing interactional approaches to code-switching do not account for the processing difficulties of code-switched utterances and the information distribution of CS in interaction.

Meaning predictability in CS

- Language processing difficulty is positively correlated with surprisal, i.e., the information content of the upcoming word

Meaning predictability in CS

- Language processing difficulty is positively correlated with surprisal, i.e., the information content of the upcoming word
- The predictability of the code-switch depends on both speaker-internal and speaker-external factors

Meaning predictability in CS

- Language processing difficulty is positively correlated with surprisal, i.e., the information content of the upcoming word
- The predictability of the code-switch depends on both speaker-internal and speaker-external factors
- A study by Myslín and Levy (2015) investigating both factors concluded that discourse-related contextual (I.e., speaker-external) factors such as lexical cohesion, participant constellation and meaning predictability

Research Questions

In this study, we tested the information-theoretic approach to code-switching:

- Do multilingual speakers switch languages at lower information points in dialogue?

Research Questions

In this study, we tested the information-theoretic approach to code-switching:

- Do multilingual speakers switch languages at lower information points in dialogue?
- Is high information content signalled using only one language in bilingual dialogue?

Methodology

Corpus

We used the Bangor-Miami corpus which contains 56 spontaneous conversations collected from Spanish-English bilingual speakers in Miami, USA.

Corpus

We used the Bangor-Miami corpus which contains 56 spontaneous conversations collected from Spanish-English bilingual speakers in Miami, USA.

15 conversations were excluded due to missing speaker turns.

Corpus

We used the Bangor-Miami corpus which contains 56 spontaneous conversations collected from Spanish-English bilingual speakers in Miami, USA.

15 conversations were excluded due to missing speaker turns.

Using the corpus, we measured:

1. Surprisal at code-switching points across utterances

Corpus

We used the Bangor-Miami corpus which contains 56 spontaneous conversations collected from Spanish-English bilingual speakers in Miami, USA.

15 conversations were excluded due to missing speaker turns.

Using the corpus, we measured:

1. Surprisal at code-switching points across utterances
2. Statistical likelihood of code-switching depending on the degree of informativeness and surprisal of the turn-level and dialogue context

Sampling from the corpus

We created balanced sets of mid-utterance code-switch points (n=500) and analogous points with no code-switch (n=500). These sets were created by sampling tokens as follows:

- excluded the first token and the final 5 tokens in each utterance
- excluded the first 50 utterances
- excluded separators, pauses, word fragments and other non-words
- in the CS case: the previous token is from another language

Language model-based surprisal metrics

- We use **information value (IV)** (Meister, Giulianelli, and Pimentel, 2024) as a proxy for (human) surprisal
  - **IV intuition:** Given some context, LM generations will be more similar to the actual continuation when the continuation is less surprising (given the context).
  - IV has been shown to correlate better with human surprisal when compared to more traditional LM-based metrics like perplexity (Giulianelli, Wallbridge, and Fernández, 2023)

Language model-based surprisal metrics

- We use **information value (IV)** (Meister, Giulianelli, and Pimentel, 2024) as a proxy for (human) surprisal
  - **IV intuition:** Given some context, LM generations will be more similar to the actual continuation when the continuation is less surprising (given the context).
  - IV has been shown to correlate better with human surprisal when compared to more traditional LM-based metrics like perplexity (Giulianelli, Wallbridge, and Fernández, 2023)
- We measure the IV of turn continuations following sampled mid-utterance points with *and* without the immediately proceeding dialogue context
  - `IV_turn` – surprisal of the continuation *without* considering dialogue context
  - `IV_diag` – surprisal of the continuation *with* dialogue context taken into account
  - **Context informativeness**: `CI` = `IV_turn` - `IV_diag` – how much the predictability of the continuation *depends on dialogue context*

Example LM-based surprisal metrics

@Begin
@Languages: eng, spa
@Participants: CHA Chantal Teenager, GIL Gillian Child
@Situation: Informal conversation between two cousins
@Date: 22-MAR-2008
GIL: ah Jacky está en Colombia para Navidad un año.
GIL: um yo (.) fui la día primero un poquito días de Navidad.
GIL: i and then um I stayed there for a week or like couple days or something like...
GIL: I came like a week before Christmas you know.
. . .
CHA: what are you favourite hobbies?
GIL: I u draw.
GIL: I sing.
GIL: and sometimes I go play golf.
CHA: muy chévere.
CHA: my hobbies are— I—
GIL: and swimming.
CHA: I— I love to play basketball.
CHA: I'm in my team in my school.
CHA: um I also love to swim.
CHA: you know in the pool it's very very fun.

GIL: I know.
GIL: remember the other day when we— (.) cuando nosotros fuimos
when we went
a la piscina aquí?
to the pool here?

We always include the first part of the transcript for minimal context and to expose the model to the transcription format.

This is the sampled point (in this case a CS point)

Example LM-based surprisal metrics

@Begin
@Languages: eng, spa
@Participants: CHA Chantal Teenager, GIL Gillian Child
@Situation: Informal conversation between two cousins
@Date: 22-MAR-2008
GIL: ah Jacky está en Colombia para Navidad un año.
GIL: um yo (.) fui la día primero un poquito días de Navidad.
GIL: i and then um I stayed there for a week or like couple days or something like...
GIL: I came like a week before Christmas you know.
. . .
CHA: what are you favourite hobbies?
GIL: I u draw.
GIL: I sing.
GIL: and sometimes I go play golf.
CHA: muy chévere.
CHA: my hobbies are— I—
GIL: and swimming.
CHA: I— I love to play basketball.
CHA: I'm in my team in my school.
CHA: um I also love to swim.
CHA: you know in the pool it's very very fun.

GIL: I know.
GIL: remember the other day when we— (.)

Actual continuation:
cuando nosotros fuimos a la piscina aquí?
when we went to the pool here?

Generated without dialogue context:
1. *when you asked me if I had seen this girl at school.* **0.35**
2. *when we were talking about Colombia?* **0.45**
3. *we were talking about it and you were like* **0.40**

IV_turn
0.40

mean cosine distance
from actual continuation

Example LM-based surprisal metrics

@Begin
@Languages: eng, spa
@Participants: CHA Chantal Teenager, GIL Gillian Child
@Situation: Informal conversation between two cousins
@Date: 22-MAR-2008
GIL: ah Jacky está en Colombia para Navidad un año.
GIL: um yo (.) fui la día primero un poquito días de Navidad.
GIL: i and then um I stayed there for a week or like couple days or something like...
GIL: I came like a week before Christmas you know.
. . .
CHA: what are you favourite hobbies?
GIL: I u draw.
GIL: I sing.
GIL: and sometimes I go play golf.
CHA: muy chévere.
CHA: my hobbies are— I—
GIL: and swimming.
CHA: I— I love to play basketball.
CHA: I'm in my team in my school.
CHA: um I also love to swim.
CHA: you know in the pool it's very very fun.

GIL: I know.
GIL: remember the other day when we— (.)

Actual continuation:
cuando nosotros fuimos a la piscina aquí?
when we went to the pool here?

Generated with dialogue context:
1. *we went to the pool.* **0.20**
2. *when we were playing basketball?* **0.35**
3. *when we went swimming?* **0.30**

Example LM-based surprisal metrics

@Begin
@Languages: eng, spa
@Participants: CHA Chantal Teenager, GIL Gillian Child
@Situation: Informal conversation between two cousins
@Date: 22-MAR-2008
GIL: ah Jacky está en Colombia para Navidad un año.
GIL: um yo (.) fui la día primero un poquito días de Navidad.
GIL: i and then um I stayed there for a week or like couple days or something like...
GIL: I came like a week before Christmas you know.
. . .

CHA: what are you favourite hobbies?
GIL: I u draw.
GIL: I sing.
GIL: and sometimes I go play golf.
CHA: muy chévere.
CHA: my hobbies are— I—
GIL: and swimming.
CHA: I— I love to play basketball.
CHA: I'm in my team in my school.
CHA: um I also love to swim.
CHA: you know in the pool it's very very fun.

GIL: I know.
GIL: remember the other day when we— (.)

Actual continuation:
cuando nosotros fuimos a la piscina aquí?
when we went to the pool here?

Generated without dialogue context:
1. *when you asked me if I had seen this girl at school.* **0.35**
2. *when we were talking about Colombia?* **0.45**
3. *we were talking about it and you were like* **0.40**

Generated with dialogue context:
1. *we went to the pool.* **0.20**
2. *when we were playing basketball?* **0.35**
3. *when we went swimming?* **0.30**

context informativeness

IV_turn **0.40** − IV_diag **0.30** = CI **0.10**

Analysis and Results

Statistical Analysis

To test if code-switching is likely when the information in the dialogue context is lower, we fitted a binomial mixed effects model with Context informativeness (CI) and Informations Values (IV) as the predictor to model the code-switching as the outcome

Statistical Analysis

To test if code-switching is likely when the information in the dialogue context is lower, we fitted a binomial mixed effects model with Context informativeness (CI) and Informations Values (IV) as the predictor to model the code-switching as the outcome

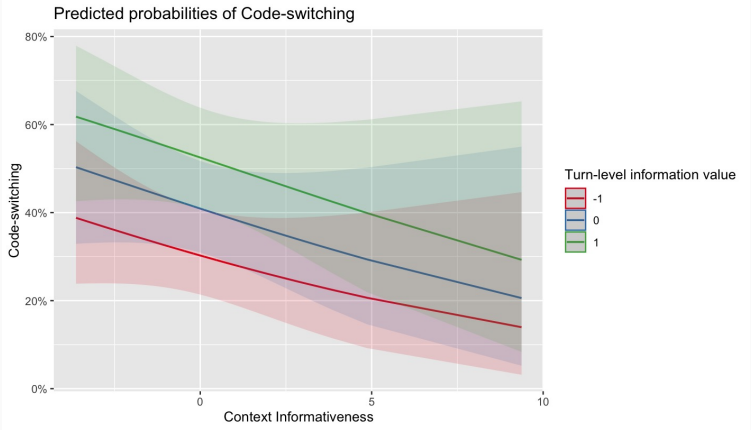
The results showed that the likelihood of code-switching increases when context-informativeness (CI) along with turn-level information value (IV_turn) decreases.

Statistical Analysis

To test if code-switching is likely when the information in the dialogue context is lower, we fitted a binomial mixed effects model with Context informativeness (CI) and Informations Values (IV) as the predictor to model the code-switching as the outcome

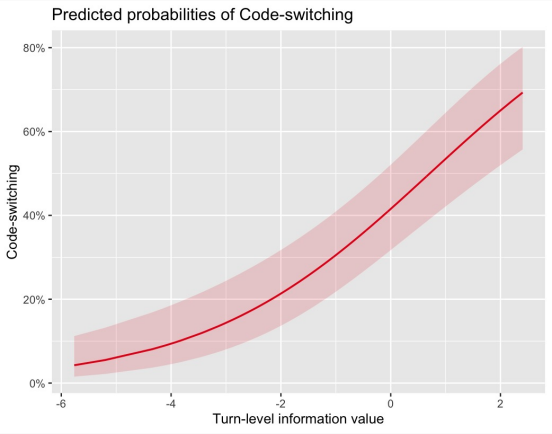
The results showed that the likelihood of code-switching increases when context-informativeness (CI) along with turn-level information value (IV_turn) decreases.

Results: Effect of Context Informativeness on Code-switching



(β = 0.47, p ¡ .001, 95% CI [0.30, 0.64])

Results: Turn-level surprisal



The probability of code-switching is higher in turns when the surprisal increases

(β = 0.47, p ¡ .001, 95% CI [0.30, 0.64])

Conclusion

Conclusion and future research

- Multilingual speakers code-switch in interaction when the information available in the preceding context is not very informative for the interlocutor to predict the upcoming utterance i.e., **whenever the speakers want to incrementally change the context, they switch languages**.

Thank you!

Questions?

fahima.ayub.khan@gu.se

bill.noble@gu.se

References i

Calvillo, Jesús et al. (2020). **"Surprisal predicts code-switching in Chinese-English bilingual text".** In: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 4029–4039.

Cromdal, Jakob and Karin Aronsson (Aug. 2000). **"Footing in Bilingual Play".** In: *Journal of Sociolinguistics* 4.3, pp. 435–457. ISSN: 1360-6441, 1467-9841. DOI: 10.1111/1467-9481.00123.

Deuchar, Margaret (2010). **BilingBank Spanish-English Bangor Miami Corpus.** DOI: 10.21415/T5J01D.

Giulianelli, Mario, Sarenne Wallbridge, and Raquel Fernández (Dec. 2023). **"Information Value: Measuring Utterance Predictability as Distance from Plausible Alternatives".** In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 5633–5653. DOI: 10.18653/v1/2023.emnlp-main.343. (Visited on 02/15/2024).

References ii

- Hlavac, Jim (2011). **"Hesitation and monitoring phenomena in bilingual speech: A consequence of code-switching or a strategy to facilitate its incorporation?"** In: *Journal of Pragmatics* 43.15, pp. 3793–3806.
- Meister, Clara, Mario Giulianelli, and Tiago Pimentel (Nov. 2024). **"Towards a Similarity-adjusted Surprisal Theory".** In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 16485–16498. DOI: 10.18653/v1/2024.emnlp-main.921. (Visited on 02/17/2025).
- Myslín, Mark and Roger Levy (2015). **"Code-switching and predictability of meaning in discourse".** In: *Language* 91.4, pp. 871–905.
- RStudio Team (2023). **RStudio: Integrated Development Environment for R.** RStudio, PBC. Boston, MA. URL: http://www.rstudio.com/.